Journal of Official Statistics
vol. 36, 3 (Sep 2020)

sciendo

Journal of Official Statistics, Vol. 36, No. 3, 2020, pp. 463–468, http://dx.doi.org/10.2478/JOS-2020-0024

Preface

The response rate is frequently seen as the most important criterion for assessing the quality of a survey. Every survey methodologist, however, knows that the response rate of a survey tells only part of the story. There is more than meets the eye and both response rates and the composition of survey nonresponse should be evaluated critically. There are multiple reasons for this.

First, response rates can be calculated in numerous ways and a first step in quality assurance is clearly reporting which disposition codes are used in the calculation of the response to a specific survey. The American Association for Public Opinion Research provides a comprehensive report of standard definitions and a response rate calculator on its website (AAPOR 2016). Transparency in nonresponse reporting is extremely important as a response percentage can be easily enhanced artificially, while the representativity of the responding sample is not enhanced at all. A clear example is redefining the target population to those persons, who already have indicated in a previous survey that they would be quite willing to participate in another survey. Fieldwork procedures and traditions also influence response rates. For example, response rates will increase when nonrespondents can be replaced by more available or more cooperative family members or neighbours, or by focusing additional fieldwork efforts on the easiest cases. These strategies can be an official part of the research protocol, but should always be reported. It is also possible to enhance response rates in illegal ways, by not allowed substitutions or plain falsification of interviews.

Second, there is a difference between nonresponse and nonresponse bias. Nonresponse bias is a function of the response rate and the difference between respondents and nonrespondents on the variables of interest. A high response rate reduces the risk of nonresponse bias. However, when respondents differ considerably from nonrespondents with regard to the core variables of the survey, bias due to nonresponse can be considerable even with high response rates. Finally, there are many other factors besides nonresponse that determine the quality of a survey. A poorly designed questionnaire or one with ill translated questions will not provide any useful results, despite a high response rate.

Still, decreasing response rates are a major concern in the survey world, for academic, governmental, and market research surveys. One of the major reasons is nonresponse bias. Other reasons are the perceived legitimacy of surveys associated with response rates and the small number available for analysis when only a minor part of the invited sample persons participate. Finally, increased efforts to maintain acceptable response rates have consequences for survey costs and may cause practical problems in setting up a survey when response rates are expected to be low or unknown.

In the past decades, the nonresponse problem has received wide attention among survey methodologists and statisticians. This was triggered by the founding of the International

Workshop on Household Survey Nonresponse and the resulting research. In July 1989, Bob Groves, then at the Survey Research Center at the University of Michigan, sent a letter to three Swedish statisticians, outlining his ideas for collaborative work on survey nonresponse. In April 1990, they, now joined by Vladimir Andreyenkov from the former USSR, introduced the idea of a Nonresponse Workshop to interested colleagues around the world. The initiative met great approval and the International Workshop on Household Survey Nonresponse was founded in 1990 in Stockholm by Bob Groves (USA), Lars Lyberg (Sweden), and Bob Barnes (UK).

From the onset, the aim of the workshop was to bring together scientists from different countries, different disciplines, and different organizations, in order to pool knowledge and to stimulate a coordinated research agenda and international collaboration. The results can be found in numerous publications on nonresponse. Examples are the pioneering work of Morton-Williams (1993), the groundbreaking book of Groves and Couper (1998), three special issues of the Journal of Official Statistics (JOS 15(2) 1999; JOS 17(2) 2001; JOS 27(4) 2011), an edited book on international perspectives of nonresponse (Laaksonen 1996), a special issue of ZUMA-Nachrichten (Koch and Porst 1998), and a special issue of the Annals of the American Academy of Political and Social Science (2013, 645). Other important contributions are a monograph on survey nonresponse edited by Groves et al. (2002), and the Hunt for the Last Respondent (Stoop 2005). In addition, a very large number of articles have been published in major journals, as well as book chapters in monographs and handbooks of survey methodology.

Since its founding in 1990, the Workshop has been hosted by different countries each year. In 1999, the Workshop extended to a large, international conference on nonresponse, which took place in Portland, Oregon, United States. For 2020, the thirtieth meeting was again planned in Sweden, this time at Örebro University. At this moment, we are still in the midst of the COVID-19 pandemic; many countries are in partial lock-down and at present large scale international travel is not possible. Therefore, we will not be able to meet in person in Örebro, but will meet with the aid of modern technology at a virtual Nonresponse Workshop online. The jubilee meeting of International Workshop on Household Survey Nonresponse coincides with another jubilee. This year it is 35 years since the first issue of the Journal of Official Statistics (JOS) was published by Statistics Sweden. JOS was founded by Lars Lyberg, who also served as chief editor for the first 25 years. So, it only seems fitting to celebrate both jubilees with a special issue of JOS on nonreponse, and so honour the founding fathers of the international nonresponse workshop: Lars Lyberg, Bob Groves, Vladimir Andreyenkov, and Bob Barnes, and the founder and first editor of JOS: Lars Lyberg.

In the last three decades much has changed. Stoop (2016) presents an overview of developments in the nonresponse research agenda over the years. Where the first attempts to understand nonresponse focused on studying correlates of nonresponse, later this shifted to attempts to better understand nonresponse (Groves et al. 1992; Hox et al. 1995) and the development of theories on survey participation (e.g., Groves et al. 2000; Singer 2011; Dillman et al. 2014; Dillman forthcoming). Likewise, studies into international nonresponse trends (De Leeuw and De Heer 2002; Beullens et al. 2018) and survey climate (Loosveldt and Joye 2016), led to the nagging question why response rates vary so much across countries (see also Stoop et al. 2010). The early collection and analysis of

interviewer observations (e.g., Campanelli, et al. 1997) has widened into the collection and analysis of paradata (Kreuter 2013). Surveys of interviewers are still being conducted in special projects shedding some light on the question why some interviewers perform better than others (e.g., Japec 2008; Blom et al. 2011).

However, it should be noted that in western countries interviewer-mediated surveys are costly and are reserved for official statistics and special projects, such as the European Social Survey (ESS) and the Surveys on Health and Aging across Europe (SHARE). Also, the characteristics and motivation of respondents has been extensively studied; here the focus shifted from studying the difficult to contact and difficult to convince respondents (e.g., Stoop 2005, chap. 7, 8, and 9) to the development of fieldwork adaptations and responsive and adaptive survey design (e.g., Schouten et al. 2017). For an overview and extensive discussion, see the special issue and special section of JOS on adaptive designs (JOS 33(3) 2017; JOS 34(3) 2018). Finally, the attention shifted from studying mere response propensity to the relationship between response propensity and measurement error.

So where are we now? We still wonder why response rates show such a wide variation across countries. We are still concerned that we may not be aware of crucial details in data collection that may enhance or reduce response rates. We know that organizational factors may have a large impact, for example, when funds diminish or fieldwork is stopped because another survey gets priority. We may agree with Brick (2013, 346, 347) that the ". . . central problem, in our opinion, is that even after decades of research on nonresponse we remain woefully ignorant of the causes of nonresponse at a profound level". We appreciate that research ethics and data protection are becoming increasingly important, although they sometimes may make data collection more complicated. We do agree that it is paramount (Brick 2013, 346, 347) to advertise the importance of high quality survey data (and the fact that this requires sufficient funding).

The contributions to this special issue of JOS cover a wide area and touch on several of the problems highlighted above.

Changes in response over time and across countries and attempts to understand these are central in three contributions. Luiten, Hox and De Leeuw describe developments in international response trends, and try to explain differences between countries, thus incorporating differences in survey culture. Larsen, Fane Lineback and Reist pay attention to the survey climate, especially economic and political conditions, and the increase in refusals to governmental surveys in the US. Finally, Friedel and Birkenbach study retention rates and R-indicators and focus on changes in the composition of a cross-national longitudinal survey (SHARE).

Three contributions focus on interviewers. Min, Schaeffer, Garbarski and Dykema analyse the action structure of recruitment calls and the impact of interviewers on acceptance of a request for an interview. Wuyts and Loosveldt measure the consequences of interviewer workloads and explore its effects on interviewer performance. Related is the contribution by Peeters, De Coninck, Wuyts, and Loosveldt, who assess interviewer performances in re-approaching reissued initial nonrespondents.

Survey design, the respondent, and adaptive surveys are addressed in two contributions. Van Berkel, Van der Doef, and Schouten describe the implementation of an adaptive survey design in the Dutch Health Survey, using multiple modes. Fowler, Brenner,

Roman, and Hargraves study a mixed mode design and compare telephone and call-in IVR surveys of address-based samples on nonresponse bias.

Finally, response propensity and error are discussed. Bethlehem presents an overview showing how response probabilities can be estimated, even when there is no complete set of auxiliary variables for respondents and nonrespondents. The analyses of Roberts, Vanderplas and Herzing delve into potential limitations of R-indicators and the suitability of auxiliary data used for estimating these, and Andridge and Little introduce a proxy pattern-mixture analysis for the assessment of the impact of nonresponse for binary variables.

This special issue aims to provide some additional evidence why nonresponse is a problem, what could be the causes, and why and how it should be fought. Still, the battle continues and we hope that this special issue stimulates further research that will deepen our understanding of nonresponse, its causes and consequences, and helps us to improve our tools for reducing nonresponse by improved fieldwork designs and for sophisticated statistical adjustment of nonresponse bias.

Edith de Leeuw
Guest Editor

Annemieke Luiten
Ineke Stoop
Guest Associate Editors

References

AAPOR. 2016. *Standard Definitions' Final Dispositions of Cases Coded and Outcome Rates for Surveys. Revision 2016*. Available at: www.aapor.org/Standards-Ethics/-Standard-Definitions-(1).aspx (accessed April 2020).

Annals of the American Academy of Political and Social Science. 2013. "The Nonresponse Challenge to Surveys and Statistics." *Annals of the American Academy of Political and Social Science*, vol. 645. https://doi.org/10.1177/0002716212456815.

Beullens, K., G. Loosveldt, C. Vandenplas, and I. Stoop. 2018. "Response Rates in The European Social Survey: Increasing, Decreasing, or a Matter of Fieldwork Efforts?" *Survey Methods: Insights from the Field*. DOI: https://doi.org/10.13094/SMIF-2018-00003.

Blom, A.G., E.D. de Leeuw, and J. Hox. 2011. "Interviewer Effects on Nonresponse in the European Social Survey." *Journal of Official Statistics* 27(2): 359–377. Available at: https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/interviewer-effects-on-nonresponse-in-the-european-social-survey.pdf (accessed April 2020).

Brick, J.M. 2013. "Unit Nonresponse and Weighting Adjustments: A Critical Review." *Journal of Official Statistics* 29: 329–352. DOI: https://doi.org/10.2478/jos-2013-0026.

Brick, M.J. and D. Williams. 2013. "Explaining rising response rates in cross-sectional surveys." *Annals of the American Academy of Political and Social Sciences* 645: 36–59. DOI: https://doi.org/10.1177/0002716212456834.

Campanelli, P.C., P. Sturgis, and S. Purden. 1997. *Can You Hear Me Knocking? An Investigation into the Impact of Interviewers on Survey Response Rates.* London: SCPR. Available at: https://www.researchgate.net/publication/312910100_Can_you_hear_me_knocking_and_investigation_into_the_impact_of_interviewers_on_survey_response_rates (accessed April 2020).

De Leeuw, E., and W. de Heer. 2002. Trends in Household Survey Nonresponse: A Longitudinal and International Comparison. In *Survey Nonresponse*, edited by R.M. Groves, D.A. Dillman, J.L. Eltinge, R.J.A. Little, 41–54. New York: Wiley.

Dillman, D.A. (forthcoming). "Towards Survey Response Rate Theories that no longer pass each other like strangers in the night." To appear in: Brenner, Philip (Ed). *Understanding Survey Methodology: Sociological Theory and Applications*. Springer books.

Dillman, D.A., J.D. Smyth, and L.M. Christian. 2014. *Internet, Phone, Mail and Mixed-Mode Surveys: The Tailored Design Method*. John Wiley: Hoboken, NJ.

Groves, R.M., R.B. Cialdin, and M.P. Couper. 1992. "Understanding the Decision to Participate in a Survey." *Public Opinion Quarterly* 56(4): 475–495. DOI: https://doi.org/10.1086/269338.

Groves, R.M. and M.P. Couper. 1998. *Nonresponse in Household Interview Surveys*. New York: Wiley. DOI: https://doi.org/10.1002/9781118490082.

Groves, R.M., D.A. Dillman, J.L. Eltinge, and R.J.A. Little, (Eds.) 2002. *Survey Nonresponse*. New York: Wiley.

Groves, R.M., E. Singer, and A. Corning. 2000. "Leverage-Saliency Theory of Survey Participation. Description and an Illustration." *Public Opinion Quarterly* 64: 299–308. DOI: https://doi.org/10.1086/317990.

Hox, J.J., E.D. de Leeuw, and H. Vorst. 1995. "Survey Participation as Reasoned Action; A Behavioral Paradigm for Survey Nonresponse?" *Bulletin de Méthodologie Sociologique (BMS)* 48: 52–67. DOI: https://doi.org/10.1177/075910639504800109.

Japec, L. 2008. Interviewer Error and Interviewer Burden. In J.M. Lepkowski, C. Tucker, J.M. Brick, E.D. de Leeuw, L. Japec, P.J. Lavrakas, M.W. Link, and R.L. Sangster (Eds.), *Advances in Telephone Survey Methodology*: 187–211. Hoboken: Wiley.

JOS. 1999. "Special Issue on Nonresponse." *Journal of Official Statistics* 15(2). available at: https://www.scb.se/en/documentation/statististical-methods/journal-of-official-statistics-jos/ (accessed April 2020).

JOS. 2001. "Special Issue on Nonresponse." *Journal of Official Statistics* 17(2). Available at: https://www.scb.se/en/documentation/statististical-methods/journal-of-official-statistics-jos/ (accessed April 2020).

JOS. 2011. "Special Issue on Nonresponse." *Journal of Official Statistics* 27(4). Available at: https://www.scb.se/en/documentation/statististical-methods/journal-of-official-statistics-jos/ (accessed April 2020).

JOS. 2017. "Special Issue on Responsive and Adaptive Survey Design: Looking Back to See Forward." *Journal of Official Statistics* 33(3). Available at: https://www.scb.se/en/-documentation/statististical-methods/journal-of-official-statistics-jos/ (accessed April 2020).

JOS. 2018. "Special Section on Responsive and Adaptive Survey Design." *Journal of Official Statistics* 34(3). Available at: https://www.scb.se/en/documentation/statististical-methods/journal-of-official-statistics-jos/ (accessed April 2020).

Koch, A. and R. Porst. 1998. *Nonresponse in Survey Research*. ZUMA SPEZIAL, Volume 4, Mannheim: ZUMA. Available at: https://www.gesis.org/en/services/publications/archive/zuma-and-za-publications/zuma-nachrichten-spezial (accessed April 2020).

Kreuter, F. ed. 2013. *Improving Surveys with Paradata: Analytic Uses of Process Information*. John Wiley & Sons, Inc., Hoboken, New Jersey. DOI: https://doi.org/10.1002/9781118596869.

Laaksonen, S. 1996. *International perspectives on Nonresponse*. Helsinki, Statistics Finland. Available at: https://www.doria.fi/bitstream/handle/10024/176203/xtut_219_dig.pdf?sequence=1&isAllowed = y (accessed April 2020).

Loosveldt G. and D. Joye. 2016. Defining and Assessing Survey Climate In: C. Wolf, D. Joye, T.W. Smith, & Y. Fu (Eds.) *The SAGE Handbook of Survey Methodology*: 67–76. SAGE Publications Ltd. DOI: https://doi.org/10.4135/9781473957893.n6.

Morton-Williams, J. 1993. *Interviewer Approaches*. Aldershot: Dartmouth.

Singer, E. 2011. Towards a Cost-Benefit Theory of Survey Participation: Evidence, Further Test, and Implications. *Journal of Official Statistics* 27(2): 379–392. Available at: https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/toward-a-benefit-cost-theory-of-survey-participation-evidence-further-tests-and-implications.pdf (accessed June 2020).

Schouten, B., A. Peytchev, and J. Wagner. 2017. *Adaptive Survey Design*. *Series on Statistics Handbooks*. Chapman & Hall/CRC. DOI: https://doi.org/10.1201/9781315153964.

Stoop, I.A.L. 2005. *The Hunt for the Last Respondent*. The Hague, Social and Cultural Planning Office.

Stoop, I.A.L. 2016. Unit Nonresponse. In: C. Wolf, D. Joye, T.W. Smith, & Y. Fu (Eds.) *The SAGE Handbook of Survey Methodology*: 409–424. SAGE Publications Ltd. DOI: https://doi.org/10.4135/9781473957893.n27.

Stoop, I., J. Billiet, A. Koch, and R. Fitzgerald. 2010. *Improving Survey Response. Lessons Learned from the European Social Survey*. Chichester, John Wiley & Sons. Ltd. DOI: https://doi.org/10.1002/9780470688335.

sciendo

Journal of Official Statistics, Vol. 36, No. 3, 2020, pp. 469–487, http://dx.doi.org/10.2478/JOS-2020-0025

Survey Nonresponse Trends and Fieldwork Effort in the 21st Century: Results of an International Study across Countries and Surveys

Annemieke Luiten[1], *Joop Hox*[2], *and Edith de Leeuw*[2]

For more than three decades, declining response rates have been of concern to both survey methodologists and practitioners. Still, international comparative studies have been scarce. In one of the first international trend analyses for the period 1980–1997, De Leeuw and De Heer (2002) describe that response rates declined over the years and that countries differed in response rates and nonresponse trends. In this article, we continued where De Leeuw and De Heer (2002) stopped, and present trend data for the next period 1998–2015 from National Statistical Institutes. When we looked at trends over time in this new data set, we found that response rates are still declining over the years. Furthermore, nonresponse *trends* do differ over countries, but not over surveys. Some countries show a steeper decline in response than others, but all types of surveys show the same downward trend. The differences in (non)response trends over countries can be partly explained by differences in survey design between the countries. Finally, for some countries cost indicators were available, these showed that costs increased over the years and are negatively correlated with noncontact rates.

Key words: Response trend; noncontact; refusal; survey design; fieldwork; costs.

1. Introduction

Response rates are often seen as a major quality indicator for surveys by both data users and survey organizations (Stoop 2005). Statisticians point out that nonresponse threatens the sample selection mechanism of a survey where each member of the populations has a known and non-zero probability of being included, and as a consequence the validity of inference about the population may be at stake (Bethlehem et al. 2011; Brehm 1993; Groves and Couper 1998). Furthermore, nonresponse results in smaller realized sample

[1] Statistics Netherlands, P.O. Box 4481, 6401 CZ Heerlen. the Netherlands. Email: a.luiten@cbs.nl
[2] Department of Methodology and Statistics, Utrecht University, P.O. Box 80140, 3508 TC, Utrecht, the Netherlands. Emails: joophox@xs4all.nl and e.d.deleeuw@uu.nl

Acknowledgments: The authors thankfully acknowledge all researchers who so very generously gave their time to complete our nonresponse questionnaire – without their effort this study could not have been completed. We also thank Barry Schouten for his helpful comments on the data collection procedure and the participants of the International Workshop on Household Survey Nonresponse 2016, 2017 and 2018 for their extremely useful feedback on earlier versions. Thanks are also due to the reviewers and editors for their insightful comments. Finally, we thank and honour Wim de Heer, who designed and completed the first international nonresponse study. The views expressed are those of the authors and do not necessarily reflect the policies of Statistics Netherlands.

sizes, and may result in longer fieldwork periods, and increased costs per completed interview.

Nonresponse occurs when a sample unit does not respond to a request to be surveyed. Two major components of unit nonresponse are non-contacts and refusals, and besides response rates, contact rates and refusal rates are important indicators for the evaluation of surveys and their fieldwork (AAPOR 2016). Therefore, it is not surprising that survey methodologists and statisticians have been concerned with the contactability and willingness of sample units (Groves and Couper 1998; Stoop 2005). This distinction was emphasized by Norman Bradburn, who in his 1992 presidential address to the American Association of Public Opinion Research declared "We all *believe* strongly that response rates are declining and have been declining for some time." Bradburn (1992, 392) then explicitly pointed out that "Part of the problem is locating respondents and part of the problem is getting respondents." A similar conviction was expressed by Groves (1989, 182) who stated that: "Participation in social surveys appears to be declining in the United States over time. This is true for government, academic, and commercial surveys."

One of the first trend studies in the United States was by Steeh (1981), who reported that between 1952 and 1979 refusals did increase in two major academic face-to-face interview studies (the US National Election Studies and the US Consumer Attitudes Survey); however Steeh did not study noncontacts. Curtin et al. (2005) investigated response trends for the US Survey of Consumer Attitudes covering the later period 1979 to 2003. During that period, data were collected using telephone interviews. Like Steeh (1981), Curtin et al. (2005) found that overall response rates for the Survey of Consumer Attitudes were decreasing over time; in addition, they showed that this decrease was partly caused by an increase in refusals and partly by an increase in noncontacts, especially after 1985.

A large comparative study for several governmental health and economic interview surveys in the United States was performed by Atrostic et al. (2001), who investigated nonresponse trends for six major US household surveys: the Current Population Survey (CPS), the Consumer Expenditure Survey (CE), the Diary and Quarterly Surveys (CED and CEQ), the National Health Interview Survey (NHIS), the National Crime Victimization Survey (NCVS) and the Survey of Income and Program Participation (SIPP). Although these surveys differ in design and fieldwork, they also have design and data collection features in common, and the first interview is always conducted by a personal visit after an introductory letter is sent to the household's address. Atrostic et al. (2001) concluded that overall nonresponse increased for all six surveys over the period 1990 to 1999 with a clear increase in 1994. Again, this increase in nonresponse was caused by an increase in both refusals and noncontacts, with the noncontact rates showing the greatest relative increase.

Building on these earlier studies, Williams and Brick (2018) updated nonresponse trends for nine face-to-face household surveys conducted in the United States since 2000. These surveys covered a variety of topics, mostly on health related issues but also on economics, crimes and social attitudes and include the CPS, NHIS and NCVS mentioned by Atrostic et al. (2001), but also the General Social Survey (GSS) and the National Survey on Drug Use and Health (NSDUH), the Medical Expenditure Panel Survey (MEPS), the Medicare Current Beneficial Survey (MCBS), the National Health and Nutrition Examination Survey (NHANES) and the National Survey of Family Growth

(NSFG). Most surveys had high response rates in the 1990s, but response was slowly decreasing over time. Williams and Brick (2018) concluded that in the United States, the overall response rates for nine major face-to-face surveys clearly decreased in the period 2000–2014, and are at present at 70% to 80%. This decrease in response is attributed to a clear increase in both the noncontacts and the refusals; with refusals as the main reason for nonresponse.

The still relatively high response rates of 70–80% for face-to-face surveys in the United States are in sharp contrast to those for telephone interviews. Curtin et al. (2005) for the Survey of Consumer Attitudes reported a decline from 72% in 1979 to 48% in 2003, and ten years later in 2013 the response had dropped to 16% (Dutwin and Lavrakas 2017). Furthermore, Pew Research (2012) reported a decrease in response rates from 36% in 1997 to 9% in 2012. Again, both contactability and willingness to cooperate declined. From 1997 to 2012, contact rates dropped from 90% to 62% and cooperation rates more than halved from 43% to 14%. This is corroborated by results for the Gallup Poll Social Survey Series (Marken 2018), which showed a drop in response rates from 28% in 1997 through 9% in 2012 to 7% in 2017.

Traditionally, studies on nonresponse trends were mainly done in the United States; European and international trend data on nonresponse are much scarcer. An early study by Hox and De Leeuw (1994) summarized data from 45 mode comparison studies in Europe, the United States and Canada for the period 1947–1992. Using meta-analytic techniques, they concluded that although face-to face surveys still obtained the highest overall response rates, there was a downward trend over the years. A similar downward trend in response was found for telephone surveys, but not for postal mail surveys. The relative stable trend for mail surveys was attributed to the increased research effort and attention for improving mail surveys in that period (e.g., Dillman 1978), counteracting a potential downward trend.

To accommodate the need for comparable international response data, an initiative was started at the first International Workshop on Household Survey Nonresponse to collect longitudinal data. An international nonresponse questionnaire was developed that was sent yearly to contacts at governmental survey agencies in different countries. The goal was to collect long-term, cross-national comparable data from governmental survey agencies on surveys that were regularly conducted. The questionnaire contained questions on response, noncontact and refusal rates, as well as questions on sampling and survey design, fieldwork and survey organization. This resulted in a final data set covering the period 1980–1997 and time series were available for 16 countries and 10 different interview surveys (De Heer 1999; De Leeuw and De Heer 2002). The main data collection mode used was the face-to-face interview, although some National Statistical Institutes (Finland, Sweden) used telephone interviews. A multilevel analysis by De Leeuw and De Heer (2002) showed that response rates had indeed been declining. Both noncontact rates and refusal rates increased over the period 1980–1997, with an average of 0.2% per year for noncontact rates and 0.3% per year for refusal rates. Furthermore, countries and surveys differed in the acceleration of refusal rates, while noncontact rates decreased to the same extent across countries.

Recently, Beullens et al. (2018) analyzed seven rounds of the European Social Survey (ESS). Response data for the ESS, which is a face-to-face survey, from 36 European countries were available. As the ESS is fielded every two years, the data set covered a time

period of 12 years (2002–2014). Beullens et al. (2018) report a tendency for response rates to decrease over time. They also conclude that their findings support the results reported by De Leeuw and De Heer (2002) for international official statistics. Furthermore, Beullens et al. (2018) point out that contrary to other nonresponse trend studies, the noncontact rates appear to be more or less stable in the European Social Survey, probably due to the increased fieldwork efforts over the years in many countries for the ESS. Therefore, they attribute the declining response rates in the ESS to the increase in refusal rates and conclude that obtaining cooperation has become increasingly difficult over time.

In 2015, an initiative was taken at the yearly International Nonresponse Workshop to collect new international response data following up the analyses by De Leeuw and De Heer (2002), who described the period 1980–1997. The new data set covered the period 1998–2015. For the Labour Force Survey (LFS) the response trend data (1980–1997) from De Leeuw and De Heer (2002) were still available. De Leeuw et al. (2018) combined the new LFS response trend data (1998–2015) with the old LFS response trend data (1980–1997). They concluded that for the Labour Force Survey, the trends visible in De Leeuw and De Heer (2002) continue almost unchanged over time with possibly a small deceleration in refusal rates. The response for the LFS decreased by an average of 0.73% each year. Both the noncontacts and the refusals increased over the total time period, but countries differed in refusal and noncontact rates. De Leeuw et al. (2018) also detected that the difference in year trends for refusals between the old (1980–1997) data and the recent (1998–2015) data was significant. This indicates that although refusals for the LFS are still increasing in the new millennium, the rate of increase is slightly smaller. This study could not investigate the differences in trends any further due to a lack of descriptive variables, as the full 1980–1997 data set and the completed questionnaires unfortunately were no longer archived. The *still available* 1980–1997 data set contained, besides response trend data, only the information whether the survey was mandatory or not. As could be expected, mandatory surveys did result in fewer refusals than voluntary surveys and hence in a higher response, but the response *trends* in mandatory and voluntary surveys were comparable (De Leeuw et al. 2018).

In sum: In the US response rates have been declining over the years for a variety of surveys. This trend was stronger for telephone than for face-to-face interviews. The growing nonresponse was partly caused by an increase in noncontacts and partly by an increase in refusals. In addition, for international surveys a clear decline in response could be found and an increase in refusals could be clearly discerned. However, Beullens et al. (2018) did not find an increase in noncontacts for the European Social Survey. On the other hand, De Leeuw et al. (2018) do detect an increase in both noncontacts and nonresponse for a very long international time series concerning the Labour Force Survey.

Neither Beullens et al. (2018) nor De Leeuw et al. (2018) could directly link fieldwork and context variables to nonresponse data in order to explore the nonresponse trends and differences found. Furthermore, the type of survey (ESS vs LFS) could play a role too, as the LFS and ESS clearly differ in topic and type of questions asked, which raises the questions of the generalizability of the findings. In order to investigate this further, we analyzed the new nonresponse data set collected in 2015 (Luiten et al. 2016). This new data covers the period 1998–2015 only, but does contain data for a variety of surveys and

information on meta-variables describing fieldwork and design. This enables us to address the following research questions:

1. In the first two decades of the new millennium, can decreasing response trends in noncontacts and refusals be observed for international surveys, as in the United States?
2. Are these trends generalizable over different surveys or do surveys differ in nonresponse trends?
3. Are these trends different for different countries?
4. Which factors in survey design and fieldwork effort are related to nonresponse trends?

2. Methods

2.1. Data Collection

In 2015, a new version of the International Questionnaire on Nonresponse was developed (Luiten et al. 2016). This new questionnaire was based on the original questionnaire as developed and used by De Heer (1999, see also De Leeuw and De Heer 2002), with added questions on mixed-mode-designs, fieldwork efforts, and costs. These variables were included based on their potential influence on noncontacts and refusals as reported in the literature on nonresponse (Biemer and Lyberg 2003; Dillman et al. 2002; Dillman et al. 2014; Groves and Couper 1998; Groves et al. 2004; Stoop 2005, 2016). Two versions of the questionnaire were available, one for the Labour Force Survey (LFS) and one for another important social survey of choice. The two versions overlap in questions asked, but for replicability and transparency both questionnaires are fully available as online supplemental material, Section 1.

The data were collected in 2016 and the two questionnaires were sent out to National Statistical Institutes (NSI) in Europe, Australia, Canada, and the United States. Respondents were asked to report on response, refusal and contact rates for the period 1998 to 2015 for the LFS and for one other important social survey at their institute. In addition, they were asked to give details on design, fieldwork efforts and costs, and on any changes made from 1998 to 2015.

In total, 25 countries participated, both European and non-European countries: Austria, Australia, Belgium, Bulgaria, Canada, Croatia, Iceland, Finland, France, Germany, Hungary, Italy, Latvia, Lithuania, Malta, the Netherlands, Norway, Poland, Portugal, Slovakia, Slovenia, Sweden, Switzerland, the United Kingdom, and the United States. All countries provided information on the LFS.

Several countries also provided information for one or more social surveys. Nine countries completed the full questionnaire with longitudinal response data (1998–2015) for the Household Budget Survey (HBS), and three countries returned a completed questionnaire and reported response time series (1998–2015) for the Survey of Income and Living Conditions (SILC). Furthermore, one county provided data for the European Social Survey (ESS), one for the National Travel Survey (NTS), one for the Consumer Barometer (CB), and one for the Survey on Social Cultural Changes (SCC). A detailed description of these seven surveys, number of institutes, and a summary of design and fieldwork procedures can be found in the online supplemental material, Section 2.

Countries not only differ in survey design and fieldwork procedures but also in demographics and economic conditions. These macro-level factors, which are not under the control of the researcher, may influence response (Groves and Couper 1998; De Leeuw and De Heer 2002; Bethlehem et al. 2011). Therefore, in addition to the available questionnaire data, we collected economic and demographic data for all participating countries over the relevant time period 1998–2015. Available data from the Worldbank, the Organization for Economic Cooperation and Development (OECD), and Eurostat, included (1) 'employment rate, persons 15–64', (2) 'GDP per capita' (gross domestic product by population), which is an indicator of a country's standard of living, (3) the GINI-coefficient, which measures inequality in income and wealth, (4) the life expectancy at birth, and (5) the percentage of single households in a country.

2.2. Data Cleaning and Index Construction

In a preliminary analysis, the data were screened for missingness and variance. Variables that contained a large fraction of missing values (such as interviewer workload) were omitted. In addition, variables that showed (almost) no variance across countries and years (such as the use of an advance letter) were omitted. See also the online supplemental material, Section 2.

2.2.1. Index Construction

There are five economic and demographic macro-level variables available for the analysis of 25 countries. In general, multivariate analyses require about ten cases per variable (Tabachnick and Fidell 2013), so we can use only two to three variables. To summarize the macro-level information, we carried out a latent class analysis (Muthén and Muthén 2017, 8.2.) with the five macro-level variables as class indicators and two levels: years within countries. At the within-country level, one year was added as a control variable, and a full covariance matrix was estimated. A two-class model had the best fit and had a straightforward interpretation: the two classes clearly distinguished between countries with low and high economic development. Countries in Class 1 were Bulgaria, Croatia, Hungary, Latvia, Lithuania, Malta, Poland, Portugal, Slovakia, and Slovenia. Countries in Class 2 were Australia, Austria, Belgium, Canada, Finland, France, Germany, Iceland, Italy, the Netherlands, Norway, Sweden, Switzerland, the United Kingdom, and the United States. A class membership variable was added as contextual variable to the data.

2.3. Data Analysis

The literature on nonresponse (Bethlehem et al. 2011; Biemer and Lyberg 2003; Dillman et al. 2002; De Heer 1999; De Leeuw and De Heer 2002; Groves and Couper 1998; Groves et al. 2004; Stoop 2005; Stoop et al. 2010) distinguishes three theoretical groups of variables that influence nonresponse. (1) Context or macro level variables that describe the social and economic environment and are not under the influence of the researcher; an example is the economic conditions of a country. (2) General design variables that are mostly under the control of the researcher or agency, such as sampling procedures. (3) Fieldwork organization/effort related variables that are under the control of the researcher, but may be restricted by available budget, such as use of incentives.

After screening on missingness and variance, the available predictor variables were assigned to the three groups. The assignment was based on the literature cited. The latent class indicator economic development was used as a 'country context' variable. Available 'general design' variables that may influence the response are, for instance, mandatory versus voluntary survey, mixed-mode used, type of sample, oversampling, substitution allowed, proxies allowed. Examples of available 'effort in field work' variables are interviewer rewards, refusal conversion, reassignment of interviewers, respondent incentives. All variables were recorded in such a way that a high score indicates a high frequency or amount of the referred attribute. For a complete list of explanatory variables and their classification in the three groups, see Appendix, Section 5.

A cross-classified multilevel analysis (Hox, et al. 2018) was carried out with years nested within the cross-classification of countries and surveys. Dependent variables were response rate (N = 535), noncontact rate (N = 485), and refusal rate (N = 485). These rates were expressed as proportions and because the distributions were rather skewed, a logit transformation was performed to increase normality of the distributions. To model the trend over time, year of data collection was included. For ease of interpretation, this was coded as 1998 = 0, 1999 = 1 and so on. As a control variable, we added a variable that indicated whether ineligibles were excluded or not in the response rate as reported by the agencies (RRcorrected).

3. Results: International Response Trends 1998–2015

3.1. International Trends: Preliminary Analyses

We have data on the LFS and a variety of other social surveys. To analyze whether the LFS and the other social surveys show the same response trends, we performed separate multilevel analyses on response, noncontact, and refusal as dependent variable and with year as explanatory variable, for the LFS and the other surveys. This enabled us to formally test the equality of the regression coefficients of the LFS versus the social surveys (Guilford and Fruchter 1978, 148).

The analyses showed that only the intercepts differed between surveys, but not the trends over time. The differences in intercepts indicated a higher overall response and fewer refusals for the Labour Force Survey than for the other social surveys ($p < .01$), but no differences in noncontacts. However, the trends over time for overall response, noncontact, and refusal rates did not differ between the different types of surveys. The results indicate that we may combine the different surveys into one analysis, provided we accommodate the differences in intercepts for refusal and response. Therefore, the substantive analyses are specified as cross-classified nested data structures with years nested within a cross-classification of surveys and countries (Hox et al. 2018). In addition, we include the slope variance for years in countries and surveys. For details of the preliminary analyses, see online supplemental material, Section 3.

3.2. International Trends: Response Rates, Noncontacts, and Refusals

We investigate whether international surveys, like US-based surveys, show decreasing response rates in the first two decades of the new millennium. In our analysis of

international trend data for the period 1998–2015, we focus on three dependent variables. Besides general response rate, we distinguish between the two main components of nonresponse: noncontacts and refusals. Both noncontacts and refusals contribute to the overall nonresponse, but different factors influence each source (Bradburn 1992; Groves and Couper 1998; Stoop 2005). Furthermore, contact rates and refusal rates are important indicators for the evaluation of surveys and their fieldwork (AAPOR 2016).

In addition, we investigate if these trends generalize over different surveys and if these trends are different for different countries. Finally, we explore which factors in survey design and fieldwork effort are related to nonresponse trends.

Analyses were performed separately for the logit of proportion overall response, noncontacts, and refusals. We start with the 'null'-model, a model without any explanatory variables. The results for the null model are summarized in Table 1.

The significant variances over countries and over surveys estimated in this null-model show that the absolute (non)response rates differ between countries and between surveys. Furthermore, the first column shows that the variance for countries (0.56) is larger than for surveys (0.15). This means that the differences in absolute response rates between the countries are far larger than the differences in response rates between different surveys. The same conclusions hold for noncontact rate (Column 2) and refusal rate (Column 3): there are larger differences between countries concerning noncontact and refusal rates than between surveys.

The response does differ between countries and between surveys, but is there a general downward trend in survey response internationally over the years 1998–2015 (Research Question 1). To answer this research question, we added the explanatory variable 'year' to the (null) model. Again, the dependent variables were response, noncontact, and refusal, and all surveys (LFS and social surveys) were analyzed together. First, we tested if year had a significant variance component at the country or survey level. The analysis showed that there is no significant variance for year between surveys (smallest $p > .08$), but there is a significant variance between countries (largest $p < .01$ for all three tests). This is an important result, indicating that, although surveys do differ in absolute response, the trend over time is the same for all surveys. Thus, although the topic of the survey may influence the achieved response rate (Dillman et al. 2002; Groves and Couper 1998; Stoop 2005), it does not influence response trends. Furthermore, the significant variances for year between

Table 1. Nonresponse: 1998–2015.
Null-model for response, noncontact, and refusal logits over surveys and countries

Fixed part	Response logit	Noncontact logit	Refusals logit
Intercept	1.07 (.21)	− 2.60 (.22)	− 1.91 (.31)
Random part (Variances)			
Over surveys	0.15 (.09)	0.17 (.12)	0.52 (.32)
Over countries	0.56 (.17)	0.41 (.13)	0.68 (.21)
Residual	0.13 (.01)	0.22 (.01)	0.28 (.02)
N Countries/surveys	25/7	23/7	23/7

Note: Dependent variables are response rate, noncontact rate and refusal rate. Parameter estimates are on a logit scale and are not proportions. Standard errors in parentheses. All estimates are significant at $p < 0.05$.

countries indicate that the trends over time do indeed differ over countries. To investigate this further, a cross-classified multilevel model was specified with slope variation for year at the country level only. The results are summarized in Table 2.

The first column of Table 2 shows that the parameter estimate for 'year' is significant. This means that there is a negative trend in response from year to year for international surveys. The response rate indeed decreases over the period 1998–2015 (Research Question 1). The regression coefficient of -0.03 for the logit of the response in Table 2 translates into a decrease in response rate by approximately 0.59 percentage point for each year. The decrease in response rates over the years is caused by significant increases in both noncontacts (regression coefficient 0.03) and refusals (regression coefficient 0.03).

With the predictor variable 'year' added to the model, the variance over surveys is no longer significant compared to the null model, but the variance over countries still remains significant. An important finding is that the variance of the regression coefficients for year over countries is small but significant. The variance for year over countries for the response rate in Column 1 of Table 2 is 0.001, which indicates small but significant differences in the response trends between countries. The variation in response trends is related to significant variation for year over countries in the trends for both noncontacts and refusals, as shown in Column 2 and 3 of Table 2. In sum: the downward trend for response rates does not differ between different surveys (Research Question 2), but the trends are different between different countries (Research Question 3). Similarly, the increasing trends for noncontact and refusal rate do not differ between surveys, but they do differ between countries.

3.3. *International Response Trends: Country Context, Survey Design and Fieldwork Efforts*

An important finding was that for international surveys there is a downward trend in response and that there is a significant cross-country variation in the response trends.

Table 2. Response trends in the period 1998–2015.
Effects of year on response, noncontact and refusal logits over surveys and countries

Fixed part	**Response logit**	**Noncontact logit**	**Refusal logit**
Intercept	1.38 (.23)	− 2.97 (.25)	− 2.21 (.34)
Year	− 0.03 (.01)	0.03 (.01)	0.03 (.01)
Random part (Variances)			
Over surveys	0.18 (.11)^{ns}	0.23 (.15)^{ns}	0.63 (.38)^{ns}
Over countries	0.59 (.18)	0.55 (.19)	0.40 (.16)
Year over countries	0.001 (.0004)	0.001 (.0005)	0.002 (.0007)
Residual	0.08 (.01)	0.16 (.01)	0.22 (.01)
N countries/surveys	25/7	23/7	23/7

Note: Dependent variables are response rate, noncontact rate and refusal rate. Parameter estimates are on a logit scale and are not proportions. Year is coded 1998=0. Standard errors in parentheses. All estimates are significant at $p<0.05$, unless indicated by *ns*.

Countries do differ in response, noncontact, and refusal trends, which give rise to the question which factors in survey design and fieldwork effort are related to these nonresponse trends (Research Question 4). Three groups of variables were available as explanatory variables to include in the trend models:

1. *Context Variable.* On the country level, a latent class indicator describing economic development was created (see Subsection 2.2.1.). Although not under the control of a research institute or researcher, the economic conditions of a country may influence response propensity (De Leeuw and De Heer 2002; Groves and Couper 1998; Harris-Kojetin and Tucker 1999; Larsen et al. 2020, this issue).
2. *General Survey Design Variables.* Whether or not a survey is mandatory may reduce refusals and enhance general response (De Leeuw and De Heer 2002). Type of sample may influence response in several ways; a household/address sample is expected to have lower noncontact rates than a person sample, but it may have lower general response rates (Stoop et al. 2010, chap. 2). Likewise, under- and oversampling of certain groups may influence response. Rules for allowing proxy respondents and for substitution of noncontacts and refusals differ for different surveys. In general, it is expected that a more lenient approach in design and allowing for proxies and for substitution enhances response (Stoop 2016; Vehovar 1999). Multiple, or mixed-mode, studies are at present used as a means to improve response at affordable costs, but the empirical evidence of their effect on response is still scarce (De Leeuw 2018). Representativity, or R-indicators, have been introduced as a quality indicator of response and are used at Statistical Institutes to compare the response to different surveys (Bethlehem et al. 2011). Although, these representativity indices do not directly influence response, they may support better monitoring and targeting nonresponse (Schouten et al. 2011). Finally, allowing for interventions in fieldwork and adapting to the situation may improve response (Chun et al. 2018; Tourangeau et al. 2016).
3. *Effort in Fieldwork Variables.* The position and working conditions of interviewers may play an important role in attained response. De Heer (1999) already pointed out that employment condition of interviewers plays a role; furthermore, motivating interviewers through a reward or sanction unwanted interviewer behaviour also may influence (non)response (Groves and Couper 1998). Refusal conversion and the use of special 'refusal' letters to motivate initial nonrespondents appear to have a beneficent effect on response; for an overview see AAPOR (2014). The limited number of studies on reassignment of initial nonrespondents to new interviewers shows conflicting evidence. For an overview, see Peeters et al. 2020, this issue. Finally, meta-analyses and overviews show that the use of respondent incentives has a positive effect on response (e.g., Singer 2002; Singer and Ye 2013). Adding these variables as predictors to the model leads to the results summarized in Table 3. As some agencies omitted ineligibles in their response rate calculation, we also included a control variable (RRcorrected) indicating whether the RR reported in the questionnaire included this correction.

Although 19 time-varying variables describing country context, survey design, and fieldwork effort were added to the model, and several of these variables are

Table 3. Predicting response, noncontact and refusal logits over surveys and countries: 1998–2015.

Fixed part	**Response logit**	**Noncontact logit**	**Refusal logit**
Intercept	0.60 (.66)^{ns}	−2.88 (.78)	−0.26 (.99)^{ns}
Year	−0.03 (.01)	0.03 (.01)	0.03 (.01)
Country context variable			
Economic development (high)	−0.26 (.38)^{ns}	0.68 (.47)^{ns}	−0.20 (.50)^{ns}
General design variables			
Mandatory	0.80 (.31)	−1.27 (.27)	−1.11 (.36)
HH/address sample (1 versus person sample 0)	0.90 (.32)	0.05 (.28)^{ns}	−2.05 (.33)
Undersampling	−0.59 (.38)^{ns}	−0.25 (.37)^{ns}	0.65 (.46)^{ns}
Oversampling	0.42 (.13)	−0.13 (.19)^{ns}	−0.52 (.16)
Proxy allowed	−0.32 (.25)^{ns}	0.54 (.35)^{ns}	0.82 (.38)
Subst. noncont	−1.44 (.31)	0.18 (.40)^{ns}	2.23 (.42)
Subst. refusal	1.77 (.34)	−0.17 (.44)^{ns}	−2.47 (.45)
Mixed-mode	−0.11 (.05)	0.17 (.08)^{ns}	0.10 (.07)^{ns}
R indicators used	−0.74 (.33)	1.60 (.42)	−0.54 (.36)^{ns}
Interventions allowed	0.88 (.25)	0.19 (.26)^{ns}	−1.07 (.26)
Effort in fieldwork variables			
Own interviewers (1 versus External 0)	0.24 (.24)^{ns}	−0.11 (.35)^{ns}	−0.30 (.45)^{ns}
Reward interviewers	−0.05 (.33)^{ns}	−0.39 (.38)^{ns}	0.23 (.43)^{ns}
Sanction interviewers	−0.39 (.30)^{ns}	0.02 (.36)^{ns}	0.97 (.42)
Refusal conversion	−0.10 (.13)^{ns}	−0.01 (.19)^{ns}	−0.06 (.16)^{ns}
Reassignment	−0.09 (.16)^{ns}	−0.92 (.21)	0.13 (.20)^{ns}
Special refusal letters	0.38 (.30)^{ns}	0.31 (.41)^{ns}	−0.38 (.44)^{ns}
Incentives	−0.01 (.22)^{ns}	−0.60 (.32)^{ns}	0.85 (.36)
Control variable			
RRcorrected for ineligibles	−0.13 (.21)^{ns}	Not applicable	Not applicable
Random part	**Variance**	**Variance**	**Variance**
Over surveys	0.48 (.33)^{ns}	0.12 (.12)^{ns}	0.64 (.48)^{ns}
Over countries	1.02 (.35)	1.23 (.49)	1.70 (.64)
Year over countries	0.001 (.000)	0.001 (.000)	0.001 (.001)
Residual	0.05 (.00)	0.11 (.01)	0.07 (.00)
N countries/surveys/records	25/7/470	23/7/432	23/7/434

Note: Dependent variables are response rate, noncontact rate and refusal rate. Parameter estimates are on a logit scale and are not proportions. Year is coded 1998=0. All variables are coded 0=no, not applicable, 1=yes, applicable. Standard errors in parentheses. All estimates are significant at $p < 0.05$, unless indicated by ns.

significantly related to nonresponse, the nonresponse trends over years retain their significance.

Which factors are related to the nonresponse trend? (Research Question 4). The patterns of significant time-varying predictors of response, noncontacts and refusals clearly show that mainly the general design variables are related to the nonresponse trends. Disappointingly, the fieldwork effort variables are hardly related to nonresponse, and the socio-economic development of the countries is not related to any of the nonresponse trends.

The significant predictors of response rate are mostly related to refusals. With the available variables, the overall response rate is associated with the design variables 'mandatory', 'type of sample (person vs household/address)', 'oversampling', 'substitutions allowed', 'use of more than one mode (mixed-mode)', 'representativity (R)-indicators used', and 'interventions in fieldwork allowed'. No effort variables were significantly associated with response rate.

Refusal rate can be predicted better than noncontact rate. Refusal rate is again associated with the design variables 'mandatory', 'type of sample', 'oversampling', 'substitutions allowed', and 'intervening in fieldwork'. The design variable 'proxies allowed', and the effort variable 'incentive for respondents' are also associated with refusal rate, but not with overall response rate.

Finally, noncontact rate is associated with the design variables 'mandatory', 'R-indicators used', and the fieldwork effort variable 'reassignment of cases to special interviewers'.

It is not surprising that mandatory surveys result in higher response rates, due to both lower noncontacts and lower refusals; however, this variable is not under the control of an individual researcher. When households are sampled, this results in a higher response rate than when persons are sampled, due to lower refusals. Similarly, oversampling leads to a higher response, again due to lower refusals. Interestingly, allowing substitution for noncontacts does not influence noncontact rate, but is associated with an increase in refusals and, as a result, a decrease in response rate. Allowing substitutions for refusals and allowing interventions in the field leads to increased response rates and decreased refusal rates.

It should be noted that the results reported above in Table 3 are descriptive, and that correlation alone does not imply causality. This is illustrated by two counterintuitive results. When an agency uses a mixed-mode design or R-indicators, this is associated with a lower response rate. This does not mean that the use of R-indicators causes lower response rates. It is far more likely that when agencies encounter decreasing response rates, they want to know if it affects the representativity and check this with R-indicators. Similarly, adopting a mixed-mode design may be a reaction to growing nonresponse. We come back to this in the discussion.

3.4. Nonresponse and Costs

Because only a small number of countries reported on costs, we could not include this variable in our analyses of response trends. However, as costs are an important topic in the survey literature (e.g., Groves 1989), we include a brief description of cost trend below.

Six countries (Belgium, Hungary, Norway, Portugal, Sweden and the United States) provided a trend report of raw costs per case in their own currency. This was corrected for inflation over the years and subsequently indexed as follows: the value reported for the first year (1998) was set at 100 for each country, thereby resolving differences in currencies. The dependent cost variable thus represents indexed cost *per case*, corrected for inflation. All countries reporting on costs report on the Labour Force Survey.

A multilevel analysis showed that this costs variable has no significant variance over surveys or countries ($p > .10$), therefore a single level analysis suffices. This analysis shows that cost per case does increase over the years ($r = 0.39$, $p = .001$). A follow-up

correlational analysis showed that for these six countries, the cost variable correlates negatively with noncontact ($r = -0.48$, $p < .001$), but does not show a statistically significant correlation with refusal ($r = 0.02$, $p = .89$) or overall response ($r = -0.02$, $p = .89$). Thus, increasing cost is associated with lower noncontact rates only.

Costs can be seen as a proxy for effort, and the results above are in line with the conclusions of Beullens et al. (2018), who remark that in the European Social Survey many countries have increased their fieldwork efforts, and thus costs, in order to prevent response rates from falling. They report that increasing effort seem to have an effect on noncontact rates, but have been less effective in reducing refusal rates.

4. Main Conclusions and Discussion

In his review on adjustment techniques, Brick (2013) concluded that, although a lot of progress has been made in many areas of nonresponse research, the '. . . central problem, in our opinion, is that even after decades of research on nonresponse we remain woefully ignorant of the causes of nonresponse at a profound level'. In this study we tried to shed some light in the darkness by collecting, besides nonresponse data, also information on design and fieldwork variables.

We found that for a variety of surveys, response has been steadily declining internationally over the period 1998–2015. This is the result of an increase in both noncontacts and refusals over time (Research Question 1).

An interesting finding is that there are no differences in response trends between different surveys, so the downward trends are generalizable across different surveys (Research Question 2). Although the Labour Force Survey did have an overall higher response rate and lower refusal rate than the social surveys, the downward *trends* did not differ between the surveys. The higher response rate of the LFS is not surprising as it is a mandatory survey in many countries, while other social surveys are far less often mandatory.

While there are no differences in nonresponse trends between surveys in the period 1998–2015, there are differences in nonresponse trends between different countries (Research Question 3). This is in line with the recent findings for the European Social Survey by Beullens et al. (2018), for the Labour Force Survey (De Leeuw et al., 2018), and by Williams and Brick (2018) for the United States. In conclusion, there is ample empirical evidence that response rates continue to decline internationally, and that these trends differ between countries.

When focusing on these different trends between countries (Research Question 4), a more complicated picture emerges. The contextual variable differences in economic development of the individual countries are not related to response trends. Perhaps differences in economic development are less important than more subjective indicators, such as survey climate and individual attitudes towards surveys (Loosveldt and Joye 2016; De Leeuw et al. 2019). Unfortunately, this information was not available; in a future version of the nonresponse questionnaire, an assessment of survey climate could be added.

In general, refusal and response rates are better explained by the variables at our disposal than noncontact rates. A striking result is that none of the fieldwork effort variables is related to overall response, while most of the general design variables are related to response trends. A potential explanation may be the nature of the data structure.

The data were collected at the survey level, and therefore do not contain information on individual contact attempts, about interviewer and respondent characteristics, and potential interactions. To obtain better insight in the latter, contact forms or process information over time are necessary. Unfortunately, this information is typically not available for most surveys and countries over a long period.

The majority of the design variables show relationships with response in the expected directions. For instance, mandatory surveys produce higher response rates than voluntary surveys. Allowing interventions in fieldwork procedures also leads to a higher response, a fact that may be related to the use of responsive or adaptive designs. Finally, using a mixed-mode design is related to a slightly lower response rate. This is understandable, since a single mode face-to-face survey is known to have the highest response rates (e.g., Bethlehem et al. 2011; Hox and De Leeuw 1994; Stoop 2005). In addition, much depends on the implementation of mixed-mode surveys, especially when online surveys are part of the mix (Dillman 2017; De Leeuw 2018).

However, several variables appear to have counterintuitive relationships. For example, the usage of R-indicators is related to lower response rates. It should be noted that decisions on survey design and fieldwork are often a reaction to declining response rates. In this case, when agencies experience decreasing response rates, the question is if this affects representativity, which can be checked with R-indicators. Another counterintuitive result is that none of the fieldwork effort variables is related to overall response. In the literature, there is ample evidence from experimental studies that show that providing incentives and converting refusals are effective in nonresponse reduction (AAPOR 2014; Singer 2002; Singer and Ye 2013). In our data, not all detail on fieldwork implementations at survey level was available or had a large amount of missing values. For instance, incentives-used was a binary variable. Information on type, amount, and implementation of incentive is lacking. Furthermore, inspection of the data showed that the lack of effect refusal conversion and incentives was not related to specific surveys, but was also found within surveys. This may indicate that agencies that encounter low or decreasing response rates increase their fieldwork efforts to counteract this effect (cf. Loosveldt 2019).

It should be emphasized that the reported relationships are descriptive (correlational). This means that a causal direction in many cases cannot be clearly established, and the absence of relations with fieldwork characteristics does not imply that fieldwork efforts are not important. This also means that survey practitioners cannot take these results directly in hand to inform on ways to improve their fieldwork designs. To determine causal directions, experimental designs need to be implemented in survey research (Lavrakas et al. 2019), and we recommend that agencies routinely use experimental designs to assess the effects of fieldwork changes.

Beside the descriptive nature of this study, a second limitation concerns the measurement quality. To obtain a long longitudinal time series on response trends, a retrospective questionnaire was sent to representatives of statistical institutes, and the retrospective nature may have influenced the data quality negatively. However, for a limited number of countries and years we were able to check our data. Eurostat has published LFS quality reports online for European countries since 2007. The correlations between equivalent variables in the two sources were extremely high (all correlations $>.95$), which gives us confidence in the quality of our data. Another measurement

problem is the question whether countries use the same definition for noncontacts and refusals. Our instruction to the representatives who completed the questionnaires stressed that the same definition should be used throughout the reporting period. Nevertheless, different definitions between countries are possible. To the extent that countries did use different definitions, this would result in a higher country level error, and thus decrease the explained variance at the country level.

Finally, although we report evidence of increased expenditure over time, the cost data are limited to six countries and three surveys. Costs per case are increasing, and higher costs are related to lower noncontact rates, but not to fewer refusals or higher response rates. Cost data are in general difficult to obtain. To enhance nonresponse research in the future, we encourage that both data on cost and fieldwork efforts are included in survey methodological reports in more detail.

5. Appendix : Explanatory Variables Used

Description	**Available records**
Country context variable	
Class indicator socio-economic development	535
General design variables	
Survey is mandatory or not	535
Person versus household/address sample	535
Undersampling	535
Oversampling	523
Proxy allowed	535
Substitution allowed for noncontact	535
Substitution allowed for refusal	535
More than one mode used	535
Any representativity (R)indicator used	535
Interventions in fieldwork	535
Effort in fieldwork variables	
Employment status (external interviewers versus own)	504
Reward good performance interviewer	517
Sanction poor performance interviewer	517
Is refusal conversion used	517
Reassignment to special interviewers	535
Use of special letters to refusals	535
Use of respondents incentives	535
Control variables	
Ineligibles controlled for in response rate	535

Note: Variables were screened in a preliminary analysis on missingness and variance. Some variables could not be used in the final analysis, because of too many missing data or almost no variance. For example: the effort variables interviewer monitoring and use of advance letters are not used because of almost no variance (almost everybody does it); effort variables number of visits and calls could not be used because of too much missing information. All binary variables were recoded into 0=no, not applicable, 1=yes, applicable.

6. References

AAPOR (American Association of Public Opinion Research) Task Force on Survey Refusals. 2014. *Current Knowledge and Considerations Regarding Survey Refusals*. Available at: http://www.aapor.org/AAPOR_Main/media/MainSiteFiles/RefusalTF_FINAL090814.pdf (accessed 20 January 2020).

AAPOR: American Association of Public Opinion Research. 2016. *Standard Definitions, Final Dispositions of Case Codes and Outcome Rates for Surveys*. Available at: https://www.aapor.org/AAPOR_Main/media/publications/Standard-Definitions20169theditionfinal.pdf (accessed 28 April 2020).

Atrostic, B.K., N. Bates, G. Burt, and A. Silberstein. 2001. "Nonresponse in US Government Household Surveys: Consistent Measures, Recent Trends, and New Insights." *Journal of Official Statistics* 17: 209–226. Available at: https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/nonresponse-in-u.s.-government-household-surveys-consistent-measures-recent-trends-and-new-insights.pdf (accessed 28 April 2020).

Bethlehem, J., F. Cobben, and B. Schouten. 2011. *Handbook of Nonresponse in Household Surveys*. New York: Wiley. DOI: https://doi.org/10.1002/9780470891056.

Beullens, K., G. Loosveldt, C. Vandenplas, and I. Stoop. 2018. "Response Rates in The European Social Survey: Increasing, Decreasing, or a Matter of Fieldwork Efforts?" *Survey Methods: Insights from the Field.* DOI: https://doi.org/10.13094/SMIF-2018-00003.

Biemer, P.P. and L.E. Lyberg. 2003. *Introduction to Survey Quality*. New York: Wiley. DOI: https://doi.org/10.1002/0471458740.

Bradburn, N. 1992. "A Response to the Nonresponse Problem: Presidential Address AAPOR." *Public Opinion Quarterly* 56: 391–387. DOI: https://doi.org/10.1093/poq/56.3.391.

Brehm, J. 1993. *The Phantom Respondents: Opinion Surveys and Political Representation.* Ann Arbor: The University of Michigan Press. DOI: https://doi.org/10.3998/mpub.9690285.

Brick, J.M. 2013. "Unit Nonresponse and Weighting Adjustments: A Critical Review." *Journal of Official Statistics* 29: 329–352. DOI: https://doi.org/10.2478/jos-2013-0026.

Chun, A.Y., S.G. Heeringa, and B. Schouten. 2018. "Responsive and Adaptive Design for Survey Optimization." *Journal of Official Statistics* 34: 581–597. DOI: https://doi.org/10.2478/jos-2018-0028.

Curtin, R., S. Presser, and E. Singer. 2005. "Changes in Telephone Survey Nonresponse Over the Past Quarter Century." *Public Opinion Quarterly* 69: 87–98. DOI: https://doi.org/10.1093/poq/nfi002.

De Heer, W. 1999. "International Response Trends: Results of An International Survey." *Journal of Official Statistics* 15: 129–142. Available at: https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/international-response-trends-results-of-an-international-survey.pdf (accessed 14 February 2020).

De Leeuw, E.D. 2018. "Mixed-Mode: Past, Present, and Future". *Survey Research Methods,* 17: 75–89. Available at: https://ojs.ub.uni-konstanz.de/srm/article/view/7402/6582 (accessed 20 January 2020).

De Leeuw, E., and W. De Heer. 2002. "Trends in Household Survey Nonresponse: A Longitudinal and International Comparison." In *Survey Nonresponse*, Edited by R.M. Groves, D.A. Dillman, J.L. Eltinge, R.J.A. Little, 41–54. New York: Wiley.

De Leeuw, E., J. Hox, and A. Luiten. 2018. "International Nonresponse Trends Across Countries and Years: An Analysis of 36 Years of Labor Force Survey Data." *Survey Methods: Insights from the Field*. Available at: https://surveyinsights.org/?p = 10452 (accessed 28 April 2020).

De Leeuw, E., J. Hox, H. Silber, B. Struminskaya, and C. Vis. 2019. "Development of An International Survey Attitude Scale: Measurement Equivalance, Reliability, and Predictive Validity." *Measurement Instruments for the Social Sciences,*1:9. DOI: https://doi.org/10.1186/s42409-019-0012-x.

Dillman, D.A. 2017. "The Promise and Challenges of Pushing Respondents to the Web in Mixed-Mode Surveys." *Survey Methodology.* Statistics Canada, Catalogue No. 12–001–X, Vol. 43, No. 1. Available at: http://www.statcan.gc.ca/pub/12-001-x/2017001/article/14836-eng.htm (accessed 5 May 2020).

Dillman, D.A. 1978. *Mail and Telephone Surveys; the Total Design Method.* New York: Wiley.

Dillman, D.A., J. L. Eltinge, R.M. Groves, and R.J.A. Little. 2002. "Survey nonresponse in design, data collection, and analysis". In *Survey Nonresponse*, Edited by R.M. Groves, D.A. Dillman, J.L. Eltinge, and R.J.A. Little: 3–26. New York: Wiley.

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

Dutwin D. and P.J. Lavrakas. 2017. *Trends in Telephone Outcomes*, Appendix D to the Future of US General Population Telephone Survey Research AAPOR Task Force Report 2017. Available at: https://www.aapor.org/Education-Resources/Reports/The-Future-Of-U-S-General-Population-Telephone-Sur.aspx?utm_source=link_newsv9&utm_campaign_item_225143&utm_medium=copy (accessed 14 February 2020).

Groves, R. 1989. *Survey Errors and Survey Costs*. New York: Wiley. DOI: https://doi.org/10.1002/0471725277.

Groves, R.M. and M.P. Couper. 1998. *Nonresponse in Household Survey Interviews*. New York: Wiley. DOI: https://doi.org/10.1002/9781118490082.

Groves, R.M., F.J. Fowler, M.P. Couper, J.M. Lepkowski, E. Singer, and R. Tourangeau. 2004. *Survey Methodology,* Second Edition. New York: Wiley.

Guilford, J.P. and B. Fruchter. 1978. *Fundamental Statistics in Psychology and Education.* New York, London: McGraw-Hill.

Harris-Kojetin, B.A. and C. Tucker. 1999. "Exploring the Relations of Economic and Political Conditions With Refusal Rates to a Government Survey." *Journal of Official Statistics* 15: 167–184. Available at: https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/exploring-the-relation-of-economic-and-political-conditions-with-refusal-rates-to-a-government-survey.pdf (accessed 14 April 2020).

Hox, J.J. and E.D. De Leeuw. 1994. "A Comparison of Nonresponse in Mail, Telephone, and Face-to-face Surveys; Applying Multilevel Models to Meta-Analysis." *Quality and Quantity* 28: 329-344. DOI: https://doi.org/10.1007/BF01097014.

Hox, J., M. Moerbeek, and R. Van De Schoot. 2018. *Multilevel Analysis. Techniques and Applications*. New York, London: Routledge. DOI: https://doi.org/10.4324/9781315650982.

Larsen, L.J, J. Fane Lineback, and B.M. Reist 2020 "Continuing to Explore the Relation between Economic and Political Factors and Government Survey Refusal Rates: 1960–2015". *Journal of Official Statistics, this issue*.

Lavrakas, P.J., M.W. Traugott, C. Kennedy, A.L. Holbrook, E.D. De Leeuw, and B.T. West. 2019. *Experimental Methods in Survey Research. Techniques That Combine Random Sampling With Random Assignment*. New York: Wiley. DOI: https://doi.org/10.1002/9781119083771.

Loosveldt, G. 2019. *Valedictory Address*. Leuven: KU Leuven, November 15, 2019. Available at: https://soc.kuleuven.be/fsw/nieuws/emeritaat-geert-loosveldt (accessed June 2020).

Loosveldt, G. and D. Joye. 2016. "Defining and Assessing Survey Climate." In *The Sage Handbook of Survey Methodology,* Edited by C. Wolf, D. Joye, T.W. Smith, and Y-C Fu, 67-76. Los Angeles: Sage. DOI: https://doi.org/10.4135/9781473957893.n27.

Luiten, A., E. de Leeuw, B. Schouten, and J. Hox. 2016. "First Results of the (new) International Questionnaire on Nonresponse: Response of the LFS." Paper Presented at International Workshop of Household Survey Nonresponse, Oslo, Norway, 31-8/1-9, 2016. Available at: https://www.nonresponse.org/uploadi/editor/DnD148714714898539LuiteneaInternationalQuestionnaire.docx (accessed June 2020).

Marken, S. 2018. *Still Listening: The State of Telephone Surveys*. Gallup. Available at: https://news.gallup.com/opinion/methodology/225143/listening-state-telephone-surveys.aspx (accessed 14 July 2019).

Muthén, L.K. and B.O. Muthén. 2017. *Mplus User's Guide*. Eight Edition. Los Angeles, CA: Muthén & Muthén.

Pew Research. 2012. *Assessing the Representativeness of Public Opinion Surveys*. Available at: https://www.people-press.org/2012/05/15/assessing-the-representativeness-of-public-opinion-surveys/ (accessed 28 April 2020).

Schouten, B., N. Shlomo, and C. Skinner. 2011. "Indicators for Monitoring and Improving Representativeness of Response." *Journal of Official Statistics*. Available at: https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/indicators-for-monitoring-and-improving-representativeness-of-response.pdf (accessed 28 April 2020).

Singer, E. 2002. "The Use of Incentives to Reduce Nonresponse in Household Surveys." In *Survey Nonresponse*, Edited by R.M. Groves, D.A. Dillman, J.L. Eltinge, and R.J.A. Little, 163–178. New York: Wiley.

Singer, E. and C. Ye. 2013. "The Use and Effects of Incentives in Surveys". *The Annals of the American Academy of Political and Social Science* 645: 112–141. DOI: https://doi.org/10.1177/0002716212458082.

Steeh, C. 1981. "Trends in Nonresponse Rates, 1952-1979." *Public Opinion Quarterly* 45: 40–57. DOI: https://doi.org/10.1086/268633.

Stoop, I. 2005. *The Hunt for the Last Respondent*. The Hague: Social and Cultural Planning Office. Available at: https://www.scp.nl/english/Publications/Publications_-

by_year/Publications_2005/The_Hunt_for_the_Last_Respondent (accessed 14 February 2020).

Stoop, I. 2016. "Unit Nonresponse". In *The Sage Handbook of Survey Methodology,* Edited by C. Wolf, D. Joye, T.W. Smith, and Y-C Fu, 409-424. Los Angeles: Sage. DOI: https://doi.org/10.4135/9781473957893.n27.

Stoop, I., J. Billiet, A. Koch, and R. Fitzgerald. 2010. *Improving Survey Response: Lessons Learned from the European Social Survey.* Chichester: Wiley. DOI: https://doi.org/10.1002/9780470688335.

Tabachnick, B.G., and L. Fidell. 2013. *Using Multivariate Statistics*. Boston, MA: Pearson.

Tourangeau, R., J.M. Brick, S. Lohr, and J. Li. 2016. "Adaptive and Responsive Survey Designs: A Review and Assessment." *Journal of the Royal Statistical Society, Series A*. 180: 203 223. DOI: https://doi.org/10.1111/rssa.12186.

Vehovar, V. 1999. "Field Substitution and Unit Nonresponse". *Journal of Official Statistics,* 15: 333–550. Available at: https://www.scb.se/contentassets/ca21efb41-fee47d293bbee5bf7be7fb3/field-substitution-and-unit-nonresponse.pdf (accessed 28 April 2020).

Williams, D. and J.M. Brick. 2018. "Trends in US Face-to-Face Household Survey Nonresponse and Level of Effort". *Journal of Survey Statistics and Methodology* 6: 186–211. DOI: https://doi.org/10.1093/jssam/smx019.

Received August 2019
Revised February 2020
Accepted May 2020

sciendo

Journal of Official Statistics, Vol. 36, No. 3, 2020, pp. 489–505, http://dx.doi.org/10.2478/JOS-2020-0026

Continuing to Explore the Relation between Economic and Political Factors and Government Survey Refusal Rates: 1960–2015

Luke J. Larsen[1], *Joanna Fane Lineback*[1], *and Benjamin M. Reist*[2]

In the United States, government surveys' refusal rates have been increasing at an alarming rate, despite traditional measures for mitigating nonresponse. Given this phenomenon, now is a good time to revisit the work of Harris-Kojetin and Tucker (1999). In that study, the authors explored the relation between economic and political conditions on Current Population Survey (CPS) refusal rates over the period 1960–1988.

They found evidence that economic and political factors are associated with survey refusals and acknowledged the need to extend this work as more data became available. In this study, our aim was to continue their analysis. First, we replicated their findings. Next, we ran the assumed underlying model on an extended time-period (1960–2015). Last, since we found that the model was not an ideal fit for the extended period, we revised it using available time series and incorporating information about the CPS sample design. In the extended, refined model, presidential approval, census year, number of jobs and not-in-labor-force rate were all significant predictors of survey refusal.

Key words: Refusal rates; response rates; nonresponse; time series.

1. Introduction

Major government survey programs have many tools at their disposal for mitigating survey nonresponse. For example, they have access to high-quality sampling frames for correctly locating potential respondents and access to staff with expertise in converting nonrespondents. Additionally, they may be able to offer multiple reporting modes, offer monetary incentives, or extend data collection. However, recently in the United States, despite access to such tools, nonresponse rates – in particular refusal rates – have been dramatically increasing for unidentified reasons (see Subsection 1.1).

Here, we explore the recent increase by continuing the work of Harris-Kojetin and Tucker (1999), which focused on potential macro-level factors of survey refusal. In that study, the

[1] U.S. Census Bureau, 4600 Silver Hill Rd., Washington, DC, 20233, U.S.A. Emails: Luke.J.Larsen@census.gov and Joanna.Fane.Lineback@census.gov

[2] NASA Headquarters, 300 E St. SW Washington, DC, 20546, U.S.A. Email: Benjamin.Reist@nasa.gov

Acknowledgments: Thanks to Isaac Dorfman for his help analyzing the CPS Contact History Instrument data. Thanks to Kevin Younes for his help collecting relevant literature. Thanks to Timothy Gilbert for providing the historical National Crime Victimization Surveys refusal rates. Thanks to Shane Ball for proofreading the article. Thanks to the U.S. Department of Agriculture's National Agricultural Statistics Service and the U.S. Census Bureau for supporting this research. Finally, thanks to our reviewers for their helpful feedback. The findings and conclusions in this publication are those of the authors and should not be construed to represent any official U.S. Department of Agriculture, U.S. Census Bureau, NASA or U.S. Government determination or policy.

authors used Current Population Survey (CPS) data and a time-series regression approach to examine economic and political influences (unemployment rate, presidential approval rating, inflation rate, consumer sentiment score, a census year indicator, and a March supplement indicator (see online Supplemental material, Appendix A (Data Sources)), for more information on these series) on CPS refusal rates over the period 1960–1988. The authors hypothesized that they would find evidence that environmental factors have an influence on the decision to participate in the CPS and suggested that a negative attitude about the government and a weak economy might decrease the likelihood of survey participation. However, they found that negative feelings about the government were associated with decreased survey participation, but that weak economic times were associated with increased survey participation.

Now, 20 years later, we replicate their findings and extend their model to the period 1960–2015. We also refine their model using available predictors and information about the CPS design. We expect to find the same relation between economic and political factors and survey refusal rates.

1.1. Increasing Refusal Rates in Major Government Surveys

Refusal rates in major government surveys in the United States have been increasing at an alarming rate, and they have been the main driver of increasing nonresponse rates. Increases in refusal rates are not unique to the United States; De Heer (1999) and De Leeuw and Luiten (2018) reported that refusal rates have been increasing since 1980 in many countries. To exemplify historical refusal rate patterns, below are plots of refusal rates over time for three such large-scale surveys: the CPS, the National Crime Victimization Survey (NCVS), and the National Health Interview Survey (NHIS). These surveys are conducted by the Census Bureau on behalf of the U.S. Bureau of Labor Statistics (BLS), the U.S. Bureau of Justice Statistics, and the National Center for Health Statistics, respectively. These surveys cover very different, but potentially sensitive subject matter: income, crime, and health, respectively. Each is primarily an in-person survey that has maintained a relatively stable design over an extended period, making the time series easy to interpret.

For interpreting the plots, it is important to understand the anatomy of these surveys' response rates. Households that were eligible for the survey but were not interviewed for some reason are referred to as Type A noninterviews. Each month, the Type A noninterview rate is calculated by dividing the total number of Type A households (refusals, temporarily absent, noncontacts, and other noninterviews) by the total number of eligible households, which includes Type A households and interviewed households. It follows that the refusal rate is the ratio of the total number of refusals to the total number of eligible households.

The CPS (U.S. Census Bureau and U.S. Bureau of Labor Statistics 2006) is the primary source of labor force statistics in the United States. Data are collected monthly in an electronic format by interviewers through in-person visits and telephone calls. As shown in Figure 1, the percentage of CPS refusals relative to the number of eligible sampled cases has been increasing over most of the 56-year period, 1960–2015. Around 1994, there was a sudden increase in refusals, as well as an increase in variability, that coincided with changes to data collection methods, including the introduction of computer-assisted in-person interviews (CAPI). Around 2010, the percentage of refusals began to increase

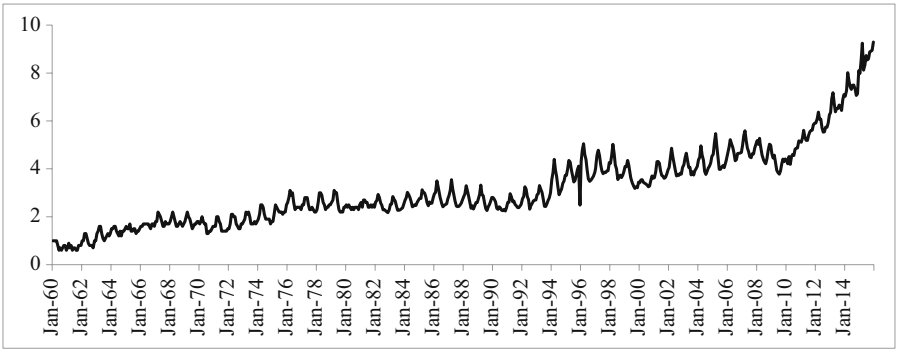

Fig. 1. CPS refusal rate by month: 1960–2015.
Source: U.S. Census Bureau, Current Population Survey, January 1960–December 2015 (unweighted).

sharply, doubling from around 4% in January 2010 to 8% in March 2014 with a high of 9.3% by the end of the study period in December 2015. As of this writing, CPS refusal rates have continued to increase, reaching a new high of 12.39% in February 2018.

Noncontacts are likely of interest to many readers, but the noncontact portion of the Type A rate was unavailable for much of the period being studied, so we can only comment on the non-refusal portion of the Type A rate. Harris-Kojetin and Tucker (1999) observed that, "Through the 1960s and 1970s, the non-refusal portion of the nonresponse rate (chiefly reflecting noncontacts) decreased at approximately the same rate as the refusal rate increased. However, when the refusal rate stabilized in the 1980s, the rate for other types of nonresponse did also." From 1989–2010, the ratio of CPS refusals to total Type A noninterviews hovered around 0.6, with the exception of the period 1999–2001 when refusals decreased as a percentage of Type A noninterviews (see Figure 2). Since 2010, refusals have increased as a percentage of total Type A noninterviews. Over the same period, the non-refusal portion of the Type A rate was increasing, but at a much slower rate.

The NCVS (NCVS 2017) is the primary source of crime victimization statistics in the United States. Data are collected in-person and by phone. The NCVS plot of the average yearly refusal rate since 1992 (shown in Figure 3) is strikingly similar to the CPS plot

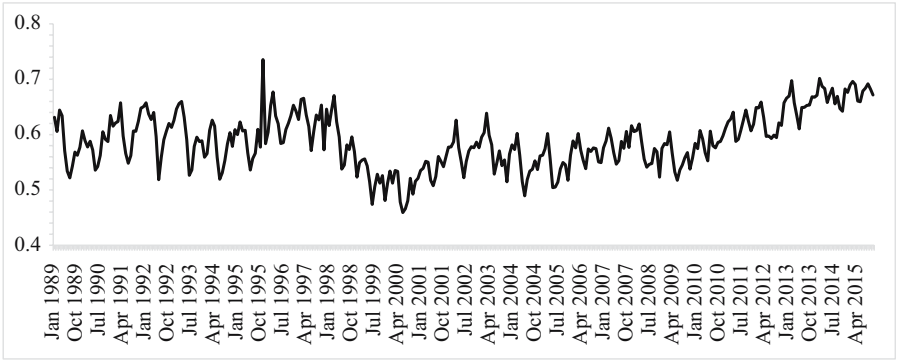

Fig. 2. Ratio of CPS refusals to total type A noninterviews (refusals, temporarily absent, noncontacts, and other noninterviews) by month: 1989–2015.
Source: U.S. Census Bureau, Current Population Survey, January 1989–December 2015 (unweighted).

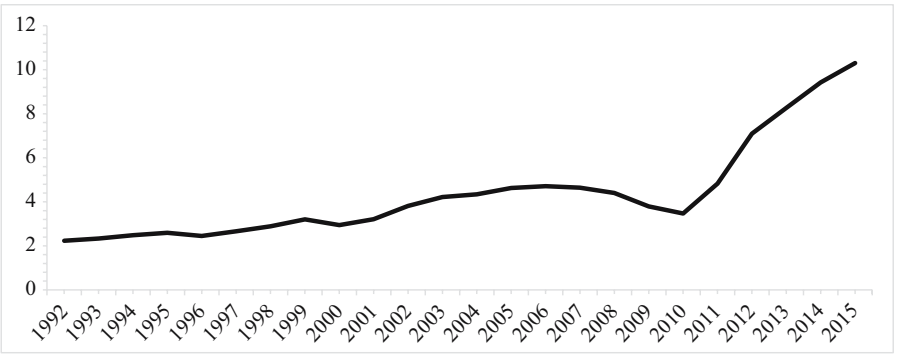

Fig. 3. NCVS average refusal rate by year: 1992–2015.
Source: U.S. Census Bureau, National Crime Victimization Survey, 1992–2015 (unweighted).

(Figure 1) over the same period. Specifically, refusal rates steadily increased until around 2010, when they began dramatically increasing.

The NHIS (National Center for Health Statistics 2016) is the main source of health statistics for the civilian, non-institutionalized population of the United States. Data are collected through in-person, household interviews. Like the CPS and the NCVS, NHIS refusal rates have seen an exponential increase in refusals over the period 2010–2015 with no signs of slowing down (see Figure 4), although NHIS refusal rates had already reached current CPS and NCVS levels around the time their refusal rates started their dramatic increase.

1.2. Changes in Data Collection Methods as a Possible Reason for Increased Refusal

From Figures 1, 3 and 4, one might wonder if the increase in refusals over time is due, at least in part, to changes in data collection methods. For instance, did working cases harder in the field lead to more contacts and ultimately more refusals?

While we do not have data before 2005 to help us answer this question, we do know that since the 1950s, the CPS has undergone regular questionnaire, sample design, estimation,

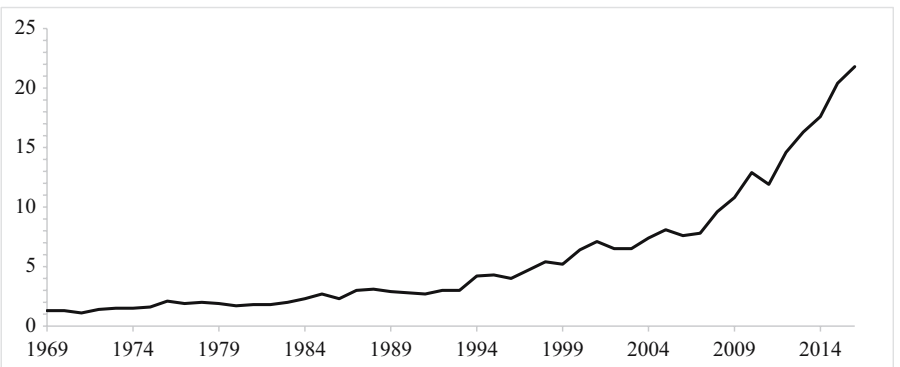

Fig. 4. NHIS refusal rate by year: 1969–2016.
Sources: 1969–2011, Gindi (2012); 2012–2016, National Center for Health Statistics, National Health Interview Survey.

and procedural data collection changes. As much as possible, these changes were planned to limit the amount of disruption to the survey. However, in many ways CPS data collection efforts have stayed the same over the years. It is still primarily an in-person survey, conducted monthly over a ten-day period, with cost-saving strategies built into field operations (such as limitations on number of contacts).

From the period 1960–1994, the Type A noninterview rate remained stable, although there were underlying changes in refusal and noncontact rates. Perhaps the most noteworthy procedural change happened in 1994, when the CPS began testing overlapping computer-assisted telephone interviewing (CATI) and CAPI. Around the same time, a new questionnaire and a sample redesign were introduced, and there was a noticeable increase in Type A noninterview rates. In late 1995 and early 1996, there was a disruption to data collection due to a government shutdown and another increase in Type A noninterview rates (U.S. Census Bureau and U.S. Bureau of Labor Statistics 2006). During 2011 and 2012, there was a restructuring of headquarters and field operations at the U.S. Census Bureau. Headquarters staff were realigned from survey-based to function-based units, with new survey directors managing each of the household surveys. Six of 12 regional offices were closed and the management structure of field operations changed. This was the first major restructuring of field operations in 50 years. Schafer (2014) found that there was no significant impact on response rates – at least for the NCVS – that could be attributed to the field restructuring.

Starting around 2005, the CPS, NCVS, NHIS, and other major government surveys conducted by the U.S. Census Bureau began collecting paradata that help analysts investigate whether changes in field efforts may have led to changes in refusal and noncontact rates. U.S. Census Bureau phone and field interviewers record information about contact attempts, including contact type, contact status, contact strategy, and reluctance-to-respond reason. From a cursory examination of these data for the CPS, which included an examination of the number of contacts and the distribution of reluctance reasons over time, there was also no evidence that data collection changes have contributed, at least since 2005, to changes in Type A noninterview rates.

1.3. Theoretical Background

Without evidence that data collection changes or any major event was the catalyst for the recent increase in refusal rates, we turn to the work of Harris-Kojetin and Tucker (1999) on large-scale factors of survey refusal. The authors ground their work in the research on an individual's decision to participate in a survey. They point out that an individual "can have well-founded rationales for not cooperating with a survey request that may be based on costs and benefits of responding." This is consistent with many of the theoretical frameworks for understanding the response process: for example, social exchange theory (Dillman et al. 2014), benefit cost theory (Singer 2011) and leverage-saliency (Groves et al. 2000). Harris-Kojetin and Tucker (1999) reason that the social, political, and economic environment may influence an individual's estimation of the costs and benefits of responding to a survey.

Based on this theoretical perspective, Harris-Kojetin and Tucker (1999) hypothesized that an individual's decision to respond to a government survey is, at least in part, related to his or her attitude about the government. They go on to propose that an individual's general

feelings about the government can be captured by their approval of the chief executive. In addition, they surmised that, since the government seeks to manage the economy, measures of the nation's economic health may also reflect an individual's feelings about the government, further influencing one's decision to participate in a government survey.

Harris-Kojetin and Tucker's hypothesis was only partially supported by their model. The model showed that presidential approval has an inverse relation with refusal rates. On the other hand, the model showed that economic strength has a direct relation with refusal rates. In other words, the refusal rate decreased during weak economic periods and increased during strong economic periods.

We propose two alternative hypotheses for the relation between refusal rates and health of the economy. Note first that during weak economic times, more individuals in the United States rely on social programs (e.g., food stamps, unemployment insurance, Medicaid). A majority of these programs are facilitated – if not directly administered – by the federal government. We suggest that this increased interaction with the government increases an individual's estimation of the value of the government, and thus the benefit of responding to a government survey. However, neither we nor Harris-Kojetin and Tucker attempt to measure refusal at the individual level. Instead, as this is not practical, we used aggregate measures that reflect the broader survey climate. On the whole, the climate and contributing factors will influence some people's decisions more than others.

Our second hypothesis is that since the CPS is primarily a labor survey, the decision to participate is more salient during an economic downturn. The awareness and benefit of the unemployment rate maybe more apparent to the general individual because of the increased media coverage of the jobs report and the increased emphasis on the unemployment rate by politicians and policy makers. If this hypothesis is true, it would suggest that the relation between economic health and refusal rates might not generalize other types of surveys, such as health or crime surveys. However, pursuit of these hypotheses is outside the scope of this particular work, so we leave the topic to future research.

2. Methodology

This section details the three stages of this project, which included replicating Harris-Kojetin and Tucker's (1999) original findings, extending their model through 2015, and refining the model using additional data sources and information about the CPS sample design. In addition to outlining the methodology, this section discusses the rationale behind the use of alternative time series.

2.1. Replicating Original Findings

Harris-Kojetin and Tucker (1999) fit monthly CPS refusal rates (from January 1960 to December 1988) to a time series regression model using a select set of monthly time series data as regressors. (Hereon, we refer to this as the H-KT model.) In time series regression, the error term of the model is assumed to be decomposable into autocorrelated error that can be modeled with (1) autoregressive moving average (ARMA) model terms and (2) uncorrelated error that is assumed to be normally distributed with mean 0 and variance σ^2 (Ostrom 1978). The aim is that, once the autocorrelated error in the model has been

controlled for, the response variable – in this case, the CPS refusal rate – can be fitted with predictor variables via typical multivariate linear regression techniques.

Most of the regressor series – CPS refusal rate, U.S. presidential approval rate, U.S. unemployment rate, and the Index of Consumer Sentiment – were differenced at both the first-order and seasonal-first-order, while the others – annual percent change of the 1982-basis consumer price index for urban consumers (CPI-U) inflation index and indicators for decennial census year and March CPS supplement month – were not differenced. (Throughout this article, we refer to the dual operations of first-order differencing followed by seasonal-first-order differencing as twice-differencing.) Harris-Kojetin and Tucker published the resulting model's coefficient estimates and the corresponding statistical significance for each model regressor, but the model's autocorrelated error structure, which takes the form of a seasonal autoregressive integrated moving average (SARIMA) model, $(p, d, q) \times (P, D, Q)_{12}$, was not identified.

In the first stage of data analysis, we attempted to replicate the original results using the same data sources over the same period. The exclusion of the autocorrelated error structure from the original article made it challenging to replicate the results of the original study exactly. Our work-around was to employ a brute-force technique to systematically fit the data under a wide variety of assumptions about the true error structure. For instance, the regression was first attempted with assumed error structure $(1,1,0) \times (0,1,0)_{12}$, then again under $(2,1,0) \times (0,1,0)_{12}$, and so forth. All data preparation and model fitting activities were conducted in R, with the SARIMA() wrapper handling the time series regression and residual diagnostics for over 200 variations of seasonal ARMA assumptions. The SARIMA() wrapper is part of the *astsa* package, which was produced for use with the textbook, *Time Series Analysis and Its Applications* (Shumway and Stoffer 2011). When adjusting the parameters of the autocorrelated error, the difference and seasonal difference parameters (d and D, respectively) were both fixed at 1, such that the first-order and seasonal-first-order differences were always in effect. However, the autoregressive (AR), moving average (MA), seasonal AR, and seasonal MA parameters (p, q, P, and Q, respectively) were allowed to vary between 0 and 3 to produce models under the various error structure assumptions.

Models in which the residuals were unstable or had significant autocorrelations were discarded from consideration; among those that remained, fit statistics – in particular, the corrected Akaike information criterion (AICc) – were used to assess good model fit, while the coefficient estimates were compared for accuracy against the "gold standard" coefficient estimates that were published in the original paper. Ultimately, one model was determined to have the best fit, while optimally minimizing the differences between the coefficient estimates and the coefficients of the original model.

2.2. *Extending the H-KT Model Through 2015*

In the second stage of data analysis, we applied the final model selected from the replication effort, including the finalized autocorrelation error structure, to an extended timeframe, January 1960 to December 2015. This particular model did not fit the new series of refusal rates very well, so we applied the brute-force procedure described previously to the extended timeframe to see if a different error structure might be more appropriate. This

exercise resulted in a model with minimal differences in coefficient estimates and a different error structure that yielded acceptably uncorrelated residuals. However, the overall model fit was not as substantial over the full 55-year period, relative to the model fit of the original 28-year period. This outcome appeared to indicate that the original model design might not be appropriate for the more recent timeframe of 1988 to 2015.

2.3. *Reevaluating the H-KT Model Construction*

In the third stage of data analysis, we attempted to refine the H-KT model in order to obtain a better fit than that afforded by the second stage of analysis. To start, we vetted data sources that were available now that may not have been available during the original study and considered changes to the structure of the original model. We only considered including series that were comparable and available across time. Unfortunately, it was difficult to find a monthly or even quarterly time series in the social or political realms for the entire timeframe, so we ultimately focused only on new economic predictors.

In the end, the following additional regressor candidates were considered for inclusion in the model: U.S. quarterly gross domestic product (GDP), U.S. not-in-labor-force rate, number of U.S. jobs, raw CPI-Uinflation index, Standard and Poor's (S&P) 500 index end-of-month value, and party composition in the U.S. Congress. (For more information about each series, refer to online Supplemental material, Appendix A. The level of S&P 500 was chosen as a measure of wealth effect on refusal rates. The GDP was chosen as a measure of the general health of the economy. The number of jobs added was chosen as an alternative to the unemployment rate. The not-in-labor-force rate was chosen as a proxy for discouraged workers, as well as a measure of saliency of the labor force survey, since a labor force survey may not be salient to people not in the labor force. The series that was considered, but ultimately not included in the model, is congressional makeup (the ratio of Republicans to Democrats in each chamber of Congress), as congressional makeup stays relatively constant across consecutive months, which does not lend itself to the twice-differencing technique used in this analysis.

The final set of new and old regressor candidates were considered together for model inclusion on the basis of their pairwise correlations, excluding some to minimize multicollinearity concerns. All the original regressors were log-transformed prior to differencing, and the response series (CPS refusal rate) received a small constant addition prior to the log transformation in order to resolve some stationarity issues in the 1960s. The brute-force, trial-by-error approach to selecting an autocorrelated error structure was eschewed in favor of a more mindful strategy that incorporated information about the CPS sample design.

The monthly CPS sample design follows a rotating-panel structure, in which participating housing units are in sample for four consecutive months, then leave the sample for the next eight months, and then return to the sample for the following four months before leaving the CPS completely. For any given survey month, the CPS sample is comprised of members from each of eight different panels (U.S. Census Bureau and U.S. Bureau of Labor Statistics 2006). The rotating-panel structure results in significant correlation among estimates derived from monthly CPS files that are specific lags apart due to the sharing of some participating households between the two files. Any two

consecutive monthly CPS files share about 75 percent of their samples by design, whereas any two CPS files that are a year apart (such as January 2000 and January 2001) share about 50% of their samples. From this structure, one can demonstrate that most CPS-based time series have significant correlation by as many as 15 lags apart. Keeping in mind that the intended response series to be modeled is actually a twice-differenced version of the monthly CPS refusal rates, the autocorrelation structure should be more complex in order to properly account for the presence of shared households between any two-point estimates of the twice-differenced series. Given the lag with significant correlation among the twice-differenced series can be as wide as 28 months, we chose a SARIMA model of $(28,1,0) \times (0,1,0)_{12}$ as the basis for the autocorrelated error of the time series regression.

3. Results

This section details the findings for each stage of this project: replicating Harris-Kojetin and Tucker's (1999) original findings, extending their model through 2015, and refining this model using alternative data sources and information about the CPS sample design.

3.1. Replicating Original Findings

After obtaining the CPS refusal rates series and the four regressor series used in the time series regression model showcased in Harris-Kojetin and Tucker (1999), some additional data preparation had to be applied prior to running the brute-force SARIMA modeling routine in R. Recall that the CPS refusal rates, presidential approval rates, unemployment rates, and consumer sentiment indices were all twice-differenced prior to modeling. In SARIMA(), all series in the model received the same differencing treatment, including those that we did not intend to difference (inflation rate, Census year indicator, and March supplement indicator). To counter this action, we applied "reverse twice-differencing" to the inflation rate series and two indicators using the diffinv() function in R. This allowed us to create a set of time series that could be twice-differenced within SARIMA() to get back to the original time series, while also applying the same differencing to the desired regressors and CPS refusal rates.

During this first stage of analysis (the primary work of which was completed between September 2015 and May 2016), we were aware that the "4-8-4" sample design employed by the CPS would be a key feature in determining which SARIMA parameters should be considered for the brute-force routine. It was known that the CPS in a given month shares a portion of its sampled housing units with other CPS sample months by up to 3 lags (months) forward and backward within a 12-month season and by up to 3 lags forward and backward after the first seasonal lag. Therefore, we decided that the brute-force routine would cycle through first-order lag parameters from 0 to 3 and seasonal lag parameters from 0 to 3. It was unknown whether the original model was strictly AR, MA, or some combination of the two types of ARMA terms, so we allowed the modeling routine to cycle both the AR- and MA-type error parameters (as well as their seasonal counterparts). From this design, the routine produced $4^4 = 256$ regression models to assess.

Each time series regression model produced by SARIMA() yielded model convergence status, coefficient estimates and corresponding standard errors, model fit statistics, and residual diagnostics charts. The residual diagnostics included the standardized residual

Table 1. Autocorrelation selection process.

Rank	Criterion	Condition	Action
1	Model convergence	Model does not converge	Eliminate model from consideration
2	Residual diagnostics	Residuals indicate autocorrelation or non-normality	Favor models with no or few residual issues
3	Coefficient estimates	Coefficient estimates not close to H-KT coefficient estimates	Favor models with estimates close to H-KT results
4	Model fit statistics	Lower AICc scores indicate better fit	Favor models with lowest AICc scores

plot over time, residual ACF plot, normal Q-Q plot of standardized residuals, and p-value plot of the Ljung-Box statistic over lags (in months). These charts were used to assess whether the standardized residuals are generally heteroscedastic, approximately normal, and contain little-to-no autocorrelation. Table 1 ranks the importance of each criterion in determining which model provided the best fit. Under this ranking, it is clear that convergence and residual behavior are critical elements in determining favorable model structures, while coefficient estimates and model fit statistics are important only when the critical elements are satisfactory. Of all 256 models under consideration, only 94 were viable choices in that they satisfied the critical convergence and residual behavior criteria. From these 94 options, the remaining criteria were assessed to make a selection of the autocorrelated error structure that yielded the best model fit.

Table 2 shows the original coefficient estimates (Harris-Kojetin and Tucker 1999) alongside those we obtained after determining the best-fit autocorrelated error structure: $(3,1,1) \times (2,1,1)_{12}$. One can observe that the point estimates were strikingly close to those of the original model, and yet two of the factors had differences in terms of statistical significance: Presidential approval rating was no longer significant under the replication effort, while the March supplement indicator was now significant. Nevertheless, the model based on $(3,1,1) \times (2,1,1)_{12}$ was convergent, featured acceptable residual behavior, and had a low AICc score of -524.90; therefore, we determined that this model structure

Table 2. H-KT versus replication model results.

	H-KT model: error structure unknown		Replication model: $(3,1,1) \times (2,1,1)_{12}$	
Predictor	Estimate	Std. Error	Estimate	Std. Error
Presidential approval (D)	−0.0026**	0.0011	−0.0013	0.0012
Inflation rate	0.0000	0.0000	−0.0004	0.0002
Unemployment rate (D)	−0.0590**	0.0180	−0.0540**	0.0201
Consumer sentiment (D)	0.0042**	0.0016	0.0043*	0.0020
Decennial year	0.0084	0.0047	0.0095	0.0196
March supplement	0.0120	0.0073	0.0112*	0.0046

Sources 1960–1988: Harris-Kojetin and Tucker (1999); U.S. Census Bureau; U.S. Bureau of Labor Statistics; University of Michigan; Gallup. All series are based on data from January 1960 to December 1988. (D) indicates a differenced time series. $N = 348$. *$p < 0.05$, **$p < 0.01$.

satisfactorily replicated the efforts of Harris-Kojetin and Tucker (1999) to fit the 1960–1988 monthly CPS refusal rate series to the featured set of predictors.

3.2. Extending the H-KT Model Through 2015

Next, we investigated whether the replication model identified in the previous section could adequately fit the refusal rates when the series was expanded to include monthly data up to December 2015. These expanded series contained several clear features, such as spikes in presidential approval following the 9/11 attacks and the 2008 presidential election, a sharp increase in unemployment rates during the Great Recession, and the relatively flat annual inflation rate since the 1990s. However, the CPS refusal rates after 1988 were particularly notable for a sustained growth trend until 2010, when the refusal rate series began a sharp increase, approaching 10% by the end of 2015. (Refer to online Supplemental material, Appendix B, for the plots of the expanded series for presidential approval, the inflation rate, the unemployment rate, and the index of consumer sentiment. Refer to Table 1 for the expanded CPS series.)

Because the trends in the original 1960–1988 window appear to be different from the trends in the 1989–2015 window, we compared pairwise correlations between the variables of interest for the entire 1960–2015 span with the correlations of the original span (Online Supplemental material, Appendix C). Most of the correlations were similar between the two efforts in terms of approximate magnitude, direction, and statistical significance. However, there were notable differences as well. For instance, the correlation between inflation and CPS refusal rate shifted direction significantly from 0.46 to −0.30, while the correlation between presidential approval and consumer sentiment shifted direction significantly from −0.60 to 0.48.

We ran the "best fit" model from Subsection 3.1 on the data from 1960–2015. Residual analysis did not indicate any problems, but the statistical significance of model coefficients for all factors except unemployment rate had shifted when comparing the 1960–1988 period to the 1960–2015 period (see Table 3).

Table 3. "Best fit" model parameters: 1960–1988 versus 1960–2015.

	1960 – 1988 $(3,1,1) \times (2,1,1)_{12}$		1960 – 2015 $(3,1,1) \times (2,1,1)_{12}$	
Predictor	Estimate	Std. error	Estimate	Std. error
Presidential approval (D)	−0.0013	0.0012	−0.0024*	0.0012
Inflation rate	−0.0004	0.0002	−0.0007**	0.0002
Unemployment rate (D)	−0.0540**	0.0201	−0.0714**	0.0222
Consumer sentiment (D)	0.0043*	0.0020	0.0033	0.0020
Decennial year	0.0095	0.0196	0.0350**	0.0082
March supplement	0.0112*	0.0046	0.0051	0.0038

Sources 1960–1988: U.S. Census Bureau; U.S. Bureau of Labor Statistics; University of Michigan; Gallup. All series are based on data from January 1960 to December 1988. (D) indicates a differenced time series. $N = 348$. $*p < 0.05$, $**p < 0.01$.

Sources 1960–2015: U.S. Census Bureau; U.S. Bureau of Labor Statistics; University of Michigan; Gallup. All series are based on data from January 1960 to December 2015. (D) indicates a differenced time series. $N = 672$. $*p < 0.05$, $**p < 0.01$.

Table 4. "Best fit" model parameters: 1960–1988 versus 1989–2015.

Predictor	1960 – 1988 $(3,1,1) \times (2,1,1)_{12}$		1989 – 2015 $(2,1,2) \times (0,1,1)_{12}$	
	Estimate	Std. error	Estimate	Std. error
Presidential approval (D)	−0.0013	0.0012	−0.0040	0.0024
Inflation rate	−0.0004	0.0002	−0.0015	0.0008
Unemployment rate (D)	−0.0540**	0.0201	−0.0768	0.0477
Consumer sentiment (D)	0.0043*	0.0020	0.0053	0.0039
Decennial year	0.0095	0.0196	0.0532**	0.0157
March supplement	0.0112*	0.0046	−0.0054	0.0084

Source 1960–1988: U.S. Census Bureau; U.S. Bureau of Labor Statistics; University of Michigan; Gallup. All series are based on data from January 1960 to December 1988. (D) indicates a differenced time series. $N = 348$. $*p < 0.05$, $**p < 0.01$.
Sources 1989–2015: U.S. Census Bureau; U.S. Bureau of Labor Statistics; University of Michigan; Gallup. All series are based on data from January 1989 to December 2015. (D) indicates a differenced time series. $N = 672$. $*p < 0.05$, $**p < 0.01$.

Analysis of the AICc between the two regression attempts yielded an interesting comparison: Under the 1960–1988 period, the AICc was −524.90, while under the 1960–2015 period, the AICc was −692.88. Because the 1960–1988 and 1989–2015 periods contain roughly the same amount of monthly data points (29 years for the former, 27 years for the latter), one might expect that the magnitude of the AIC for the combined periods would be approximately double that of the original period. Yet, that was not the case, which led us to suspect that the $(3,1,1) \times (2,1,1)_{12}$ error structure may not yield the best model fit for the more recent period. Another run of pairwise correlations, this time exclusively upon the 1989–2015 period, further corroborates this notion (see online Supplemental material Table 2 of Appendix C). Compared with the correlations from the 1960–1988 period, nearly half of the 15 pairwise correlations differed from the earlier period in magnitude, direction, or statistical significance.

Next, we returned to the brute-force modeling strategy from Subsection 3.1 to determine whether a different error structure might yield improved model fit for the 1989–2015 period. In this effort, the "best fit" model (see Table 4) featured an error structure of $(3,1,1) \times (2,1,1)_{12}$ with AICc $= -187.98$; though the replication model did have a better AICc of −190.87, it failed the residual assessment and was ineligible for further consideration. This is still less than half the magnitude of the AICc for the 1960–1988 replication model (AICc $= -524.90$), so the pursuit of obtaining a better fit to the more recent CPS refusal rate data may have to venture beyond the core construction of this particular model. To drive this point further, note that all but one of the coefficient estimates for the "best" model in the 1989–2015 window were not statistically significant. With only the decennial census year indicator having a significant effect upon predicted refusal rates, this clearly is not a very useful outcome.

3.3. Reevaluating the H-KT Model Construction

After determining that the original model could be reconstructed to fit the 1960–1988 data, but subsequently finding that process did not yield a comparably good fit to the

expanded 1960–2015 data, we attempted a number of modifications in pursuit of a better model fit. First, we applied a log-transformation to the CPS refusal rate series – as well as to each of the regressors in the model – prior to the twice-differencing step in an attempt to improve the stationarity of the series before fitting the regression model. Also, because the raw refusal rates in the 1960s were so low, we found that we could further improve stationarity about that time period by adding a small constant to the entire raw refusal rate series prior to the log transform.

Next, we prepared a handful of other economic variables to be candidates for regressors in the new model (detailed in online Supplemental material, Appendix A). As mentioned previously, the current number of U.S. jobs regressor from the Current Employment Statistics by BLS was introduced to replace the CPS-based unemployment rate series in the model, while the raw inflation index (CPI-U, 1982 basis) was used to replace the 12-month percent change in the same index. The three additional regressors – not-in-labor-force rate from CPS, quarterly U.S. GDP, and end-of-month closing price of the S&P 500 – serve to provide additional dimensions of U.S. economic health that may be relevant to potential CPS respondents in determining their willingness to participate in government surveys. As with the other regressors, these five variables were log-transformed and twice-differenced prior to their inclusion in the time series regression model.

Pairwise correlations among the CPS refusal rate and the expanded set of regressors for the 1960–2015 window were analyzed (see online Supplemental material, Table 3 of Appendix C). Notably, all the regressors were significantly correlated with the refusal rate, while many of the regressor pairs had strong correlations between them (aside from those involving the census year and March supplement indicators). This is in line with findings from the correlational analysis done for Subsections 3.1 and 3.2. One should keep in mind that a few of these correlations are exceptionally strong, indicating that there may be a risk of overspecification in the model.

Finally, we reevaluated the manner of selection for the autocorrelated error structure. Since we were no longer trying to replicate the H-KT model results by guessing the error

Table 5. New model using expanded set of regressors: 1960–2015.

Regressor	Coefficient estimate	Standard error
Presidential approval (LD)	−0.0171*	(0.0104)
Consumer sentiment (LD)	0.0322	(0.0302)
Decennial year	0.0054**	(0.0018)
March supplement	−0.0001	(0.0019)
Number of jobs (LD)	0.9510**	(0.3345)
Inflation (LD)	−0.1851	(0.2687)
Not in labor force (LD)	0.5029*	(0.2160)
U.S. GDP (LD)	0.1133	(0.2017)
S&P 500 (LD)	−0.0040	(0.0231)

Sources: U.S. Census Bureau; U.S. Bureau of Labor Statistics; U.S. Bureau of Economic Analysis; University of Michigan; Gallup; Standard and Poor's. All series are based on data from January 1960 to December 2015. Results shown are for log-differenced CPS refusal rates.
(LD) indicates a log-differenced series (first order and seasonal first order). $N = 672$. $*p < 0.10$, $**p < 0.01$.

Table 6. New model fit: 1960–1888 versus 1989–2015.

	Early era (1960–1988)		Recent era (1989–2015)	
	Replicated model	Refined model	Replicated model	Refined model
Model fit (AICc)	−524.90	−1364.72	−187.98	−1324.75
Significant regressors	Unemployment, consumer sentiment, march supplement	Jobs, NILF rate, consumer sentiment	decennial year	Jobs, U.S. GDP, decennial year

Sources: U.S. Census Bureau; U.S. Bureau of Labor Statistics; U.S. Bureau of Economic Analysis; University of Michigan; Gallup; Standard and Poor's. All series are based on data from January 1960 to December 2015. $N = 348$ for 1960–1988. $N = 322$ for 1989–2015.

structure used in that study, the "brute force" iterative method used in Subsections 3.1 and 3.2 was not appropriate. Instead, we applied information about the CPS sample design to make reasonable assumptions about the autocorrelation present in the CPS refusal rate series and subsequently, the transformed refusal rate series to be fit by the time series regression model. Ultimately, we determined that the error structure for this model should be $(28,1,0) \times (0,1,0)_{12}$. (See Subsection 2.3 for more details.)

With these changes to specifications in place, the refined model was convergent and yielded satisfactory residual diagnostics. Table 5 shows the coefficient estimates of the expanded regressor set under this new model. We expected to see more evidence that positive feelings towards politicians are associated with a decrease in refusal rates and positive feelings about the economy are associated with an increase in refusal rates. In fact, four of the series are statistically significant predictors of CPS refusal rates – presidential approval and census year indicator from the original set of predictors, and number of jobs and not-in-labor-force status from the new set of predictors. From these results, increases in the number of U.S. jobs and the share of the population that is not in the labor force were predictive of increases in refusal rates. Being in a decennial census year was also linked to higher CPS refusal rates. However, increases in presidential approval were predictive of lower refusal rates. Compared with the Subsection 3.2 results, there was not a notable change in the point estimates of the coefficients – differences in statistical significance aside – but a comparison of the model fit statistics was particularly interesting.

Under the final model decided upon in Subsection 3.2, we found that the AICc was about −693, but under the new model described here, the AICc was about −2799 – about four times greater in magnitude. Note that the log transformation is the most likely driver of this difference in AICc values, so that difference in of itself is not an indicator of improved model fit between this effort and that of the replicated model in Subsection 3.2. However, the reader may also recall that one of the problems with the replicated model was that the model fit from the "recent era" (1989–2015) was not as good as the model fit from the "early era" (1960–1988) – the relevant AICc statistics were −188 and −524, respectively (Table 6). To see how the newer model shown here might compare, we re-ran the model for the two shorter timeframes and found that the AICc statistics between the two were roughly the same: −1365 for the early years and

− 1325 for the recent years. This finding indicates that the refined model fits the refusal rates series about equally well in either of the shorter timeframes – a substantial improvement over the previous effort.

4. Discussion

Harris-Kojetin and Tucker (1999) initially considered the effect of large-scale political and economic factors on survey refusal rates using CPS refusal rates and relevant predictors over the period 1960–1988. They proposed that negative feelings about politics and and a weak economy would be associated with an increase in refusal rates. They found that disapproval of the president was associated with an increase in refusal rates, but a weak economy was associated with a decrease in refusal rates. With the rapid increase in government surveys' refusal rates over the past decade, it seemed like the ideal time to replicate and extend the work by Harris-Kojetin and Tucker (1999).

First, we replicated the results from the original H-KT model using similar time series methods and the same set of predictors (unemployment rate, presidential approval rating, inflation rate, consumer sentiment score, census year indicator, and March supplement indicator). We also found that presidential approval and unemployment rate were both negatively associated with refusal, while consumer sentiment was positively associated with refusal.

Next, we extended this model to the period 1960–2015, but found that the model did not extend well to the period 1989–2015. After refitting the model, the statistical significance of all model factors except unemployment rate changed from the original to the longer period. Presidential approval, inflation, and decennial year were all significant factors in the extended model, while consumer sentiment and March supplement month were no longer significant. These results may, in part, reflect that the original model was developed for a much more stable period of refusal rates. However, given the poor model fit, we have little confidence in these results.

Last, we refined the model using a modified set of predictors (presidential approval rating, consumer sentiment score, census year indicator, March supplement indicator, number of jobs, inflation rate, not in labor force rate, GDP, and the S&P 500 index). We achieved increased model fit over the original model. Increases in presidential approval were associated with lower CPS refusal rates, while U.S. jobs, the percentage of the population not-in-labor-force, and decennial year were all associated with higher refusal rates. It might not be obvious that strong economic times would lead to increased refusal, but if one considers that somehow people may feel less connected to the government during a strong economy, then this is reasonable. This result makes even more sense in the context of a labor force survey, such as the CPS.

It is important to underscore that these results may not be generalizable. The focus of this study was a United States labor force survey. The results may not extend to other countries. Within the United States, the results may not generalize to non-government surveys, which have very different response rates, and they may not even generalize to other government surveys.

Along these lines, a logical next step would be to replicate this analysis for other government surveys. The methodology of surveys like the NHIS and NCVS has

stayed relatively stable for many years, giving us additional time series to study. At the same time, we should continue to take a closer look at the theory on survey nonresponse and collect or otherwise obtain measures that will help us understand more about the social aspect of the social-political-economic construct that is missing from these analyses.

In sum, we explored the recent increase in government surveys' refusal rates by continuing the work of Harris-Kojetin and Tucker (1999), which focused on potential macro-level factors of survey refusal. We refined and extended their model, and showed that presidential approval, census year, number of jobs and not-in-labor-force rate were all significant predictors of CPS refusal. While this model does not explain the changes in refusal rates, it can be used as a tool for monitoring possible causes of survey refusal over time. And while the recent spike in refusal rates is alarming, the good news for surveys like the CPS, NHIS, and NCVS is that overall response rates are still high. Government surveys, at least in the United States, still see response rates that far surpass response rates of most non-government surveys.

5. References

De Heer, W. 1999. "International Response Trends: Results of an International Survey." *Journal of Official Statistics* 15(2): 129–142. Available at: https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/international-response-trends-results-of-an-international-survey.pdf (accessed may 2020).

De Leeuw, E., J. Hox, and A. Luiten. 2018. "International Nonresponse Trends across Countries and Years: An analysis of 36 years of Labour Force Survey data." *Survey Insights: Methods from the Field*. DOI: http://doi.org/10.13094/SMIF-2018-00008.

Dillman, D.A., J.D. Smyth, and L.M. Christian. 2014. *Internet, Phone, Mail, and Mixed Mode Surveys: The tailored design method (4th ed.)*. Hoboken, NJ, US: John Wiley & Sons Inc.

Gindi, R.M. 2012. "Responsive Design on the National Health Interview Survey: Opportunities and challenges." Paper presented at FCSM Research Conference, Washington, DC, US. December 4th 2012. Available at: https://slideplayer.com/slide/7836641/ (accessed August 2018).

Groves, R., E. Singer, and A. Corning. 2000. "Leverage-Saliency Theory of Survey Participation: Description and an illustration." *Public Opinion Quarterly* 64(3): 299–308. DOI: https://doi.org/10.1086/317990.

Harris-Kojetin, B. and C. Tucker. 1999. "Exploring the Relation of Economic and Political Conditions with Refusal Rates to a Government Survey." *Journal of Official Statistics* 15(2): 167–184. Available at: https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/exploring-the-relation-of-economic-and-political-conditions-with-refusal-rates-to-a-government-survey.pdf (accessed May 2020).

National Center for Health Statistics. 2016. "National Health Interview Survey Description." Centers for Disease Control and Prevention, U.S. Department of Health and Human Services. Available at: https://nhis.ipums.org/nhis/resources/srvydesc2016.pdf.

NCVS, National Crime Victimization Survey. 2017. "National Crime Victimization Survey, 2016: Technical Documentation." Bureau of Justice Statistics, U.S. Department of Justice. Available at: https://www.bjs.gov/content/pub/pdf/ncvstd16.pdf (accessed August 2018).

Ostrom, C.W. 1978. *Time Series Analysis: Regression techniques*. Thousand Oaks, CA, US: SAGE Publications Inc.

Schafer, J.L. 2014. "Modeling the Effect of Recent Field Interventions in the National Crime Victimization Survey." *Research Report Series:* Statistics 2014-02, U.S. Census Bureau. Available at: https://www.census.gov/srd/papers/pdf/rrs2014-02.pdf (accessed March 2019).

Shumway, R.H. and D.S. Stoffer. 2011. *Time Series Analysis and Its Applications (3rd ed.)*. New York, NY, US: Springer.

Singer, E. 2011. "Toward a Benefit-Cost Theory of Survey Participation: Evidence, Further Tests, and Implications." *Journal of Official Statistics*, 27(2): 379–392. Available at: https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/-toward-a-benefit%20cost-theory-of-survey-participation-evidence-further-tests-and-implications.pdf (accessed June 2020).

U.S. Census Bureau and U.S. Bureau of Labor Statistics. 2006. "Design and Methodology: Current Population Survey." *Technical Paper:* 66. Available at: https://www.census.gov/prod/2006pubs/tp-66.pdf (accessed March 2019).

Received August 2018
Revised May 2019
Accepted November 2019

sciendo

Journal of Official Statistics, Vol. 36, No. 3, 2020, pp. 507–527, http://dx.doi.org/10.2478/JOS-2020-0027

Evolution of the Initially Recruited SHARE Panel Sample Over the First Six Waves

Sabine Friedel[1] *and Tim Birkenbach*[2]

Attrition is a frequently observed phenomenon in panel studies. The loss of panel members over time can hamper the analysis of panel survey data. Based on data from the Survey of Health, Ageing and Retirement in Europe (SHARE), this study investigates changes in the composition of the initially recruited first-wave sample in a multi-national face-to-face panel survey of an older population over waves. By inspecting retention rates and R-indicators, we found that, despite declining retention rates, the composition of the initially recruited panel sample in Wave 1 remained stable after the second wave. Thus, after the second wave there is no further large decline in representativeness with regard to the first wave sample. Changes in the composition of the sample after the second wave over time were due mainly to mortality-related attrition. Non-mortality-related attrition had a slight effect on the changes in sample composition with regard to birth in survey country, area of residence, education, and social activities. Our study encourages researchers to investigate further the impact of mortality- and non-mortality-related attrition in multi-national surveys of older populations.

Key words: R-indicator; wave nonresponse; mortality- and non-morality-related attrition; panel sample composition.

1. Introduction

Panel surveys of older populations in Europe have become the focus of widespread interest in recent decades. Falling fertility rates (Myrskylä et al. 2013) and greater life expectancy (Leon 2011) bring many challenges for Western European societies. To investigate these dynamic processes, researchers need data that allow them to provide evidence of changes over time (Olsen 2018). In contrast to cross-sectional surveys, panel surveys fulfil this

[1] University of Mannheim, SFB 884 "Political Economy of Reforms", B6, 30-32, 68131 Mannheim, Germany. Email: s.friedel@uni-mannheim.de
[2] Max Planck Institute for Social Law and Social Policy, Munich Center for the Economics of Aging (MEA), Amalienstrasse 33, 80799 Munich, Germany. Email: birkenbach@mea.mpisoc.mpg.de
Acknowledgments: Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 139943784 – SFB 884. This article uses data from SHARE Wave 1 (DOI: https://doi.org/10.6103/SHARE.w1.600). The SHARE data collection has been primarily funded by the European Commission through FP5 (QLK6-CT-2001-00360), FP6 (SHARE-I3: RII-CT-2006-062193, COMPARE: CIT5-CT-2005-028857, SHARELIFE: CIT4-CT-2006-028812) and FP7 (SHARE-PREP: N°211909, SHARE-LEAP: N°227822, SHARE M4: N°261982). Additional funding from the German Ministry of Education and Research, the Max Planck Society for the Advancement of Science, the U.S. National Institute on Aging (U01_AG09740-13S2, P01_AG005842, P01_AG08291, P30_AG12815, R21_AG025169, Y1-AG-4553-01, IAG_BSR06-11, OGHA_04-064, HHSN271201300071C) and from various national funding sources is gratefully acknowledged (see www.share-project.org). The authors would like to gratefully thank Annette Scherpenzeel for her ideas, exceptional support, and feedback, Michael Bergmann, Thorsten Kneip, Peter Lugtig, and Annelies Blom for their advice and feedback, and Thomas Klausch for his advice and expertise on R-indicators.

requirement because they repeatedly collect data from the same respondents over time (Lynn 2009).

However, a major detracting feature of panel surveys is the risk of attrition – that is, the loss of panel members from the initially recruited sample over time (Binder 1998). Panel attrition is a frequent phenomenon that has been observed during the last decades (Fitzgerald et al. 1998; Watson 2003; Buck et al. 2006). Attrition may occur because panel members are no longer able or willing to participate or because they can no longer be located or contacted (Lynn and Lugtig 2017). The largest amount of drop out occurs in the second wave (Watson and Wooden 2009; Schoeni et al. 2013). When attrition occurs, changes over time cannot be observed from the beginning to the end of the panel because one measure is missing in two consecutive waves (Lynn and Lugtig 2017). This absence of data can lead to restrictions when researchers want to analyze changes in the data. Thus, we need to inform researchers about attrition in the data they use.

Particularly in panel surveys of older populations, researchers are faced with a greater risk of attrition due to death. In an investigation of characteristics associated with attrition in the English Longitudinal Study of Ageing (ELSA) and the U.S. Health and Retirement Study (HRS), Banks et al. (2011) found that the mortality rate between two waves among panel members aged 70–80 years was 15%, and that among 55–64 year-old panel members it was 4%. In contrast, for the Panel Study of Income Dynamics (PSID), which is a household panel survey, Watson (2003) reported a mortality rate of only 0.5% between two waves. Thus, the risk of mortality-related attrition is much higher in panel surveys of older populations compared to those that collect data on younger populations.

Deaths in panel surveys of older populations are not problematic per se. Older populations are not fixed, and all older populations are affected by deaths (Smith et al. 2009). Deaths occur both in the population and in the sample, and thus deaths of panel members change the composition of the data sample and of the population about which researchers want to draw conclusions. In both settings – the population and the sample – individuals who have a lower risk of dying, for example because they have a high socioeconomic and health status, are more likely to survive to old age than individuals with a low socioeconomic and health status (Banks et al. 2011). Thus, we assume that mortality in panel surveys of older populations is selective. However, deaths reflect changes in the composition of the population to which the data refer, and, as Smith et al. (2009, 29) noted, "as long as these [deaths] can be identified and distinguished from nonresponse, they are easily incorporated in analyses by using a code for dead units."

In contrast to mortality-related attrition, respondents who drop out for other reasons are still present in the population, and their non-participation changes only the composition of the sample. Changes in these individuals' outcomes of interest can no longer be observed in the survey data, although they are occurring in the population. However, this type of attrition is not problematic per se, either, unless it is selective, and thus can affect the validity and interpretation of estimates (Watson and Wooden 2019).

The present study focuses on the Survey of Health, Ageing and Retirement in Europe (SHARE) (Börsch-Supan et al. 2013), a biennial panel study based on people in Europe aged 50 years and older. With its harmonized collection of data in many European countries, SHARE is unique and offers many opportunities to analyze dynamic processes in the European societies. Although previous research has shown that attrition occurs in the

SHARE panel (Bergmann et al. 2019), little research has investigated in more detail the changes in the composition of the initially recruited panel sample over time (e.g., Bristle et al. 2019). Moreover, little is known about the relation between attrition and the changes in the panel composition over waves when mortality is particularly considered. Both aspects can inform researchers about the impact of attrition on the evolution of the SHARE panel.

To obtain a clear picture of how the composition of the SHARE panel has evolved over waves, we define two samples of interest:

- A: the initially recruited SHARE sample (i.e., the sample first interviewed in Wave 1), and
- B: the initially recruited SHARE sample, excluding respondents who were reported to have died.

Whereas Sample A is fixed over waves and includes all respondents who dropped out, Sample B is dynamic over waves and excludes for each wave separately respondents who were reported to have died before the corresponding wave started. For instance, Sample B in Wave 2 is based on the initially recruited SHARE sample, excluding respondents who were reported to have died before the second wave started, or Sample B in Wave 3 is based on the initially recruited SHARE sample, excluding respondents who were reported to have died before the third wave started. Thus, Sample A investigates total attrition (non-mortality-related and mortality-related), whereas Sample B investigates non-mortality-related attrition only.

With these two definitions of the samples of interest, we aim to answer the following research questions:

1. How has the initially recruited first-wave sample (A and B) evolved over the survey waves?
2. Has the evolution of the initially recruited first-wave sample (A and B) over waves varied across countries?
3. What variables/characteristics have played the most important role in the evolution over waves of the sample that excludes reported deaths (Sample B)?

The remainder of this article is organized as follows: In the next section, we describe our SHARE dataset and the variables considered in our analyses. We then answer Research Questions 1 and 2 by applying two aggregate-level measures (retention rates, R-indicators). In Section 4, we apply two variable-level measures (subgroup retention rates, logistic regressions) to answer Research Question 3. Thus, the methods and results for the first two research questions and the methods and results for the third question are presented separately. The article concludes with a summary of the findings and discussion for all three research questions.

2. Data and Variables

2.1. Data

We used data from the Survey of Health, Ageing and Retirement in Europe (SHARE) (Börsch-Supan 2017). SHARE is a biennial multidisciplinary, cross-national panel survey

that collects microdata on the health, socio-economic status, and social and family networks of individuals aged 50 years and older and of their partners, regardless of their age. The target persons and their partners are interviewed face-to-face using computer-assisted personal interviewing (CAPI) (Börsch-Supan et al. 2013). The first wave of SHARE was conducted in 2004 in eleven European countries and in Israel. Samples from each country are based on a probability sample that is representative of the non-institutionalized population aged 50 years and older (De Luca et al. 2013). The initial individual response rates (RR1, American Association for Public Opinion Research, AAPOR 2016) ranged between 27.9% and 58.8% (Bergmann et al. 2019).

For our analyses, we used the first-wave data about respondents' individual and household characteristics and supplemented these data with information about whether or not the respondents had participated in later waves. We restricted our sample to countries that participated in all six observed waves. This selection criterion reduced the sample to nine countries (Austria, Belgium, Denmark, France, Germany, Italy, Spain, Sweden, and Switzerland). Moreover, we restricted our sample to respondents aged 50 years or older. Together, these restrictions decreased the sample to 21,227 panel respondents (Table 1, Respondents aged 50+). About 5% of the respondents could not be considered because they did not know or refused to report the answer to questions that were used to measure variables included in the analyses. As a consequence, the first analysis sample of Sample A consisted of 20,236 respondents. The sample size by country ranged from 898 in Switzerland to 3,521 in Belgium (see Table 1, Analysis Sample A).

To study further non-mortality-related attrition, we excluded respondents who were reported to have died before a given wave. This exclusion resulted in a dynamic Analysis Sample B (see Table 1, Analysis Sample B, Wave 1–Wave 6). However, the quality of information we used to identify deaths differs between countries. This is due mainly to the fact that most European countries lack a national mortality register or similar records. Therefore, SHARE cannot reliably ascertain the vital status of nonrespondents who drop out because they cannot be located or contacted or because they refuse to be re-interviewed (Bergmann et al. 2019). Thus, the dynamic Analysis Sample B may include unreported deaths.

Table 1. Sample selection of initially recruited first-wave SHARE respondents.

	Respondents aged 50+	Analysis sample A	Analysis sample B					
Country	Wave 1	Wave 1 – 6	Wave 1	Wave 2	Wave 3	Wave 4	Wave 5	Wave 6
Austria	1,516	1,487	1,487	1,442	1,361	1,287	1,213	1,174
Belgium	3,631	3,521	3,521	3,474	3,356	3,237	3,120	3,017
Denmark	1,597	1,527	1,527	1,480	1,390	1,310	1,220	1,134
France	2,955	2,706	2,706	2,650	2,519	2,428	2,298	2,221
Germany	2,909	2,768	2,768	2,718	2,648	2,545	2,508	2,486
Italy	2,495	2,406	2,406	2,353	2,268	2,189	2,081	1,984
Spain	2,232	2,075	2,075	1,984	1,884	1,769	1,655	1,547
Sweden	2,961	2,848	2,848	2,778	2,640	2,486	2,349	2,268
Switzerland	931	898	898	882	860	839	816	788
Total	21,227	20,236	20,236	19,761	18,926	18,090	17,260	16,619

2.2. Variables

Investigating the evolution of the SHARE panel offered the possibility of including a rich set of variables in the models. To examine the evolution of the panel, we selected 23 first-wave key variables from the areas of demographics, social embeddedness, health, and economics, and three survey-specific variables of the questionnaire design (Table 2).

Table 2. Operationalization of information used to examine the evolution of the SHARE panel.

Variable	Operationalization
Demographics	
Gender	0: male; 1: female
Age	1: 50–59 years; 2: 60–69 years; 3: 70–79 years; 4: 80+ years
Born in survey country	1: yes; 0: no
Education level	1: low; 2: medium and other; 3: high
Household (HH) size	1: 1-person HH; 2: 2-person HH; 3: 3+-person HH
Partner in HH	0: no; 1: yes
Area of residence	1: city/large town; 2: small town; 3: rural village
Social embeddedness variables	
Residential proximity of child(ren)	1: no children; 2: child living in household; 3: child living = 1 km away; 4: child living > 1 km away
Social activities	0: no activities; 1: at least one activity
Received help from others	0: no; 1: yes
Gave help to others	0: no; 1: yes
Health variables	
Health status	0: good or better; 1: fair or poor
Chronic diseases	0: none; 1: at least one chronic disease
Depression (Euro-D)	0: no or insufficient symptoms; 1: 4 or more depressive symptoms
Maximum grip strength	1: item nonresponse; 2: 1st quartile; 3: 2nd quartile; 4: 3rd quartile; 5: 4th quartile
Memory recall ability	0: recalled less than half of the words; 1: recalled more than half of the words
Hospital overnight stays in last 12 months	0: no; 1: yes
Currently smoking	0: no; 1: yes
Currently drinking	0: never; 1: less than once a week; 2: 1–6 times a week; 3: daily
Limitation of instrumental activities of daily living (IADL)	0: no IADL limitation; 1: at least one IADL limitation
Economic variables	
Employment status	1: retired; 2 working; 3: not working and other
Make ends meet	0: difficulties; 1: no difficulties
Total household income	1: item nonresponse; 2: 1st quartile; 3: 2nd quartile; 4: 3rd quartile; 5: 4th quartile
Interview process variables	
Financial respondent	0: no; 1: yes
Family respondent	0: no; 1: yes
Household respondent	0: no; 1: yes

When selecting variables to investigate changes in the composition of the initially recruited sample over waves, care was taken to ensure that they represented the main publication domains, related to key survey items, and/or related to survey-specific motives for nonresponse (Schouten et al. 2011).

We included sociodemographic and socioeconomic variables (gender, age, education, citizenship, number of children, and income) in our models. Researchers have used these individual characteristics in almost all models for their substantive analyses based on SHARE data (SHARE-ERIC 2018). Additionally, some of these variables have been found to predict attrition in SHARE (Bristle et al. 2019). As Bristle et al. (2019) showed that item nonresponse to financial questions in SHARE negatively affected cooperation in the next wave, we supplemented the income quartiles with an additional category indicating that respondents did not answer the household income question.

We also included information on household composition, area of residence, employment status, and making ends meet, because this information has been widely used in economic research (SHARE-ERIC 2018) and has been found to predict cooperation in SHARE (Bristle et al. 2019). We included several key health variables that have been extensively used in the literature because researchers have also used SHARE data to study health (SHARE-ERIC 2018). Moreover, research has shown that persons with poor health tend to cooperate less than healthy persons (Bristle et al. 2019). Our selection of health variables included self-assessed health, chronic diseases, depression symptoms (Euro-D), limitations of instrumental activities of daily living (IADL), smoking and drinking behavior, and two objective health measurements/tests (grip strength and recall memory). As SHARE data are also used by researchers in the field of family and social networks, well-being, and charity, we included information on the spatial proximity of children, giving help to others, and receiving help from others. Additionally, as the literature shows that being socially active can predict cooperation in longitudinal studies (Bianchi and Biffignandi 2019), information on the number of social activities was also included.

Furthermore, research has shown that respondent burden in the previous SHARE wave influenced cooperation in the next wave (Bristle et al. 2019). In SHARE, selected household members serve as so-called family, financial, or household respondents and answer specific questions on behalf of the whole household. Being selected for one of these roles means that the duration of the interview is usually longer than average and that the respondent provides more information. To capture this respondent burden, we selected three interview process variables (financial, family, and household respondent).

3. Evolution of the SHARE Panel Sample Over Waves and Across Countries

3.1. Analytical Approach

Addressing Research Questions 1 and 2, we examined changes in the composition of the initially recruited SHARE sample over waves and across countries by calculating retention rates and estimating R-indicators for Analysis Samples A and B (the latter excludes reported deaths before the start of the corresponding wave and potentially includes unreported deaths). To investigate changes in the sample composition over waves, we

coded participation for each wave. We denoted by y_i the outcome for respondent i as follows:

$$y_i = \begin{cases} 0 & no\ participation \\ 1 & participation \end{cases} \quad (1)$$

where participation y_i equals 1 if respondent i participated in the survey and 0 otherwise.

The retention rates in the present study measured the proportion of respondents who participated in each wave, conditional upon having participated in the first wave. The R-indicator (where "R" stands for representativeness) was originally designed to measure the degree to which the respondents in a sample resemble the total target population or gross sample (Schouten et al. 2009). By contrast, the R-indicators in our study measured the degree to which the respondents in Analysis Sample A resemble the initially recruited first-wave respondents over waves, and the degree to which respondents in the dynamic Analysis Sample B resembles the initially recruited first-wave respondents over waves but excluding respondents who were reported to have died before a given wave.

Researchers have used R-indicators to assess the extent to which a net sample is representative of the target population or a gross sample. For instance, data of recruited samples have been compared with census, administrative, or population register data (e.g., Moore et al. 2016; Schouten et al. 2012; Luiten and Schouten 2013; Roberts et al. 2014). R-indicators can also be used as indicators for representativeness in panel studies (Schouten et al. 2012). Bianchi and Biffignandi (2017) used R-indicators to compare the panel sample of the UK household longitudinal study Understanding Society over waves with administrative data to assess population representativeness. In sum, they showed that R-indicators were a valuable measure of representativeness.

R-indicators are estimated as follows (Schouten et al. 2009):

$$\hat{R}_{\hat{\rho}} = 1 - 2\,\hat{S}_{\hat{\rho}}, \quad (2)$$

where $\hat{S}_{\hat{\rho}}$ is the estimated standard deviation of the individual response propensities. Therefore, the R-indicator is a measure of variation in response propensities. The estimated R-indicator $\hat{R}_{\hat{\rho}}$ ranges between 1 and 0, where the value 1 denotes strong representativeness and the value 0 denotes the maximum deviation from strong representativeness.

Our approach differed from that of Schouten et al. (2009) with respect to the meaning of the term "representativeness." Schouten et al. (2009) designed R-indicators to assess the extent to which a net sample is representative of the total target population or a gross sample, whereas we used R-indicators to compare the composition of the initially recruited sample in Wave 1 of SHARE with the composition of the sample in subsequent waves, including any recruitment bias that might have existed in the original sample. The main advantage of our approach was that a rich set of individual-level data could be used rather than the sparse data that are available at population level. For our analyses of the evolution of the panel sample, all information already provided by the participants in the first wave could be used. This approach allowed for the detection of systematic dropout from the panel with respect to many important and substantive survey variables, and not only with respect to a few demographic variables available at the population level.

Thus, we adapted Schouten and colleagues' concept (2009) to examine changes in the composition of the initially recruited SHARE sample over waves. We defined a panel response subset of variables X as "fully representative" if the average propensity to participate again over these categories of X was constant for all possible values of X (Equation 2). For Analysis Sample A, samples in later waves were "fully representative" if their propensities to participate again were equal over the categories of X. As a consequence, the distributions of the selected respondent and household characteristics X remained identical as in the first observed wave. For the dynamic Analysis Sample B, samples in later waves were "fully representative" if their propensities to participate again were equal over the categories of X when reported deaths before a given wave were excluded. As a consequence, the distributions of the selected respondent and household characteristics X remained identical as in the first observed wave excluding reported deaths before a given wave. The estimated R-indicator $\hat{R}_{\hat{\rho}}$ (Equation 2) in our study also ranged between 1 and 0. However, 1 means no change in the composition of the original sample and 0 means total change. Confidence intervals for each R-indicator in each wave were estimated at the five percent level.

The probability that the R-indicators would reach high values differed for our two analysis samples. We expected that the exclusion of reported deaths in the Analysis Samples B would lead to higher R-indicator values for the dynamic Analysis Sample B compared to the fixed Analysis Sample A because we assumed that respondents who dropped out because they died belonged to a selective group of respondents. In contrast, if we had perfect response or if we had equal response propensities over waves, the value of the R-indicator of both analysis samples (A and B) would remain at 1.

To estimate the R-indicators, we used a specially adapted tool provided by the Representative Indicators for Survey Quality Project (RISQ 2015). In more detail, to compute R-indicators, we used a version of Version 2.1 of RISQ that was adapted for our purposes by the RISQ team. RISQ recommends that representativeness be analyzed by using categorical information rather than continuous information, we applied a categorical approach to describe and explore the evolution of the SHARE panel. We fitted several R-indicator models with the 26 selected variables based on participation outcome as the dependent variable. First, we estimated overall R-indicators for all countries (Analysis Sample A). Second, we estimated overall R-indicators that excluded reported deaths before a given wave for all countries (Analysis Sample B) to focus on non-mortality-related attrition. Third, we estimated the R-indicator based on Analysis Sample A and the R-indicator based on the dynamic Analysis Sample B for each country separately.

3.2. Results

To answer the first research question as to how the composition of the initially recruited first-wave sample evolved over waves, we calculated retention rates and estimated R-indicators for each wave, averaged across all countries.

The overall retention rate of Analysis Sample A declined almost linearly over the waves from 69% to 42% (Figure 1), with a kink at the first follow-up interview. Around 30% of the initially recruited first-wave respondents (Analysis Sample A) did not participate in the second wave. Also in the case of the R-indicator (Analysis Sample A), the largest decrease

in the value was observed from the first to the second wave (−.16). However, in contrast to the retention rate, the R-indicator (Analysis Sample A) decreased weakly over time afterwards. After six waves, Analysis Sample A reached an R-indicator value of .72. Thus, after the second wave, no further large decline in representativeness of the initially recruited first-wave sample and only few changes in the sample composition were observed.

Comparing the R-indicator for Analysis Sample A with that for Analysis Sample B, where we excluded reported deaths, we saw that the R-indicators of Analysis Sample B followed the same trend over waves as of Analysis Sample A – a substantial decrease in value after the first wave, and relatively stable values after the second wave. Moreover, we noted that the R-indicators for Analysis Sample B differed significantly from that of Analysis Sample A (see Figure 1). After six waves, the R-indicator for – and thus the representativeness of – Analysis Sample A was .72, whereas the R-indicator for the dynamic Analysis Sample B was .80. Thus, a decline in retention rate is not automatically linked to strong changes in the sample composition. In particular, when we eliminated the selective mortality-related attrition in Analysis Sample B, the representativeness of the sample was reasonably strong.

To answer Research Question 2 as to whether the evolution of the initially recruited sample over waves differed across countries, we calculated retention rates and estimated R-indicators for each country separately. Overall, the same pattern of declining retention rates and stabilizing R-indicators after the second wave was observed (Figure 2). Retention rates in Analysis Sample A ranged from 55% to 75 % across countries in Wave 2 and from 24% to 50% in Wave 6. By contrast, the values of the R-indicators in Wave 2 ranged across countries from .76 to .85 for Analysis Sample A and from .77 to .86 for Analysis Sample B. At the last observed wave (Wave 6), R-indicators ranged across countries from .61 to .74 for Analysis Sample A and from .69 to .85 for Analysis Sample B. Despite the fact that the gap between retention rates and R-indicators varied across countries, the observed pattern

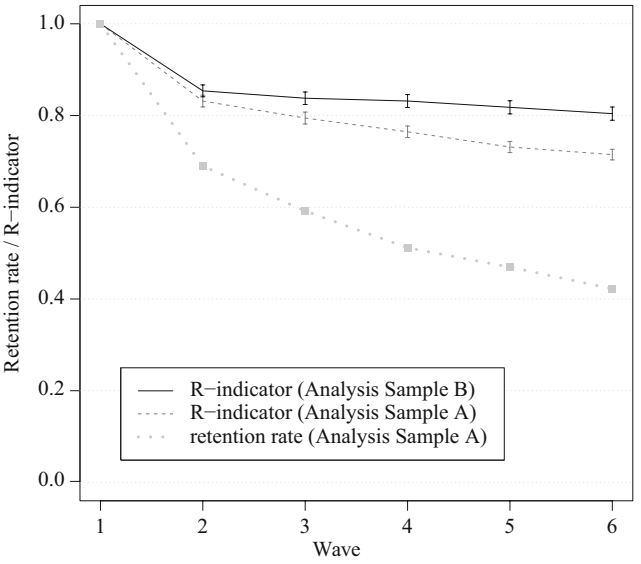

Fig. 1. Evolution of the initially recruited SHARE sample over waves.

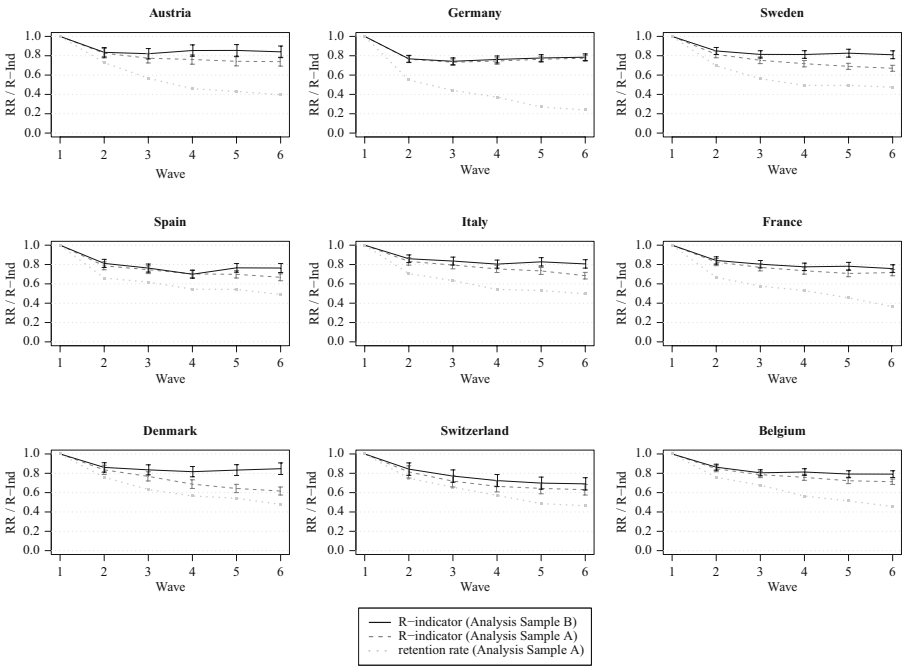

Fig. 2. Evolution of the initially recruited SHARE sample over waves, by country.

of change in the composition of the initially recruited first-wave sample (A and B) measured by R-indicators tended to be similar for all countries.

4. Variable-Level Analysis of Non-Mortality-Related Attrition in SHARE

4.1. Analytical Approach

Research Question 3 aims at understanding non-mortality-related attrition and asked what variables/characteristics played the most important role in the evolution of Analysis Sample B (which excludes reported deaths before a given wave) over waves across all countries. To answer this question, we calculated subgroup retention rates and estimated logistic regression models across all countries.

We defined several attrition scenarios for Research Question 3:

- Scenario 1 (W2): attrition in Wave 2,
- Scenario 2 (W3|W2): attrition in Wave 3, conditional upon participation in Wave 2,
- Scenario 4 (W6): attrition in Wave 6,
- Scenario 5 (W6|W3): attrition in Wave 6, conditional upon participation in Wave 4.

These scenarios will inform researchers about the changes in the composition of the initially recruited first-wave SHARE sample in later waves. For further exploration, we also defined and analyzed a number of other scenarios (see online Supplemental data, Table 1).

We compared subgroup retention rates for the defined scenarios with the first-wave subgroup proportions, excluding reported deaths before the given wave (dynamic Analysis

Sample B). Only deviations of one percentage point or more are reported in the corresponding Figures (see Supplemental data, Figures 1–3), and only deviations of two percentage points or more are discussed in what follows.

In addition to the univariate subgroup retention rates, we explored non-mortality-related attrition within a multivariate framework because multivariate analyses allow several respondent and household characteristics to be taken into account at once. We estimated logit equations to examine which selected key variables have played the most important role in the evolution of the panel for the various selected scenarios. In contrast to the subgroup retention rates, the coding of y_i was reversed intentionally for the multivariate logits. It allows for an interpretation of the results related to attrition rather than participation. Thus, the attrition propensity ρ_i for a panel respondent i is defined as follows:

$$\rho_i(X) = P(y_i = 1 \mid X = x_i). \tag{3}$$

For a respondent $i = 1, \ldots, N$, y_i refers to the binary nonresponse outcome, which equals 1 if panel respondent i dropped out and 0 otherwise. The outcome y_i can be different for each of the six waves; x_i is a vector of the 26 selected SHARE key variables for panel respondent i (Table 2).

As standard coefficients in logistic models indicate only the effect direction and provide no information about effect size, we estimated average marginal effects (AME) to evaluate the logistic regression coefficients more appropriately. AMEs represent the average change in probability when the variable predictor increases by one unit (Mood 2010). Moreover, by examining the z-scores of the logistic regression models we could quantify the impact of the individual and household characteristics on non-mortality-related attrition (Analysis Sample B). This examination deepened the understanding of which variables actually led to a decline of the R-indicators in Subsection 3.2.

4.2. Results

To answer Research Question 3 as to what variables/characteristics played the most important role in the changes in the composition of the initially recruited first-wave sample (Analysis Sample B, which excludes reported deaths before a given wave) over waves, we calculated subgroup retention rates on participation and ran logistic regression models on attrition for the selected scenarios.

4.2.1. Wave 2

In the subgroup retention rates in Wave 2, where respondents who were reported to have died before Wave 2 were excluded, we observed a deviation of two or more percentage points from the initially recruited Analysis Sample B in Wave 1 only for social activity (see online Supplemental data, Figure 1). The share of respondents who were socially active in Wave 1 increased by 2.4 percentage points in Wave 2, whereas the share of those who were not socially active increased by the same amount of percentage points.

The multivariate analyses of the Analysis Sample B in Wave 2, that excludes reported deaths before Wave 2, in Table 3 showed that, after controlling for other respondent and household characteristics, the association of being socially active with not participating in the second wave was statistically significant ($p < .001$; z-score $= -5.11$). The probability

of attrition in Wave 2 decreased by four percentage points if respondents were socially active in Wave 1. However, the significant association of social activity with not participating in Wave 2 was not the strongest association observed. Rather, the strongest association of attrition in Wave 2 was observed with residing in a rural village ($p < .001$; z-score $= -9.12$). The probability of dropping out in Wave 2 was seven percentage points lower for respondents residing in rural villages than for those living in cities or large towns.

Other strong associations with attrition in Wave 2 were found for respondents who had participated in the grip strength test and who had reported their total household income in Wave 1, regardless of the value in measure ($p < 0.001$; z-scores between -3.75 and -5.93). They were less likely to drop out in Wave 2 than respondents who had not provided these measures. The decrease in probability to drop out ranged from five to eight percentage points (Table 3).

The multivariate analyses additionally showed that other numerous individual and household characteristics of Analysis Sample B in Wave 2 were significantly associated with attrition in the second wave (Table 3). The probability to drop out increased significantly with having received help from others, smoking, and having at least reported one limitation in IADL in the first wave. In addition to these positive significant associations with attrition in the second wave, we observed several negative significant associations with attrition in the second wave. Respondents who were between 60 and 69 years old in Wave 1 were less likely to drop out in Wave 2 than respondents who were between 50 and 59 years old in Wave 1. A respondent born in the survey country was less likely to attrite in Wave 2 than a respondent born outside the survey country. Highly educated respondents were less likely to attrite than low educated respondents, and respondents who resided in a small town in Wave 1 had a lower probability to drop out than respondents residing in a city or large town in Wave 1. Having children, among all groups of residential proximity of the child in Wave 1, decreased the probability to drop out in Wave 2 compared to having no children. Moreover, the probability to drop out in Wave 2 decreased significantly at the five percent level with giving help to others, having at least reported to have one chronic disease, having reported at least four depression symptoms, having a larger memory recall ability, and drinking, regardless of the frequency of alcohol consumption in the first wave.

4.1.2. Wave 3

The subgroup retention rates of Analysis Sample B in Wave 3, conditional upon participation in Wave 2, showed no deviations larger than two percentage points from the initially recruited respondents in the first wave when we excluded respondents that that were reported to have died before the third wave. Only one deviation larger than one percentage point was observed from respondents who resided in the city or large town. Their share was 1.1 percentage points lower compared to their share in Wave 1 (result not shown).

Multivariate analyses showed that strong predictors of attrition in Wave 3, conditional upon participation in Wave 2, were: high educational level ($p < .001$; z-score $= -4.98$) compared to a low educational level, social activity in Wave 1 ($p < .001$; z-score $= -3.92$), age between 60 and 69 years in Wave 1 ($p < .001$; z-score $= -3.48$) compared to age between 50 and 59 years in Wave 1, birth in survey country level ($p < .001$;

Table 3. Estimated average marginal effects (AME) from logistic regressions of attrition by individual and household characteristics.

	W2	W3|W2	W6	W6|W3
Gender: male (ref.)				
– female	−.01	−.01	−.04**	−.01
	(−1.15)	(−.46)	(−2.96)	(−.99)
Age: 50–59 years (ref.)				
– 60–69 years	−.02*	−.04***	−.03*	−.00
	(−2.23)	(−3.48)	(−2.38)	(−.38)
– 70–79 years	.00	−.03*	−.01	.01
	(.06)	(−2.11)	(−.72)	(.58)
– 80+ years	−.00	.01	.11***	.10***
	(−.16)	(.81)	(4.78)	(3.42)
Born in survey country: no (ref.)				
– yes	−.05***	−.05***	−.09***	−.07***
	(−3.86)	(−3.38)	(−6.27)	(−4.05)
Education level: low (ref.)				
– medium	−.00	−.02	−.01	−.01
	(−.51)	(−1.81)	(−1.00)	(−.46)
– high	−.04***	−.05***	−.07***	−.05***
	(−4.14)	(−4.98)	(−6.43)	(−4.20)
HH size: 1–person (ref.)				
– 2–person HH	.00	.03	.02	.03
	(.13)	(1.47)	(1.26)	(1.21)
– 3+ person HH	.01	.02	.02	.04
	(.64)	(.92)	(.76)	(1.37)
Partner in HH: no (ref.)				
– yes	.01	.01	.02	−.01
	(.73)	(.48)	(.89)	(−.35)
Area of residence: city/large town (ref.)				
– small town	−.03***	−.03**	−.05***	−.05***
	(−4.26)	(−3.08)	(−5.56)	(−4.44)
– rural village	−.07***	−.03**	−.07***	−.03*
	(−9.12)	(−3.00)	(−6.90)	(−2.54)
Residential proximity of child(ren): no children (ref.)				
– child in HH	−.08***	−.03	−.07***	−.04*
	(−5.61)	(−1.69)	(−4.23)	(−2.33)
– child = 1 km away	−.06***	−.01	−.08***	−.04*
	(−4.39)	(−.72)	(−5.04)	(−2.03)
– child > 1 km away	−.04***	−.01	−.05***	−.02
	(−3.95)	(−1.10)	(−3.86)	(−1.17)
Social activities: no activities (ref.)				
– at least one activity	−.04***	−.03***	−.06***	−.04***
	(−5.11)	(−3.92)	(−7.06)	(−4.12)
Received help from others: no (ref.)				
– yes	.02*	−.02*	−.02	−.02
	(2.08)	(−2.06)	(−1.86)	(−1.66)
Gave help to others: no (ref.)				
– yes	−.02*	−.01	−.01	−.00
	(−2.44)	(−1.86)	(−.74)	(−.06)

Table 3. Continued

	W2	W3|W2	W6	W6|W3
Health status: good/better (ref.)				
– poor or fair	.01	.01	.02	.02
	(1.62)	(1.08)	(1.84)	(1.90)
Chronic diseases: none (ref.)				
– 1+ chronic diseases	−.02*	−.02*	−.01	−.02*
	(−2.25)	(−2.22)	(−1.32)	(−2.25)
Depression (Euro-D): insufficient **symptoms (ref.)**				
– 4+ symptoms	−.02**	−.00	−.02*	−.01
	(−2.96)	(−.39)	(−2.33)	(−.57)
Maximum grip strength: item nonresponse (ref.)				
– 1st quartile (very weak)	−.08***	−.01	−.04	−.02
	(−4.84)	(−.71)	(−2.65)	(−.68)
– 2nd quartile	−.07***	−.01	−.06**	.01
	(−4.60)	(−.29)	(−1.85)	(.50)
– 3rd quartile	−.08***	−.00	−.07**	−.01
	(−4.73)	(−.05)	(−3.20)	(−.33)
– 4th quartile (very strong)	−.08***	.00	−.04	−.00
	(−4.52)	(.10)	(−3.08)	(−.18)
Memory recall ability:				
– less than half of the words (ref.)				
– more than half of the words	−.03***	−.01	−.02**	−.02
	(−4.27)	(−1.66)	(−2.62)	(−1.84)
Hospital overnight stay in last 12 months: no (ref.)				
– yes	−.00	−.00	−.00	−.01
	(−.39)	(−.01)	(−.40)	(−.80)
Currently smoking: no (ref.)				
– yes	.03***	.01	.04***	.01
	(3.95)	(1.48)	(4.43)	(.77)
Currently drinking: never (ref.)				
– less than once a week	−.03**	−.02	−.05***	−.03*
	(−2.70)	(−1.65)	(−3.81)	(−2.41)
– 1–6 times a week	−.04***	−.02	−.05***	−.05***
	(−4.07)	(−1.65)	(−4.68)	(−3.64)
– almost every day	−.04***	−.01	−.04***	−.02
	(−4.33)	(−.89)	(−3.77)	(−1.62)
IADL: no IADL limitations (ref.)				
– 1+ IADL limitations	.03**	.03*	.04***	.04*
	(3.03)	(2.16)	(3.41)	(2.33)
Employment status: retired (ref.)				
– working	.01	−.00	−.01	−.01
	(.97)	(−.03)	(−.95)	(−.64)
– not working and other	−.00	−.00	.00	.01
	(−.34)	(−.47)	(.15)	(.39)
Making ends meet: difficulties (ref.)				
– no difficulties	−.01	−.00	.04***	−.02
	(−1.65)	(−.13)	(4.43)	(−1.67)

Table 3. Continued

	W2	W3|W2	W6	W6|W3
Total household income: item nonresponse (ref.)				
– 1st quartile	−.07***	−.04**	−.05***	−.04*
	(−5.54)	(−3.08)	(−3.39)	(−2.11)
– 2nd quartile	−.07***	−.05***	−.06***	−.04*
	(−5.93)	(−3.37)	(−4.51)	(−2.49)
– 3rd quartile	−.06***	−.06***	−.06***	−.02
	(−4.92)	(−4.21)	(−4.64)	(−1.51)
– 4th quartile	−.05***	−.04**	−.04**	−.00
	(−3.75)	(−2.83)	(−2.79)	(−.29)
Family respondent: no (ref.)				
– yes	−.01	−.00	−.01	.01
	(−.92)	(−.19)	(.62)	(.80)
Financial respondent: no (ref.)				
– yes	.02	−.00	−.01	−.02
	(1.25)	(−.04)	(−.95)	(−1.27)
Household respondent: no (ref.)				
– yes	−.03	−.00	−.01	−.00
	(−1.95)	(−.03)	(−.57)	(−.08)
N	19,761	13,466	16,619	10,412

Note. W2 = attrition in Wave 2; W3|W2 = attrition in Wave 3, conditional upon participation in Wave 2; W6 = attrition in Wave 6; W6|W3 = attrition in Wave 6, conditional upon participation in Wave 3. Z statistics in parentheses. HH = Household; IADL = instrumental activities of daily living; all models additionally include country dummies; $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$.

z-score = −3.38), and reporting the total household income among all income groups ($p < .01$; z-scores between −2.83 and −4.21) compared to item nonresponse in the total household income in Wave 1 (Table 3).

Other significant negative associations with attrition were observed for respondents who were between 70 and 79 years old, resided in a small town or rural village, received help from others, and reported at least one chronic disease in the first wave compared to corresponding reference category. Other positive significant associations with attrition in the third wave were observed with having reported at least one IADL limitation in Wave 1 (Table 3).

Some significant effects of individual and household charactersitics on attrition we found in the model for the second wave, that excluded reported deaths before Wave 2, could not be found in the conditional model for the third wave, where we excluded reported deaths before Wave 3 (Table 3).

4.1.3. Wave 6

The proportion of respondents who were born in the survey country, and of respondents who self-assessed their health in Wave 1 as good or better, and of respondents who were socially active in Wave 1 was between 2.40 and 3.64 percentage points larger for the panel members who participated in Wave 6 compared to the respective Wave 1 proportions. Moreover, the proportion of respondents who had a medium educational level was 3.06 percentage points smaller compared to the respective Wave 1 proportion (see

Supplemental data, Figure 2). In the conditional Wave 6 scenario (attrition in Wave 6, conditional upon participation in Wave 3) no larger deviation than two percentage points were observed (see Supplemental data, Figure 3).

Examining multivariate attrition in Wave 6, we observed for the unconditional scenario that many individual and household characteristics significantly predicted the drop out in the sixth wave (Table 3). Strong positive associations with attrition were found for respondents who smoked ($p < .001$; z-score $= 4.43$), made ends meet with no difficulties ($p < .001$; z-score $= 4.43$), and reported at least one IADL limitation in the first wave ($p < .001$; z-score $= 3.41$) compared to respondents who did not smoke, made ends meet with difficulties, and reported no IADL limitation in the first wave. The probability to drop out increased by four percentage points for each of these characteristics (smoking, making ends meet, and having at least one IADL limitation). Strong negative associations with attrition were found for respondents who were socially active ($p < .001$; z-score $= -7.06$), had a high educational level ($p < .001$; z-score $= -6.43$) compared to low educational level, were born in survey country ($p < .001$; z-score $= -6.27$), resided in a rural village ($p < .001$; z-score $= -6.90$) or small town ($p < .001$; z-score $= -5.56$) compared to city or large town. The decrease in probability to drop out for these groups ranged between four and ten percentage points. For further negative and positive associations in Wave 6 (with a lower significance level than 99.9% or with a smaller absolute value in z-score than 5) please see Table 3.

For attrition in Wave 6, conditional upon participation in Wave 3, we observed at the significance level of 99.9%, that highly educated and socially active respondents in Wave 1, and who were born in the survey country were less likely to drop out in Wave 6 than low-educated and socially inactive respondents and those, who were born outside the survey country (Table 3). Furthermore, residing in a small town and drinking between one and six drinks peer week, compared to residing in a city or large town and not drinking in Wave 1 decreased the probability of dropping out by five percentage points for the respective characteristics (Table 3). For further negative and positive associations (with a lower significance level than 99.9%) please see Table 3.

Comparing the conditional Wave 6 attrition model with the unconditional Wave 6 attrition model, we noted that far fewer individual and household characteristics were significantly associated with attrition in the conditional model. However, the age group 80 + in Wave 1, who were aged 92 + years in Wave 6, had a relatively large positive impact in both Wave 6 attrition models. The probability to drop out increased by eleven percentage points in the unconditional model and by ten percentage points in the conditional model for those old respondents (Table 3).

5. Summary and Discussion

This study examined the evolution of the initially recruited SHARE first-wave sample. With its specific target population, SHARE has a relatively large proportion of respondents who are at a high risk of attrition because of death. As we assumed that people who die are a selective group of the population and of the panel sample, we investigated the evolution of the SHARE panel with two defined samples. We used Analysis Sample A to study total attrition (non-mortality-related and mortality-related attrition), and Analysis

Sample B to study exclusively non-mortality-related attrition. We applied different methods to answer our research questions.

We answered Research Question 1 "How has the initially recruited SHARE first-wave sample (A and B) has evolved over waves" by calculating retention rates and estimating R-indicators. We detected declining retention rates with a major loss of respondents in the second wave. This finding is in line with previous literature (Lepkowski and Couper 2002; Schoeni et al. 2013; Lugtig 2014). Moreover, the retention rates observed in our study are about the same as those for second-wave response in other studies of older populations (Banks et al. 2011). In addition, we observed that the values of the R-indicators of the initially recruited SHARE sample (Analysis Sample A and B) dropped in the second wave but remained stable afterwards. Thus, we could show that, despite declining retention rates, the composition of the first-wave sample changed, but was maintained over waves with respect to many individual or household characteristics after the second wave. Furthermore, the results showed, when we excluded respondents that had been reported as dead before a given wave (Analysis Sample B, Wave 1 − Wave 6), that, the observed changes in the sample composition over time were mainly due to deaths (with the exception of Wave 2).

As SHARE collects data in various countries, it has to deal with country-specific differences, although it is harmonized ex ante. Therefore, we further investigated the evolution of the SHARE panel by Research Question 2 "Has the evolution of the initially recruited first-wave sample (A and B) over waves varied across countries?". We observed that the changes in the composition of the initially recruited sample over time differed across countries, although the differences were small. All countries followed the same trend, with a stable R-indicator value after the second wave (Analysis Samples A and B). However, comparing R-indicator values for Analysis Sample B (excluding deaths before a given wave) revealed larger differences across countries. These differences may be due to the quality of the respective death reports.

To answer Research Question 3 as to what characteristics and variables played the most important role in the changes in the composition of the initially recruited first-wave sample (dynamic Analysis Sample B) over waves, we examined various attrition scenarios by calculating subgroup retention rates and estimating multivariate logistic regression models on attrition. The results of the subgroup retention rate analyses were supported by those of the multivariate analyses. In all multivariate models, first-wave respondents who were born in the survey country, were residing in a rural area or small town, had a high level of education, and were socially active were less likely to attrite than first-wave respondents who were not born in the survey country, who were residing in a city or a large town, who had a low level of education, and were socially inactive. We did not observe that health-related variables, such as illness or age, were strong predictors of non-mortality-related attrition. Only very old respondents (aged 80+ in the first wave) had a high risk of attrition in later waves. Overall, birth in survey country, area of residence, education, and social activities played an important role in the non-mortality related attrition and their impact led to a decline of the R-indicators.

Comparing logit models from early waves with those from later waves, we noted that some significant associations declined to statistical insignificance in the multivariate models, especially in the models for attrition conditional upon participation in a specified

previous wave. This change in significance is reflected in the stabilizing R-indicator after the second wave.

The present study has a number of limitations. To draw conclusions from panel data about the general population aged 50 years or older, researchers need to consider and investigate initial nonresponse – that is, nonresponse that occurs in the recruitment stage of the panel. As the focus of the present study was on the evolution of the initially recruited first-wave sample over waves, we did not consider initial nonresponse. However, as initial nonresponse is an important factor for understanding the overall nonresponse process in SHARE and might have an impact on the data researchers use for analyzing dynamic processes in the European societies, future research should take it into account.

Another limitation of this study relates to the reporting of deaths. The SHARE countries included in the study differed in the share of reported deaths in the initially recruited sample over the course of the panel. Unlike the U.S. Health and Retirement Survey (HRS) or the English Longitudinal Survey of Ageing (ELSA) in England, SHARE cannot be linked to a mortality register because national mortality registers are lacking in most European countries (Bergmann et al. 2019). A comparison of the share of reported deaths in the initially recruited first-wave sample in SHARE with the mortality rate among persons aged 50+ years between 2004 and 2015 in Eurostat data (Eurostat 2004–2015) showed that only in a minority of the SHARE countries in our study was the share of respondents who died over the course of the panel lower than the estimated share of persons in the corresponding population group who died between 2004 and 2015 (Supplement data, Table 2). Thus, we may have underestimated the number of deaths in SHARE for a few countries due to a lack of information. However, we expected the share of deaths in Eurostat and SHARE to differ to some extent because SHARE excludes the hospitalized population from the sampling frame.

Notwithstanding these limitations, our study shows that, despite declining retention rates, the composition of an initially recruited panel sample can remain stable over later waves. The representativeness of the first wave sample (fixed Analysis Sample A and dynamic Analysis Sample B) did not decline further after the second wave. Moreover, this study informs researchers who wish to analyze dynamic processes over time about the impact of mortality-related and non-mortality-related attrition on the composition of the initially recruited first-wave SHARE sample over time. To further inform researchers wishing to analyze dynamic processes in SHARE over time, future research should examine the impact of mortality- and non-mortality-related attrition on cross-sectional and longitudinal estimates.

6. References

AAPOR (American Association for Public Opinion Research). 2016. *Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys* (9th ed.). AAPOR. Available at: https://www.aapor.org/AAPOR_Main/media/publications/Standard-Definitions20169theditionfinal.pdf (accessed August 2018).

Banks, J., M. Alastair, and J.P. Smith. 2011. "Attrition and Health in Ageing Studies: Evidence from ELSA and HRS." *Longitudinal and Life Course Studies* 2: 1–29. DOI: https://doi.org/10.14301/llcs.v2i2.115.

Bergmann, M., T. Kneip, G. De Luca, and A. Scherpenzeel. 2019. *Survey Participation in the Survey of Health, Ageing and Retirement in Europe (SHARE), Wave 1–7*. SHARE Working Paper Series 31-2017. Munich, Germany: Munich Center for the Economics of Aging (MEA), Max Planck Institute for Social Law and Social Policy. Available at: http://www.share-project.org/uploads/tx_sharepublications/WP_Series_41_2019_Bergmann_et_al.pdf (accessed February 2019).

Bianchi, A. and S. Biffignandi. 2017. "Representativeness in Panel Surveys." *Mathematical Population Studies* 24: 126–143. DOI: https://doi.org/10.1080/08898480.2016.1271650.

Bianchi, A. and S. Biffignandi. 2019. "Social Indicators to Explain Response in Longitudinal Studies." *Social Indicators Research* 141: 931–957. DOI: https://doi.org/10.1007/s11205-018-1874-7.

Binder, D. 1998. "Longitudinal Surveys: Why Are These Surveys Different from All Other Surveys?" *Survey Methodology* 24: 101–108. Available at: https://www150.statcan.gc.ca/n1/en/catalogue/12-001-X19980024347 (accessed June 2020).

Börsch-Supan, A., M. Brandt, C. Hunkler, T. Kneip, J. Korbmacher, F. Malter, B. Schaan, S. Stuck, and S. Zuber. 2013. "Data Resource Profile: The Survey of Health, Ageing and Retirement in Europe (SHARE)." *International Journal of Epidemiology* 42: 992–1001. DOI: https://doi.org/10.1093/ije/dyt088.

Börsch-Supan, A. 2017. *Survey of Health, Ageing and Retirement in Europe (SHARE) Wave 1. Release version: 6.0.0*. SHARE-ERIC. DOI: https://doi.org/10.6103/SHARE.w1.600.

Bristle, J., M. Celidoni, C. Dal Bianco, and G. Weber. 2019. "The Contributions of Paradata and Features of Respondents, Interviewers and Survey Agencies to Panel Co-Operation in the Survey of Health, Ageing and Retirement in Europe." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 182: 3–35. DOI: https://doi.org/10.1111/rssa.12391.

Buck, N., J. Burton, H. Laurie, P. Lynn, and S.C.N. Uhrig. 2006. *Quality Profile: British Household Panel. Survey Version 2.0: Waves 1 to 13: 1991–2003*. Colchester, UK: University of Essex, Institute for Social and Economic Research.

De Luca, G., C. Rossetti, and F. Malter. 2013. "Sample Design and Weighting Strategies in SHARE Wave 5." In *SHARE Wave 5: Innovations & Methodology*, edited by F. Malter and A. Börsch-Supan, 75–84. Munich, Germany: Munich Center for the Economics of Aging, Max Planck Institute for Social Law and Social Policy.

Eurostat. 2004–2015. Available at: "https://ec.europa.eu/eurostat/web/population-demography-migration-projections/population-projections-/database (accessed June 2019).

Fitzgerald, J., P. Gottschalk, and R. Moffitt. 1998. "An Analysis of Sample Attrition in Panel Data: The Michigan Panel Study of Income Dynamics." *The Journal of Human Resources* 33: 251–299. DOI: https://doi.org/10.2307/146433.

Leon, D.A. 2011. "Trends in European life expectancy: a salutary view." *International Journal of Epidemiology* 40: 271–77. DOI: https://doi.org/10.1093/ije/dyr061.

Lepkowski, J.M. and M.P. Couper. 2002. "Nonresponse in the Second Wave of Longitudinal Household Surveys." In *Survey Nonresponse*, edited by R.M. Groves, D.A. Dillman, J.L. Eltinge, and R. J.A. Little, 259–272. New York: John Wiley & Sons.

Lugtig, P. 2014. "Panel Attrition." *Sociological Methods & Research* 43: 699–723. DOI: https://doi.org/10.1177/0049124113520305.

Luiten, A. and B. Schouten. 2013. "Tailored fieldwork design to increase representative household survey response: an experiment in the Survey of Consumer Satisfaction." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 176: 169–189. DOI: https://doi.org/10.1111/j.1467-985X.2012.01080.x.

Lynn, P. 2009. "Methods for Longitudinal Surveys." In *Methodology of Longitudinal Surveys*, edited by P. Lynn, 1–19. Chichester, UK: John Wiley & Sons.

Lynn, P. and P. Lugtig. 2017. "Total Survey Error for Longitudinal Surveys." In *Total Survey Error in Practice*, edited by P.P. Biemer, E. De Leeuw, S. Eckman, B. Edwards, F. Kreuter, L.E. Lyberg, N.C. Tucker, and B.T. West, 279–298. Chichester, UK: John Wiley & Sons.

Mood, C. 2010. "Logistic Regression: Why We Cannot Do what We Think We Can Do, and What We Can Do about It." *European Sociological Review* 26: 67–82.

Moore, J.C., G.B. Durrant, and P.W. F. Smith. 2016. "Data Set Representativeness during Data Collection in Three UK Social Surveys: Generalizability and the Effects of Auxiliary Covariate Choice." *Journal of the Royal Statistical Society: Series A (Statistics in Society)*: 229–248. DOI: https://doi.org/10.1111/rssa.12256.

Myrskylä, M., J.R. Goldstein, and Y.A. Cheng. 2013. "New Cohort Fertility Forecasts for the Developed World: Rises, Falls, and Reversals." *Population and Development Review* 39: 31–56. DOI: https://doi.org/10.1111/j.1728-4457.2013.00572.x.

Olsen, R.J. 2018. "Respondent Attrition Versus Data Attrition and Their Reduction." In *The Palgrave Handbook of Survey Research*, edited by D.L. Vannette and J.A. Krosnick, 155–158. Cham, Switzerland: Springer.

RISQ. 2015. *Representative Indicators for Survey Quality-Tools*. Manchester, UK: University of Manchester. Available at: https://www.cmist.manchester.ac.uk/research/-projects/representative-indicators-for-survey-quality/tools/ (accessed August 2018).

Roberts, C., C. Vandenplas, and M.E. Stähli. 2014. "Evaluating the Impact of Response Enhancement Methods on the Risk of Nonresponse Bias and Survey Costs." *Survey Research Methods* 8: 67–80. DOI: https://doi.org/10.18148/srm/2014.v8i2.5459.

Schoeni, R.F., F. Stafford, K.A. McGonagle, and P. Andreski. 2013. "Response Rates in National Panel Surveys." *Annals of the American Academy of Political and Social Science* 645: 60–87.

Schouten, B., J. Bethlehem, K. Beullens, Ø. Kleven, G. Loosveldt, A. Luiten, Ka. Rutar, N. Shlomo, and C. Skinner. 2012. "Evaluating, Comparing, Monitoring, and Improving Representativeness of Survey Response through R-Indicators and Partial R-Indicators." *International Statistical Review* 80: 382–399. DOI: https://doi.org/10.1111/j.1751-5823.2012.00189.x.

Schouten, B., F. Cobben, and J. Bethlehem. 2009. "Indicators for the Representativeness of Survey Response." *Survey Methodology* 35: 101–113.

Schouten, B., N. Shlomo, and C. Skinner. 2011. "Indicators for Monitoring and Improving Representativeness of Response." *Journal of Official Statistics* 27: 1–24. Available at: https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/indicators-for-monitoring-and-improving-representativeness-of-response.pdf (accessed May 2020).

SHARE-ERIC. 2018. SHARE Publications: Journal Articles.Available at: http://www.share-project.org/share-publications/journalarticles00.html (accessed August 2018).

Smith, P., P. Lynn, and D. Elliot. 2009. "Sample Design for Longitudinal Surveys." In *Methodology of Longitudinal Surveys*, edited by M. Groves, G. Kalton, J.N. Rao, N. Schwarz, C. Skinner, and P. Lynn, 21–33. Chichester, UK: John Wiley & Sons.

Watson, D. 2003. "Sample Attrition between Waves 1 and 5 in the European Community Household Panel." *European Sociological Review* 19: 361–378. DOI: https://doi.org/10.1093/esr/19.4.361.

Watson, N. and M. Wooden. 2009. "Identifying Factors Affecting Longitudinal Survey Response." In *Methodology of Longitudinal Surveys*, edited by P. Lynn, 157–181. Chichester, UK: John Wiley & Sons.

Watson, N. and M. Wooden. 2019. "Chasing Hard-to-Get Cases in Panel Surveys: Is it Worth it?" *methods, data, analyses* 13: 199–222. DOI: https://doi.org/10.12758/mda.2018.03.

Received August 2020
Revised June 2019
Accepted February 2020

sciendo

Journal of Official Statistics, Vol. 36, No. 3, 2020, pp. 529–559, http://dx.doi.org/10.2478/JOS-2020-0028

The Action Structure of Recruitment Calls and Its Analytic Implications: The Case of Disfluencies

Bo Hee Min[1], *Nora Cate Schaeffer*[2], *Dana Garbarski*[3], *and Jennifer Dykema*[4]

We describe interviewers' actions in phone calls recruiting sample members. We illustrate (1) analytic challenges of studying how interviewers affect participation and (2) actions that undergird the variables in our models. We examine the impact of the interviewer's disfluencies on whether a sample member accepts or declines the request for an interview as a case study. Disfluencies are potentially important if they communicate the competence or humanity of the interviewer to the sample member in a way that affects the decision to participate. Using the Wisconsin Longitudinal Study, we find that although as they begin, calls that become declinations are similar to those that become acceptances, they soon take different paths. Considering all recruitment actions together, we find that the ratio of disfluencies to words does not predict acceptance of the request for an interview, although the disfluency ratio before the turning point – request to participate or a declination – of the call does. However, after controlling for the number of actions, the disfluency ratio no longer predicts participation. Instead, when we examine actions before and after the first turning point separately, we find that the number of actions has a positive relationship with participation before and a negative relationship after.

Key words: Participation; nonresponse; disfluencies; recruitment; survey introduction; interviewer-respondent interaction.

[1] Copenhagen Business School, Department of Management, Politics and Philosophy, Porcelænshaven 18B, 2000 Frederiksberg, Denmark. Email: bhm.mpp@cbs.dk
[2] University of Wisconsin-Madison, Department of Sociology, 1180 Observatory Drive Madison, WI 53706, U.S.A. Email: schaeffe@ssc.wisc.edu
[3] Loyola University Chicago, Department of Sociology, 440 Coffey Hall, 1032 W. Sheridan Rd. Chicago, IL 60660, U.S.A. Email: dgarbarski@luc.edu
[4] University of Wisconsin Survey Center, 475 No. Charter Street, Madison, WI 53706, U.S.A. Email: dykema@ssc.wisc.edu
Acknowledgments: We thank the participants in the Wisconsin Longitudinal Study for their generous contributions of time and information over many years. Some of the ideas in this article were presented at the International Workshop on Nonresponse 2016. This work was supported by a grant from the National Science Foundation (grant number SES-1230069) to Nora Cate Schaeffer. Additional support for this research was provided by the University of Wisconsin – Madison Office of the Vice Chancellor for Research and Graduate Education with funding from the Wisconsin Alumni Research Foundation to Nora Cate Schaeffer. Other support for the construction of the original data file, analysis, and collection of the data was received from the National Science Foundation (grant number SES-0550705) to Douglas W. Maynard, the Wisconsin Center for Demography and Ecology (National Institute of Child Health and Human Development Center Grant [grant number R24 HD047873]), Wisconsin Center for Demography of Health and Aging (National Institute on Aging Center Grant (grant number P30 AG017266, by the William H. Sewell Bascom Professorship, and by the University of Wisconsin Survey Center (UWSC). This research uses data from the Wisconsin Longitudinal Study (WLS) of the University of Wisconsin-Madison. Since 1991, the WLS has been supported principally by the National Institute on Aging (grant numbers AG-9775, AG-21079, AG-033285, and AG-041868), with additional support from the Vilas Estate Trust, the National Science Foundation, the Spencer Foundation, and the Graduate School of the University of Wisconsin-Madison. Since 1992, data have been collected by the University of Wisconsin Survey Center. A public use file of data from the Wisconsin Longitudinal Study is available from the Wisconsin Longitudinal Study, University of Wisconsin-Madison, 1180 Observatory Drive, Madison, Wisconsin 53706 and at http://www.ssc.wisc.edu/wlsresearch/data/.

1. Introduction

Although survey interviews have been regularly conducted by phone for decades, we know surprising little about the sequence and structure of actions within the opening of these calls or about the implications of this structure for measurement and analysis. This article contributes to filling this gap in two ways. First, by providing a detailed description of the actions of the interviewer in calls to recruit a sample member, we show how the sequence of actions in calls that end in declination differs from that in calls that end in acceptance. Second, we present a case study that shows that taking the action structure of the call seriously affects conclusions. To do this, we examine whether disfluent speech – such as "um" and other fillers – by the interviewer predicts whether a sample member accepts or declines the request for an interview. Our case study explores the conclusion of an earlier study that the likelihood of acceptance of the request to participate was greatest when interviewers were moderately disfluent (Conrad et al. 2013). That conclusion suggested that disfluencies might be consequential because they communicate the competence or humanity of the interviewer to the sample member in a way that affects the decision to participate. Our analysis makes salient that disfluencies originate in an underlying structure of actions: the level of disfluencies by interviewers in a call depends on which actions are performed, the number of those actions, and the characteristic number of words in and level of disfluency of those actions.

To describe the structure of recruiting calls, we take advantage of an existing case-control design extracted from the Wisconsin Longitudinal Study (WLS) that compared two key outcomes of the initial contact with the sample member – declinations and acceptances. A conversation analysis of calls made for the WLS was the basis for an interactional model of the recruitment call (Schaeffer et al. 2013; Maynard et al. 2010). Grounding our analysis in this earlier work, we first describe the action structure of calls that end in acceptance and declination. We next motivate an interest in disfluencies and show how they are distributed over and located in the various actions by interviewers during recruitment. As a last step, we use the case-control design to predict acceptance of the request for survey participation from the disfluencies in the interviewer's speech and other variables with which those disfluencies are highly associated.

We find that which actions are performed, how many of them, and their typical levels of disfluency differ for calls that end in acceptance or declination. Some key actions of interviewers necessarily differ in calls that end in acceptance (e.g., talk about when to begin the interview) compared to those that end in refusals (e.g., responding to a refusal). Considering all recruitment actions together, we find that the ratio of disfluencies to words does not predict acceptance of the request for an interview, although the disfluency ratio before the turning point – request to participate or a declination – of the call does. However, after controlling for the number of actions, the disfluency ratio no longer predicts participation. Instead, when we examine the relationship between the number of actions and the odds of participation before and after the first turning point separately, we find that the number of actions has a positive relationship with the odds of participation before and a negative relationship after. In order to train interviewers to be successful in

recruiting sample members, it is important to be able to identify which features of which actions engage – or disengage – sample members.

2. Challenges in Studies of Interaction During Recruitment Calls

Studies about how interviewers influence the outcome of recruitment calls face substantial challenges. Advances in recording, transcribing, and coding interaction have allowed us to observe how closely the actual events during a recruitment call match our impressions (e.g., Dijkstra and Smit 2002; Maynard et al. 2002; Maynard and Schaeffer 1997; Schaeffer et al. 2013). For example, although we are rightly concerned about how best to train interviewers to address sample members' concerns (e.g., Groves and McGonagle 2001), sample members frequently exit without providing interviewers opportunities to use those skills (e.g., Sturgis and Campanelli 1998; Schaeffer et al. 2013). Similarly, the finding that householders who ask questions are more likely to participate (Groves and Couper 1996), can be refined to distinguish between questions that come before the request to participate (associated with a lower likelihood of participation) and questions placed after the request (a higher likelihood of participating) (Schaeffer et al. 2013). We know that many sample members stay on the phone call for only a few seconds, and that implies that we need to know what constitutes the most effective first turn for the interviewer, because it is the only talk that many sample members hear (Schaeffer et al. 2018).

Examining the impact of the interviewer's talk and actions during recruitment also raises technical issues of several kinds. First, because the sample member speaks first when they answer the phone, every action by the interviewer is plausibly influenced by the preceding actions of the sample member. Attributing a causal influence to any specific action by the interviewer requires strong study design. In the absence of an experiment, it may be ultimately unclear whether, for example, interviewers deliver more scripted descriptions of the study in cases that end in acceptance because such sample members are receptive when interviewers describe the study or because such scripted requests are persuasive (Schaeffer et al. 2013). Second, although we might have theoretical reasons to think that turns, actions, words, or some other feature of interaction, such as disfluencies of speech, are likely to be a critical influence on the sample member's decision, these are all very highly correlated in practice. Third, as the descriptions below show in detail, calls that end in acceptance look very different from calls that end in declinations. For example, in a large proportion of declinations, the sample member hangs up before the interviewer issues a request for participation, but the request is delivered in all but a handful of calls that end in acceptance (Schaeffer et al. 2013, and see detail below). If different sequences of actions lead to different outcomes, this also raises questions of measurement. For example, can a measure, such as the number of words in an interaction, be meaningfully compared for declinations and acceptances, when those words are produced in very different actions?

3. Interactional Model of the Recruitment Call

Our description of the actions that interviewers and sample members perform and the sequence of some of those actions is based on the interactional model of the recruitment call proposed in Schaeffer et al. (2013) (see also refinement in Schaeffer,

forthcoming). That earlier work identified the actions in the call but did not describe their relative frequency or how their frequency differed for calls with different outcomes. The recruitment encounter begins when the sample member comes to the phone and ends with a hang-up (for declinations) or when the interview begins (for acceptances).

Table 1 shows an example of a call that was initially a declination but ended in acceptance (one of a handful of within-in call conversions in our data), with the interviewer's recruitment actions labeled. The actions comprise roughly three phases with fluid boundaries: Identification, purpose of call, and recruitment. The identification phase includes the greeting, self and institutional identification, and request to speak to the sample member. As Table 1 suggests, these three actions regularly occupy the interviewer's first turn after the sample member answers (see also Schaeffer et al. 2018). The purpose of the call may be conveyed by the institutional identification in the call opening, but it is explicit in the second phase, which includes actions that verify the identity of the sample member and refer to the advance letter or the study. The recruitment phase includes the request for participation, attempts to persuade after any subsequent declinations ("follow-up to declination"), and statements that refer to the length of the interview (which at this location in the call are about scheduling the interview).

An analysis comparing calls that end in acceptance and declination must consider the overall structure of the calls. Although the calls begin similarly, there is a "turning point" at which their paths diverge. (See definitions in Table 2.) The interviewer's first request for participation, when it occurs, serves as the first turning point in the call, after which the outcome almost always quickly unfolds. When the interviewer is not able to deliver the request, the outcome is already unfolding, and the first turning point becomes the last interviewer action before the first declination, hang-up, or acceptance. The turning point is important analytically for at least two reasons: First, because many sample members exit during or quickly after the interviewer's first turn, it is possible that actions of the interviewer very early in the call have strong effects on some sample members (e.g., Schaeffer et al. 2018). Second, recruitment calls are most comparable before the first turning point, that is, during the phases of identification and explaining the purpose of the call (when those phases occur). After the first turning point, interviewers are either scheduling the interview (for acceptances) or attempting to persuade (after a refusal).

The analysis of actions, their relative frequency and sequence allow us to describe the differences in the actions that precede two key outcomes, acceptance and declination. We then apply this analysis to understanding how a feature of interviewers' behavior that has been of interest to other investigators – disfluency during the recruitment call (e.g., Van der Vaart et al. 2006; Conrad et al. 2013; Schaeffer et al. 2013; Schaeffer et al. 2018) – might influence the sample member's participation.

4. Disfluencies in Interviewers' Recruitment Actions: A Case Study

Disfluencies are non-lexical components of speech that take several forms, potentially including fillers (predominantly "um" and "uh"); the broken-off talk and repetitions that result from false starts and repairs ("ma- may I" or "I I am coming"); pauses; and

Table 1. Call opening illustrating interviewer's recruitment actions, WLS, call ending in acceptance.

Phase	Turn number	Interviewer's recruitment action	Actor	Talk
Identification	1		SM	Hello.
	2	Greeting	INT	Hello.
	2	Self-identification		My name is (FIRST NAME LAST NAME).
	2	Institutional identification		I'm calling from the University of Wisconsin Survey Center at the University of Wisconsin Madison.
	2	Request to speak to sample member		and I'm hoping to speak to Mr. (FIRST NAME LAST NAME)
	3		SM	Yes.
Purpose of call	6	Verification of sample member	INT	Is this the (FIRST NAME LAST NAME) who was enrolled at (NAME OF HIGH SCHOOL) High School in 1957?
	7		SM	Yes.
	8	Letter reference	INT	As you um may have you may recall from our recent letter,
	8	Study reference		we're doing a follow up study (inaudible) of our sample of people who were Wisconsin high school seniors in 1957,
	9		SM	Mhmm
Recruitment	10	Request to participate	INT	and we'd like to interview you now for this important study.
	10	Request to participate		Is this a good time for you?
	11		SM	If it doesn't take too long.
	12	Length-of-interview statement	INT	It does take about an hour.
	13		SM	Then no.
	16	Follow-up to declination	INT	We can it can be done in parts so we can just do as much of it as you like and then we can reschedule um to call you back another time that's better for you.
	17		SM	Yeah. What w- what if we can do like fifteen minutes. Would be alright, I guess, but otherwise I've got things I gotta do.

Note: SM is the sample member; INT is the interviewer. Within a turn, each action is shown on a separate line. Punctuation has been added and minor actions (e.g., exclamations) deleted for readability.

Table 2. Concepts and definitions.

Type of concept	Concept	Definition and coding
Structure of call	Interviewer's recruitment actions	A unit of talk in the opening of the recruitment call that accomplishes a specific interactional task. Actions were identified as regular components of the call in a conversation analysis. A turn may include more than one action. See examples and labels in Table 1. Source: Maynard, Freese, and Schaeffer 2010; Schaeffer et al. 2013. (See also concept of "moves" in Conrad et al. 2013.)
	First turning point	For acceptances, the first turning point is the first request for participation or the last interviewer action before an acceptance. For declinations, the first turning point is the last interviewer action before the first (blocking) declination, hang up, or acceptance, whichever came first. Sources: Schaeffer et al. 2013; see also Conrad et al. 2013.
	Congruent/ incongruent actions	Actions that "follow up" declinations are common in calls that end in declination and uncommon in calls that end in acceptance; the follow-up to declination is "congruent" with the declination path. Similarly, talk about the length of the interview and how to administer it in parts are common in calls that end in acceptance and uncommon in calls that end in declination; the "statement of length of interview" is congruent with the acceptance path.
Disfluencies	Disfluency	Irregularities in speech that intrude into the smooth production of talk. We examine three possible disfluencies: fillers, broken-off talk that results from false starts and repairs, and nonpropositional elements. Source: Bortfeld et al. 2001; Schober et al. 2012.
	Fillers	A set of non-lexical tokens. Our data include "Uh Um Ah Hmm Mmm Eh Aw Er Nn Num," and we standardized their spelling to facilitate accuracy of text coding. Source: Bortfeld et al. 2001.
	Nonpropositional elements	A cover term for words that are used as "discourse markers", for example, around fillers at the beginning of a turn (e.g., "and um and"), or as "acknowledgments" (e.g., "okay") when the speaker changes. The most frequent nonpropositional elements in our corpus are "and," "okay," "alright," "so," "well," and "but." Most of the remaining were synonyms for these words and appeared four or fewer times (e.g., "excellent" instead of "okay"). We counted nonpropositional elements immediately preceding one or more fillers (that immediately precede an action). Source: Adapted from Bortfeld et al. 2001.
	Broken-off talk	Broken-off talk occurs when speaker stops talking before completing a unit of talk, usually an action. Broken-off talk was indicated by trailing "-" in the transcripts. Broken-off talk was usually followed by a restart (e.g., "I- I was"), We used the length of a pause at speaker change to distinguish broken-off talk from interruptions. If talk overlapped at the transition, we considered the break-off to be due to an interruption.Source: Our broken-off talk is similar to one component of Bortfeld et al.'s (2001) "restart." Due to our limited resources, we did not examine all restarts, which can be quite complex.

"nonpropositional elements," which include discourse markers ("but" or "well") and acknowledgment tokens ("okay"), that precede or are embedded with fillers (Bortfeld et al. 2001). See Table 2 for definitions used here.

4.1. Interviewers' Disfluencies: Theoretical Issues and Prior Research

In studies of survey interviews, most attention has been given to disfluencies of respondents during the interview itself: Disfluent respondents may be treated as having comprehension problems (Schaeffer and Maynard 2002; see also Schober and Bloom 2004), and respondents' disfluencies may indicate that an answer is less accurate or reliable (Draisma and Dijkstra 2004; Draisma et al. 2005; Schaeffer and Dykema 2011; Garbarski et al. 2011; Schober et al. 2012; Smith and Clark 1993; Mathiowetz 1999).

The impact of the interviewer's disfluencies on recruitment may depend on the perceptions of sample members (e.g., Van der Vaart et al. 2006). Sample members may ignore disfluencies; or disfluencies may affect whether the interviewer is perceived as comfortable, confused, honest, anxious, and so forth (e.g., Christenfeld 1995; Fox Tree 2002, 2007). Disfluencies may perform other tasks that are informative: Listeners may hear a disfluency as signaling that the next item mentioned may be new (Arnold et al. 2004; Arnold et al. 2007; Barr and Syfeddinipur 2010) or that a repair is forthcoming (Brennan and Schober 2001). Disfluencies may also separate "intonation units" (Clark and Fox Tree 2002) in a way that may serve as audible "punctuation" and so make speech easier to understand. "Uh(m)" may serve to delay dispreferred acts or the purpose of a call (Schegloff 2010).

In studies of interviewers' success in recruiting sample members, disfluencies, as defined here, have been studied less than other acoustic and behavioral qualities of the interviewer's speech (e.g., Groves et al. 2008). For an interviewer, being disfluent may (or may not) be associated with whether the interviewer is successful at recruiting sample members (Schaeffer et al. 2013; Schaeffer et al. 2018; Conrad et al. 2013; Oksenberg and Cannell 1988; Van der Vaart et al. 2006; Sharf and Lehman 1984). Schaeffer et al. (2013) found higher odds of participation when disfluencies were present; however, they also predicted higher odds of participation if the interviewer's mention of the advance letter or description of the study followed an available (optional) script, a practice that reduced disfluencies (results not shown). Their subsequent analysis of the interviewer's first turn indicated that the odds of participation were lower ($p < 0.10$) if that turn began with a filler (Schaeffer et al. 2018), although few first turns began in this way. In an analysis with multiple samples and a different design, Conrad et al. (2013, 201) found that participation had a curvilinear relationship with the interviewer's filler rate (fillers per 100 words): The proportion of householders who agreed to participate was lowest for interviewers with the lowest (0) or highest filler rate. Their interpretation was that interviewers with no disfluencies may sound robotic, and those with too many disfluencies may sound incompetent.

4.2. Interviewers' Disfluencies and Their Locations

As have other researchers, we observe that interviewers' disfluencies regularly occur in three locations: At the beginning of a turn, at the beginning of an action within a turn, and within an action (Boomer 1965; Shriberg 1996; Clark and Fox Tree 2002, 95). Excerpt 1

Excerpt 1. Interviewer's canonical introduction, showing fillers before next action, call that ends in declination, WLS, punctuation and capitalization added.

Line	Turn	Action	Transcript	Disfluency and location
1	4	Greeting	Hi, sir.	
2	4		uh	Filler before next action
3	4	Self-identification	My name's (FF) (LL).	
4	4	Institutional identification	I'm calling from the University of Wisconsin in Madison for the Wisconsin Longitudinal Study.	
5	4		um	Filler before next action
6	4	Letter reference	Did you happen to get our letter in the mail recently?	

shows how disfluencies are located within the actions of the call opening, which suggests ways they might function.

Excerpt 1 begins after the interviewer delivered an "efficient" introduction (which begins by confirming that he is speaking with the sample member instead of with self-identification, not shown). Excerpt 1 begins with turn 4, the interviewer's second turn, in which he adapts the "canonical" introduction (one that begins with greeting and self-identification – provided on his screen) (Schaeffer et al. 2018) and adds a reference to the advance, because WLS interviewers were authorized to treat the scripted introduction as "flexible" (Morton-Williams 1993; Houtkoop-Steenstra and Van den Berg 2002). He inserts the fillers "uh" and "um" in lines 2 and 5 before his "self-identification" and "letter reference" actions, so that the disfluencies reinforce the meaningful units within the interviewer's stream of talk.

In Excerpt 2, the request to speak to the sample member at line 8 begins with a discourse marker ("and"), followed by a filler ("uh"), and then broken-off talk ("we were wonder-"), followed by another filler ("uh") and a restart or repair ("we were wondering"). The midstream embedded disfluencies ("we were wonder- uh") do not mark transitions of speaker or action the way the "uh" at line 4 or the initial "and uh" in line 8 do.

These excerpts suggest that a disfluency in any location may indicate that the speaker is planning speech, retrieving words, or undertaking a repair. Disfluencies that occur midstream during an action or turn (e.g., line 8 in excerpt 2) may be distinct, either in their origins or in how they are perceived by listeners. These midstream or embedded disfluencies do not perform the turn-taking or transitional work performed by disfluencies that begin an action or turn; they may communicate that the speaker is searching for what to say or how to say it and so be more consequential.

4.3. Hypotheses: Interviewers' Actions, Disfluencies, and Participation

Prior investigations of the impact of the interviewer's disfluencies on participation have not accounted for how they are located within actions. Our hypotheses examine the

Excerpt 2. Interviewer's second introduction with sample member, illustrating discourse markers and fillers, call that ends in declination, WLS, punctuation and capitalization added, SM = sample member, INT = interviewer.

Line	Turn	Action	Actor	Transcript	Disfluency
1	3		SM	You spoke so fast I couldn't understand who this was.	
2	4	Change-of-state Token	INT	Oh,	Discourse marker at beginning of turn
3	4	Apology		I'm sorry.	
4	4			uh	Filler before next action
5	4	Self-identification		My name's (FF) (L).	
6	4	Institutional identification		I'm calling from the University of Wisconsin Madison for the W L S study the	
7	5		SM	Okay.	
8	6	Request to speak to sample member	INT	And uh we were wonder- uh we were wondering if we could speak to dzhu- (FF) (LLL).	Discourse marker, filler, broken-off talk, filler, restart at beginning of turn
9	7		SM	This is she.	

components of a sample member's experience of disfluencies: Disfluencies occur in actions – which can vary in frequency, length, and fluency – at various locations in the call.

We first compare the action structure of calls that end in declination and acceptance. For example, it is possible that some calls with no disfluencies end in refusal because the householder hangs up before the interviewer has much opportunity to talk (see discussion in Sturgis and Campanelli 1998) and thus to be disfluent. Similarly, the level of disfluency could be high if interviewers become flustered and increasingly disfluent when trying to persuade very resistant sample members. The model of the call, the literature briefly reviewed above, and our observations of disfluencies lead to our first prediction, that the action structure of the call opening will be similar for calls with that end in acceptances and declinations but diverge after that.

We then turn to the fluency of actions. The varied actions in the call make different demands on the interviewer. The identification phase is familiar and well-rehearsed and likely to be delivered fluently. Once the identification phase is complete, the behavior of the sample member becomes less predictable, and sometimes hostile, and the interviewer must plan and execute actions quickly. When a sample member declines, the interviewer may perceive the stakes as higher; the interviewer's actions following a declination take place in an uncertain environment, and the interviewer may be more, and differently, disfluent as a result. This leads to our second prediction, that the most routine actions, in the identification phase of the call, will have low rates of disfluency that are similar

for calls with each outcome, and most disfluencies that do occur will be placed at the beginning of actions. Later in the call, and particularly after a declination, the uncertainty of the interactional environment will lead to more and more varied disfluencies, including midstream disfluencies of broken-off talk and midstream fillers. This leads to our third prediction, that actions that follow up declinations, an action congruent with declinations, will be less fluent than talk about the length of interview (used to schedule an interview after an acceptance), an action congruent with acceptances.

If we observe these differences in action structure and in the fluency of various actions, we can appreciate in a different way the complexity of estimating the impact of the interviewer's disfluency on participation: For example, if interviewers are most disfluent when trying to persuade reluctant sample members, and these disfluencies do not appear for accommodating sample members, then we might suspect that these disfluencies result, at least in part, from the sample member's resistance (or propensity to participate) and so might be effects of a decision the sample member has already made. So, to complement our descriptive analysis, we then take advantage of our case-control design (see below) to predict participation from disfluencies. Because the structure of calls that end in declination and acceptance are most comparable before the first turning point in the call (results below), we examine the impact of our measures before and after that point. Because calls move quickly to their outcome after the first turning point, we might predict that disfluencies before the first turning point make a poor impression and reduce the likelihood of participation. We predict that calls with more disfluencies after the first turning point – for example, because interviewers get flustered attempting to persuade resistant sample members – will be more likely to end in declination. However, we expect that the number of disfluencies, words, and actions are interdependent in ways that makes it difficult to assess the impact of each, and we address this by examining the impact of disfluencies net of the number of actions.

5. Methods: Data, Variables, and Analysis

Our analysis requires detailed transcripts and some ability to estimate the impact – or, at least, predictive strength – of features of interaction on participation. Designing an experiment to address this topic in a production context would present substantial obstacles. So, we address our research questions using a recent case-control design constructed from the Wisconsin Longitudinal Study (WLS). See details of study design in Schaeffer et al. (2013). However, because we use matched pairs, our declinations and acceptances cannot be combined to estimate characteristics of the WLS. (For example, our sample, by design, has a "response rate" of 50%, but the response rate for the WLS is much higher.) Considered separately, our declinations and acceptances each constitute a collection or corpus, rather than a probability sample from a specific population, although some of our tests treat them as independent samples.

5.1. Sample

Our analysis uses digital records of phone contacts from the 2004 round of the WLS, which interviewed 80% of surviving panel members. The WLS began with a one-third

sample of 1957 Wisconsin high school graduates and followed up in 1964 (mail to parents), 1975 (telephone), 1992 (telephone and mail), and 2004 (telephone and mail). (See Hauser 2005.) The case-control study selected 257 pairs of cases (the maximum number of pairs that could be made). One member of each pair declined to be interviewed in their first contact with the interviewer (declination), and the other member of the pair accepted on the first contact with an interviewer (acceptance). Pair members matched exactly on sex and past participation in the WLS and as closely as possible on estimated propensity to participate. The model estimating the propensity to participate included education, high school class rank, high school cognitive assessments, self-reported health, sex, and past participation (See Appendix, Section 8, for additional details.). To the extent that the pairs are successfully matched on propensity to participate, differences in outcome should be largely due to the behavior of the interviewer. We recognize, however, that the matching of the pairs is subject to measurement and other errors, and we modulate our claims of causality accordingly.

5.2. *Analysis of Actions*

Actions (listed in the example in Table 1) and their features were identified in an extended conversation analysis of the call opening. The interactional model of the recruitment call summarizes this analysis and the reliability of coding is described elsewhere (Maynard et al. 2010; Schaeffer et al. 2013; Schaeffer et al. 2018; Schaeffer, forthcoming). The definitions of the concepts in the present analysis are summarized in Table 2 .

5.3. *Variables: Measuring Disfluencies*

Disfluencies have been operationalized in many ways. Our definitions (see Table 2 and Appendix) drew heavily on the concepts and operational rules described by Bortfeld et al. (2001, 131–132) because of their relevance and completeness. We developed computer code to identify and count disfluencies in transcripts that had been standardized in preparation. The summary statistics we discuss are described in Table 3, and the Appendix gives details of underlying rules for counting disfluencies.

5.4. *Analysis*

Our analysis has two parts. First, our descriptive analysis examines the components of exposure to disfluencies, which comes by way of the specific actions the interviewer performs and the frequency, length, and fluency of those actions: We first describe the action structure of the calls that end in acceptance and declination, and then we describe the fluency of those actions. This decomposed description is suitable for our corpus of acceptance and declination calls. This decomposition is useful because other populations of sample members and interviewers and other study designs could give rise to a different distribution of actions or different levels of disfluency in those actions. Results from our approach might be generalizable if these actions appear in other studies; for example, if interviewers self-identify in similar ways across a variety of populations, the effects of self-identification might then be expected to be similar. In the descriptive analysis, when

Table 3. Calculation of summary measures.

Table	Label in table	Unit	Summary statistic	Calculation
5	Mean turn number of the action, all actions of a given type	Actions of given type	Mean of turn numbers of all actions of given type	Within each call, each action of a given type was assigned its turn number. (The sample member's answer to summon is turn 1.) The mean turn number of all actions of a given type was calculated. A call may have more than one action of a given type.
5	Modal turn number of the action, all actions of a given type	Actions of given type	Mode of turn numbers of all actions of given type	Within each call, each action of a given type was assigned its turn number. (The sample member's answer to summon is turn 1.) The most common value of the turn numbers of all actions of the type is calculated. A call may have more than one action of a given type.
5	Mean of number of actions per call, for calls with the action	Calls with action of given type	Sample-level mean of number of actions of given type per call	Within each call, the number of actions of given type is counted. The mean number of actions of a given type is calculated. If a call has no action of a given type, it is excluded from the calculation.
5	Mean of mean number of words per action, for calls with the action	Calls with action of given type	Sample-level mean of call-level mean number of words in each action (excluding fillers and broken-off talks), for calls with at least one instance of given action.	Within each recruitment action, the number of words in each action (excluding fillers and broken-off talk) is counted. Within each call with at least one action of a given type, the mean of the number of words in actions of a given type is calculated. Then the sample-level mean of call-level means is calculated. If a call has no action of a given type, it is excluded from the calculation.

Table 3. Continued.

Table	Label in table	Unit	Summary statistic	Calculation
6	Initial: mean of proportion of actions that begin with any disfluency	Calls with action of given type	Sample-level mean of call-level proportion of actions of given type with initial disfluency	Within each action, a dummy variable indicates if there is a initial disfluency. Within each call, the proportion of actions of a given type with a initial disfluency is calculated. Then a sample-level mean of these proportions is calculated. If a call has no action of a given type, it is excluded from the calculation.
6	Midstream: mean of proportion of actions with any midstream filler or broken-off talk	Calls with action of given type	Sample-level mean of call-level proportion of actions of given type with midstream disfluency	Within each action, a dummy variable indicates if there is a midstream disfluency. Within each call, the proportion of actions of a given type with midstream disfluency is calculated. Then a sample-level mean of these proportions is calculated. If a call has no action of a given type, it is excluded from the calculation.
6	Mean of proportion of actions with any disfluency in any location	Calls with action of given type	Sample-level mean of call-level proportion of actions of given type with disfluency at any location	Within each action, a dummy variable indicates if there is a disfluency at any location. Within each call, the proportion of actions of a given type with any disfluency is calculated. Then a sample-level mean of these proportions is calculated. If a call has no action of a given type, it is excluded from the calculation.

Table 3. Continued.

Table	Label in table	Unit	Summary statistic	Calculation
6	Mean of mean number of disfluencies in any location	Calls with action of given type	Sample-level mean of call-level mean of number of disfluencies in actions of a given type in the call	Within each action, the number of disfluencies (fillers, broken-off talk, and discourse markers) is calculated. Within each call, the mean number of disfluencies in actions of a given type is calculated. Then a sample-level mean of these means is calculated for calls with actions of a given type. If a call has no action of a given type, it is excluded from the calculation.
7	Number of disfluencies	Calls	Total count of disfluencies in the action or call	Within each of the recruitment actions, fillers at any location in the action, broken-off talk, and discourse marker are counted. Within each call, these numbers are summed across the recruitment actions.
7	Number of words/100	Calls	Total count of words in recruitment actions or calls, excluding disfluencies, divided by 100	Within each of the recruitment actions, words (excluding fillers at any location in the action, broken-off talk, and discourse marker) are counted. Within each call, the numbers of words across recruitment actions is summed and divided by 100.
7	Number of actions	Calls	Total count of recruitment actions or actions of a given type in the call	Within each call, the number of recruitment actions (see Table 1) or actions of a given type is counted.
7	Disfluency ratio	Calls	(number of disfluencies/ (number of words/100))	Within each of the recruitment actions, the number of disfluencies (fillers at any location in action, broken-off talk, and discourse markers) and the number of words excluding disfluencies is counted. Within each call, the total number of disfluencies and the total number of words excluding disfluencies are counted. The ratio of the total number of disfluencies to the (total number of words/100) is calculated.

we provide t-tests for differences in means or proportions, we treat the acceptances and declinations as independent samples from the WLS.

In the second part of our analysis, we use the case-control design (conditional logit analyses, clogit in Stata) to predict participation as the dependent variable (see Schaeffer et al. 2013; Schaeffer et al. 2018). The following likelihood function for clogit with groups (that is, pairs of observations) was used:

$$L = \sum_{\{i \in I_1\}} \left(\sum_{\{j: y_{ij}=1\}} \left[(\mathbf{x}_{i2} - \mathbf{x}_{i1}) \left[(-1)^{I(j-2)} \boldsymbol{\beta} \right] - ln\left(1 + e^{(\mathbf{x}_{i2} - \mathbf{x}_{i1})[(-1)^{I(j=2)}\boldsymbol{\beta}]} \right) \right] \right)$$

where

- The first beta is a multiplier to the difference in the x values in the i-th group
- The bold font for the x and betas in the formula indicates that there may be more than one regressor in the model
- i is the group identifier
- ij, where $j \in \{1,2\}$, is the jth observation of the ith group
- $I_1 = \{i \,|\, y_{i1} + y_{i2} = 1\}$
- x_{ij} is the row of covariates associated with the jth observation of the ith group
- $I(j = 2)$ is the indicator function for $j = 2$

The outer summation is over all pairs in which the pair's responses contain one 0 (declination) and one 1 (acceptance). The inner summation is over the single observation within the pair in which the response is 1. (The likelihood function minimized by clogit is described on the Stata clogit page (http://www.stata.com/manuals14/rclogit.pdf). This section references several other sources, including Chamberlain (1980), which is the basis for the likelihood function above (Mark Banghart, personal communication).

Conditional logit is similar to a fixed-effect logit in which the matching characteristics (see above) are used as categorical regressors in the model. The analysis thus adjusts for characteristics that the pairs are matched on and anything else that they have in common. A conditional logit regression estimates the association between the within-pair action of interest and participation; it "conditions" the intercept for each pair out of the analysis. The intercepts for the pairs are nuisance parameters and not of substantive interest but can bias estimates if not accounted for. Because our sample size is small, and we want to identify avenues for future investigation, we discuss relationships that are significant with the relatively generous $\alpha = 0.10$ but note when results are significant by conventional standards ($\alpha = 0.05$).

6. Results

We provide descriptions that have been absent from the literature to date: (1) a detailed picture of the overall action structure of the recruitment call, and (2) variation in disfluencies by location of disfluency (at beginning of action or midstream within an action), and by type of action. We then use information about actions and disfluencies to predict acceptance of the request for participation.

6.1. Interviewers' Actions in Recruitment Calls: Overall Structure

Table 4 and Table 5 describe interviewers' actions. Table 4 gives counts of actions (Panel A) and of calls with actions (Panel B and Panel C). These counts are descriptive in themselves and also document the number of units on which the summary statistics in Table 5 and Table 6 are based. Table 5 describes other features of actions: the turns in which they are located (Panel A) and their number and length (in words) (Panel B), for calls with each outcome.

Results in both tables confirm that, overall, the identification phase is similar for declination and acceptances calls. In Table 4, Panel B demonstrates that the single significant difference (in issuing a greeting) is substantively small. In Table 5, Panel A reinforces the similarity of the actions in the identification phase: The mean (column (a)) and modal turn numbers (column (b)) of the actions are similar or identical across actions and outcomes.

After identification some declining sample members have hung up, and the trajectories of remaining calls destined for acceptance or declination increasingly diverge: the number of calls with each action after the identification phase is significantly different for declinations and acceptances (Table 4, Panel B). More calls that end in acceptance (compared to calls that end in declinations) have each of the actions in the "purpose of call" phase: the interviewer's verification that they have reached the sample member, questions about the advance letter, and descriptions of the study (Table 4, Panel B). Differences in the sequential position of some actions also begin to appear after the identification phase (Table 5, Panel A): For example, the modal turn number for the letter reference and request to participate is one turn later for calls that end in acceptance than for declinations. For the "purpose of call" actions, the mean turn numbers are higher than the mode, as the call structure becomes less conventional.

A critical difference in the action structure of calls with the two outcomes is that the request to participate occurs in almost all acceptances, but in fewer than half of declinations (Table 4, Panel B). Of the 257 declinations, 15 have hung up and another 141 have declined and hung up before the interviewer can issue a request. In two cases the first request came after the first declination (detail not shown). This massive and selective exodus of sample members very early in the call means that many who decline have almost no exposure to the interviewer – they are exposed to few interviewer actions, disfluencies, or attempts at persuasion. In Table 4, column (d) in Panel C clarifies what Panel B suggests – that interaction after the first turning point is dominated by the congruent actions: Follow-up actions in declinations (in 175 of 257 calls) and talk about the length of the interview in acceptances (in 146 of the 257 calls).

Thus far, our results suggest that exposure to disfluencies might differ for declination and acceptance calls because the outcomes are preceded by different actions. Panel B in Table 5 indicates that the different frequency (column (c)) and length in words (column (d)) of various actions could also contribute. Once again, calls with both outcomes are similar in the identification phase: Comparing declinations and acceptances, the mean number of actions is similar for the two outcomes (with a single small difference), as is the mean number of words per action. In the "purpose of call" phase, the mean number of references to the advance letter is greater and the mean number of actions discussing the

Table 4. Number of recruitment actions and number of calls with each type of recruitment action by the interviewer, by location and call outcome.

		Number of actions		Number of calls in which action occurs.. . .						
		Panel A		Panel B			Panel C			
		(a) Actions of given type		(b) Anywhere in the call			(c) Up to and including first turning point		(d) After first turning point	
Phase	Recruitment action	Declination	Acceptance	Declination	Acceptance	*p*	Declination	Acceptance	Declination	Acceptance
Identification	Greeting	434	430	257	253	**	256	253	11	3
	Self-identification	246	254	236	234		236	234	1	2
	Institutional identification	309	309	256	255		255	255	7	4
	Request to speak to sample member	230	236	208	209		207	208	2	2
Purpose of call	Verification of sample member	80	175	75	168	***	72	128	3	40
	Letter reference	205	343	157	233	***	147	224	17	24
	Study reference	247	316	128	200	***	117	182	31	44

Table 4. Continued.

Phase	Recruitment action	Number of actions		Number of calls in which action occurs.. . .						
		Panel A		Panel B			Panel C			
		(a) Actions of given type		(b) Anywhere in the call			(c) Up to and including first turning point		(d) After first turning point	
		Declination	Acceptance	Declination	Acceptance	*p*	Declination	Acceptance	Declination	Acceptance
Recruitment and persuasion	Request to participate	129	394	101	253	***	99	253	9	69
	Follow-up to declination	928	26	175	10	***	0	0	175	10
	Length-of-interview statement	51	285	32	146	***	2	1	30	146
Total number of actions or calls		2,859	2,768	257	257		257	257	257	257

Note: Cells in Panel A are number of actions in calls of the type; cells in Panels B and C are number of calls in which an action occurred at least once. Cells for actions after the turning point that are "incongruent" with the outcome are in gray font; in addition, in Panel C, an action is shown in black font in the "up to or including" or "after" panel if the action occurs in at least 9 calls for both acceptances and declinations.

Base analytic sample is matched pairs: 257 acceptances and 257 declinations. Interaction begins when sample member answers or is brought to the phone and ends when the sample member hangs up (declination) or the interview begins (acceptance). For acceptances, the first turning point is the first request for participation or the last interviewer action before an acceptance. For declinations, the first turning point is the last interviewer action before the first (blocking) declination, hang up, or acceptance, whichever came first.

Table 5. Selected characteristics of calls with recruitment actions, by call outcome.

Phase	Recruitment Action	Panel A Cells based on all actions of a given type: (a) Mean turn number of the action, all actions of a given type — Declination	Panel A — Acceptance	Panel A: (d) Mean of mean number of words per action, for calls with the action — Declination	Panel A — Acceptance	Panel B Cells based on all actions of a given type: (d) Mean of mean number of words per action, for calls with the action — Declination	Panel B — Acceptance	Panel B — p	Panel B: (d) Mean of mean number of words per action, for calls with the action — Declination	Panel B — Acceptance	Panel B — p
Identification	Greeting	1.52	1.54	1	1	1.69	1.70		2.19	2.24	
	Self-identification	1.77	1.79	2	2	1.04	1.09	*	5.44	5.37	
	Institutional identification	2.01	2.01	2	2	1.21	1.21		11.93	11.92	
	Request to speak to sample member	1.21	1.33	1	1	1.11	1.13		6.64	6.76	
Purpose of call	Verification of sample member	2.69	3.73	2	2	1.07	1.04		17.72	18.60	
	Letter reference	3.33	3.67	2	3	1.31	1.47	**	11.25	10.63	
	Study reference	4.98	5.07	3	3	1.93	1.58	*	23.71	21.29	*
Recruitment and persuasion	Request to participate	3.88	4.90	3	4	1.28	1.56	***	18.37	17.73	
	Follow-up to declination	6.90	5.65	4	4	5.30	2.60	***	14.12	21.77	*
	Length-of-interview statement	6.69	6.49	7	5	1.59	1.95	**	28.16	28.91	*

Note: In Panel A, statistics are calculated for actions of a given type; multiple actions of a given type within a call are all included; Panel A in Table 4 shows the number of actions of a given type. In Panel B, unit is calls in which an action of a given type occurred; Panel B in Table 4 shows the number of calls with an action of a given type. Tests in Panel B compare the mean for the two call outcomes. Because a call may have more than one action of a given type, means for number of words are calculated for each call in which an action occurs; the mean (of the means) for the calls that included a given action is reported in the table.

Cells for actions that are "incongruent" with the outcome are in gray font. Base analytic sample is matched pairs: 257 acceptances and 257 declinations. Interaction begins when sample member answers or is brought to the phone and ends when the sample member hangs up (declination) or the interview begins (acceptance).

*$p < 0.10$, **$p < 0.05$, ***$p < 0.01$.

study smaller for acceptance than declination calls. The recruitment phase is substantially different for the two outcomes: Column (c) shows that the mean number of requests to participate is larger for acceptance than declination calls, and thereafter the calls become even more difficult to compare. For example, in column (c) the mean number of "follow-up to declination" actions in declination calls is 5.3, and the mean number of actions about the length of interview in acceptance calls is 1.95. As shown in column (d), however, the former action is shorter on average (mean average number of words $= 14.1$) than are actions about the length of interview in acceptance calls (28.9). Thus, the relative frequency and, to a lesser extent, the relative length of different actions provide different opportunities for exposure to the interviewer's disfluency.

Taken together, the results in Table 4 and Table 5 support our first hypothesis: Calls with each outcome begin similarly but diverge sharply after the first turning point. We also find that a substantial number of sample members who decline have extremely short calls with few actions by the interviewer, and a substantial number have longer calls with multiple follow-up attempts by the interviewer. The actions that occur and their frequency and length vary for calls with different outcomes.

6.2. *Disfluencies in Interviewers' Actions*

Table 6 summarizes features of disfluencies in the recruitment actions by call outcome. (Table 3 gives details of calculations.) We examined many measures of disfluency (some of them overlapping) to understand how they differed and select among them. We present summary statistics and test the difference between acceptances and declinations. As part of our analytic approach that distinguishes the presence of an action and its features – a sort of decomposition strategy – we focus here on calls in which the action occurred in order to characterize disfluencies in various actions when they occur. Panel A of Table 6 presents the mean (across all calls with an action) of the proportion of actions that begin with a disfluency (initial disfluencies) (column (a)) or include a midstream disfluency or broken-off talk (midstream disfluencies) (column (b)). Panel B summarizes across components of disfluency, so that columns (c) and (d) of Panel B are different summaries of the information in columns (a) and (b).

In the identification phase, there is a single notable difference between declination and acceptance calls: Although a disfluency at the beginning of a greeting is rare, it appears more often in declinations (as reported in Schaeffer et al. 2018). As predicted, compared to actions later in the call, the actions in the identification phase are relatively fluent. An exception is the presence of initial disfluencies for self-identification, but the frequency is not significantly different for declinations and acceptances (Panel A and Panel B). It is plausible that an initial disfluency for self-identification simply separates it from a preceding greeting in the same turn (as in Excerpt 1, line 4). In the identification phase, midstream disfluencies are relatively frequent for "institutional identification," an action that allows interviewers to choose components of the identification, but there are no significant differences between declinations and acceptances.

In the "purpose of the call" and "recruitment and persuasion" phases, initial and midstream disfluencies each appear to be similarly frequent for both call outcomes. However, the level of disfluency appears higher in declination than acceptance calls for

Table 6. Disfluencies in recruitment actions, for calls with actions of a given type, by call outcome.

		Panel A. Components of disfluency									Panel B. Summary measures		
		(a) Initial: mean of proportion of actions that begin with any disfluency			(b) Midstream: mean of proportion of actions with any midstream filler or broken-off talk			(c) Mean of proportion of actions with any disfluency in any location			(d) Mean of mean number of disfluencies in any location		
Phase	Recruitment Action	Declinations	Acceptances	p	Declinations	Acceptances	p	Declinations	Acceptances	p	Declinations	Acceptances	p
Identification	Greeting	0.08	0.03	***	0.01	0.02		0.08	0.04	**	0.08	0.04	**
	Self-identification	0.16	0.20		0.02	0.03		0.18	0.23		0.19	0.24	
	Institutional identification	0.05	0.06		0.11	0.13		0.14	0.18		0.21	0.24	
	Request to speak to sample member	0.08	0.13	*	0.08	0.08		0.14	0.19		0.19	0.28	
Purpose of call	Verification of sample member	0.38	0.43		0.27	0.21		0.51	0.57		0.77	0.95	
	Letter reference	0.30	0.32		0.17	0.14		0.39	0.38		0.61	0.63	
	Study reference	0.23	0.21		0.32	0.26		0.46	0.39		1.08	0.75	**

Table 6. Continued.

Phase	Recruitment Action	Panel A. Components of disfluency									Panel B. Summary measures		
		(a) Initial: mean of proportion of actions that begin with any disfluency			(b) Midstream: mean of proportion of actions with any midstream filler or broken-off talk			(c) Mean of proportion of actions with any disfluency in any location			(d) Mean of mean number of disfluencies in any location		
		Declinations	Acceptances	p	Declinations	Acceptances	p	Declinations	Acceptances	p	Declinations	Acceptances	p
Recruitment and persuasion	Request to participate	0.37	0.28	*	0.27	0.23		0.52	0.42	*	0.93	0.74	
	Follow-up to declination	0.22	0.15		0.21	0.49	*	0.35	0.51		0.65	1.18	
	Length-of-interview statement	0.41	0.36		0.44	0.44		0.65	0.63		1.48	1.28	

Note: Unit of analysis is calls in which an action occurred. Ns for cells are in Table 4, Panel B. See Table 3 for calculation of variables. Tests compare these means for the two call outcomes.

Cells for actions that are "incongruent" with the outcome are in gray font. Base analytic sample is matched pairs: 257 acceptances and 257 declinations. Interaction begins when sample member answers or is brought to the phone and ends when the sample member hangs up (declination) or the interview begins (acceptance).

*$p < 0.10$, **$p < 0.05$, ***$p < 0.01$.

study references (see summary in column (d)) and the request to participate (see summary in column (c)), although the difference is not always significant. Nevertheless, disfluency in these actions could have a cumulative effect on participation for sample members who have not exited. If we compare the two congruent actions in the recruitment phase – follow-up actions for declination calls and talk about the length of interview for acceptance calls – the latter seem to be more disfluent. However, as seen in Table 5 (column (c)), the average number of such actions that a sample member who accepts the request experiences (1.95) is fewer than the average number of follow-up actions for a sample member who declines (5.3).

The description provided in Table 6 supports our second prediction that most actions in the identification phase are more fluent than those in later phases. Perhaps surprisingly, in the recruitment phase, we do not observe that follow-up actions for declinations are more disfluent than the congruent actions for acceptances, so that our third prediction is not supported. However, our description identifies several components of the sample member's exposure to the interviewer that could be important to distinguish: Which actions occur, the number of times each action occurs, the number of words in the action, and whether the action is performed disfluently.

Table 6 suggests that when similar actions occur in declination and acceptance calls, they seem to have similar levels of disfluency, and the results of tests of differences are neither consistent nor strong. If we focus on the patterns in Table 6, we could say that for some actions – such as the request to speak to the sample member – there appear to be more disfluencies in acceptance calls; for others – notably the greeting, talk about the study, and the request to participate – disfluencies appear higher for declinations. However, only two of those differences are statistically significant. Table 6 also reinforces the observation based on Table 5 that the actions that distinguish declinations (e.g., early exits and follow-ups to declinations) and acceptances (statements about the length of the interview) make the interaction in calls that continue, in some ways, fundamentally incomparable.

6.3. Predicting Acceptance from Interviewers' Disfluencies, Words, and Actions

The structural dependencies among features of talk such as the number of actions, words, and disfluencies are reflected in correlations high enough that it is difficult to distinguish their relative contributions. For example, among the number of disfluencies, words, and actions, the correlations range from 0.70 to 0.92 considering all recruitment actions and all calls (detail not shown). In addition, all of these features can be viewed as indexing the length of the interaction, which is plausibly a product of the sample member's propensity to participate more than of the actions of the interviewer. To take the correlation between the number of words and disfluencies into account, we calculate a ratio of disfluencies per 100 words (without disfluencies); this measure is similar to the "filler rate" used by Bortfeld et al. (2001); see also Conrad et al. (2013). Table 7 presents the results of this analysis. To facilitate comparisons with previous studies, we first examine the disfluency ratio and then add the number of actions to control for the length of the interaction.

We find that when all recruitment actions are considered together, the disfluency ratio is not a significant predictor of acceptance. However, when we consider only actions up to the first turning point, there is a modest positive relationship between the disfluency ratio

Table 7. Multivariate and bivariate conditional logistic regressions of acceptance of request for participation on disfluencies, words, and actions, by location relative to first turning point.

Model	Recruitment actions	Measure	Odds ratio	*p* (two-tailed)	95% CI	
					Lower	Upper
Disfluency ratio only	All	Disfluencies/(words/100)	1.03	0.23	0.98	1.09
	Up to turning point	Disfluencies/(words/100)	1.06	0.03	1.00	1.11
	After turning point	Disfluencies/(words/100)	0.98	0.21	0.95	1.01
Disfluency ratio and actions	All	Disfluencies/(words/100)	1.04	0.19	0.98	1.09
		Number of actions	0.98	0.35	0.95	1.02
	Up to turning point	Disfluencies/(words/100)	1.03	0.32	0.97	1.09
		Number of actions	1.33	0.00	1.22	1.46
	After turning point	Disfluencies/(words/100)	1.00	0.96	0.97	1.04
		Number of actions	0.87	0.00	0.82	0.92
Disfluencies only	All	Number of disfluencies	1.00	0.83	0.98	1.02
	Up to turning point	Number of disfluencies	1.09	0.00	1.04	1.15
	After turning point	Number of disfluencies	0.96	0.03	0.92	1.00
Words only	All	Number of words/100	0.97	0.66	0.83	1.12
	Up to turning point	Number of words/100	4.94	0.00	2.91	8.37
	After turning point	Number of words/100	0.73	0.00	0.59	0.90
Actions only	All	Number of actions	0.99	0.43	0.95	1.02
	Up to turning point	Number of actions	1.34	0.00	1.23	1.46
	After turning point	Number of actions	0.87	0.00	0.82	0.92

Note: Unit of analysis is 257 matched pairs of acceptances and declinations. The interviewer's disfluencies, words, and actions are summed over all the recruitment actions in Table 1. For acceptances, the first turning point is the first request for participation or the last interviewer action before an acceptance. For declinations, the first turning point is the last interviewer action before the first (blocking) declination, hang up, or acceptance, whichever came first. Interaction begins when sample member answers or is brought to the phone and ends when the sample member hangs up (declination) or the interview begins (acceptance).

and acceptance ($p = 0.03$). When we add the number of actions to the models, the picture changes: The disfluency ratio no longer predicts acceptance. Instead, the number of actions before the turning point has a large positive effect on acceptance and the number after has a large negative relationship ($p = 0.00$ for both). A model that removes the structure built into a disfluency ratio by using the number of disfluencies and number of words as separate predictors shows the same result for the disfluencies (not shown).

To put these results in context, we also estimated bivariate models for each of the number of disfluencies, number of words, and number of actions (shown in Table 7). We find that for each of these three measures the relationship is null when all recruitment actions are pooled, positive for actions before the turning point, negative after. Each of these measures suggest that longer interactions before the first turning point predict acceptance, longer interactions after the first turning point predict declination.

7. Discussion

We provide a new, detailed description of how the actions of the interviewer in initial calls to recruit a sample member differ in calls that end in acceptance and declination. This description is similar in spirit to the discussion in Sturgis and Campanelli (1998), but we are able to provide more detail. This detail clarifies some of the challenges in studying how the interviewer affects participation. Our case study shows that taking the action structure of the call seriously affects conclusions: Calls that end in declinations and acceptances are most comparable only in the identification phase of the call because subsequently they consist of different actions. Levels of disfluency that occur later in the call originate in different actions, with different numbers of words, and different levels of fluency. These facts complicate the goal of predicting how the interviewer's disfluencies influence participation, because it is not clear that disfluencies that arise in different actions (e.g., scheduling interviews vs. responding to declinations) can be compared.

Our description and predictive analysis illustrate that the challenge of how to appropriately control for the different lengths and constituent actions of calls with different outcomes does not have a simple solution. This is not merely a technical issue – it potentially matters, for example, if the impact of the interviewer's talk is due to its length, its disfluency, or the actions in which the talk occurs; but these are structurally related and difficult to distinguish. We follow earlier analyses of disfluencies in considering the entire recruitment interaction in our predictive analyses (Conrad et al. 2013). However, we also compare actions up to (and including) the first turning point to those after, and control for the number of actions; we find that the turning point is important. If we predict participation, the disfluency ratio before the first turning point has a modest positive relationship with participation, but that estimate loses significance when the number of actions is included as a predictor. It is not clear whether these results reflect the results of actions by the interviewer or simply summarize a description of the call that is driven by the sample member's propensity to participate. In addition, even when we structure our analysis to consider the first turning point, we are counting disfluencies and words in different – and arguably incomparable – actions in calls with different outcomes. Thus, additional refinements to the analysis of actions are also needed to identify sites, if any, where fluency might be particularly crucial and useful to compare across outcomes.

Concepts related to, but different from, disfluencies as examined here include acoustic measures of fluency and ratings of perceived fluency (e.g., Sharf and Lehman 1984; Van der Vaart et al. 2006), following a script (e.g., Schaeffer et al. 2013), sounding scripted, or sounding "robotic" (e.g., Conrad et al. 2013). The operationalizations that accompany these various concepts include some elements not used here, such as pauses, re-starts, perceptions of listeners, and so forth. Continued work is needed to understand which of these related concepts and operationalizations, if any, enters the decisions of sample members. Understanding the impact of the fluency of the interviewer's talk is potentially important, if interviewers could be screened for or trained for fluency. The practice of giving interviewers an "agenda" rather than a script to use in recruitment (Houtkoop-Steenstra and Van den Bergh 2002; Morton-Williams 1993) may put more importance on the interviewer's ability to be fluent in a range of actions, both rote and improvised. If, for example, descriptions of the study that are more scripted and less disfluent are more effective, increasing fluency in this task could be a focus of interviewer training.

Prior studies provided a strong foundation for our operationalizations of disfluency, but we found that our interviews included complex combinations of fillers, nonpropositional elements, and other components that challenged our coding methods. It still seems possible to us that some types of disfluencies in some actions could reduce the sample member's likelihood of participating, perhaps because the disfluencies are irritating, or slow down the interviewer's delivery, or suggest incompetence. The disfluencies of the interviewer in Excerpt 2 for example, combined with his speed, certainly do not give a positive impression. However, the intuition formed by listening to such interactions has not yet led to a discovery of when and how fluency matters. We also were not able to code the great variety of complex midstream re-starts, which could be an important type of disfluency. Developing more sensitive, and potentially more informative, measures of disfluencies requires additional qualitative work.

We note that phone contacts recruiting sample members continue to be important in longitudinal and other list samples, and for a range of other purposes. By describing the actions of the interviewer in the recruitment call in more detail and differently than has been done to date (e.g., Conrad et al. 2013), we aim to deepen the way that we think about the interviewer's actions, how they depend on actions that came before, and what those dependencies imply for quantitative analyses that must summarize over such details. Our data are from a longitudinal study whose sample members are older, homogeneous in many ways, and contacted at a time when landlines were dominant, and so some details that we observe may be specific to our case. Our data were collected when landlines were still dominant, and although we believe that the trajectory of calls on cell phones differs from the trajectory we describe, we cannot say exactly how, and we are not likely to have a comparable collection of calls on cell phones anytime soon. In addition, we expect our approach to continue to be useful for analyzing actions and thinking about the challenges of determining the extent to which participation reflects the sample member's pre-existing propensity to participate versus the interviewer's action. In any case, the unusual combination of data sources (recordings, transcripts, case-control design, and participation as a criterion) provides a laboratory for exploring what such resources can teach us.

8. Appendix

Details about sample:

Schaeffer et al. (2013) give details about the sample, estimated propensity scores, and reliability of identification of actions. The 1964 data collection had an 87% response rate (http://www.ssc.wisc.edu/wlsresearch/documentation/retention/cor1004_retention.pdf). All interviews were conducted in English at whatever telephone number (usually a landline) the sample member provided. The impact of clustering within interviewer is limited by the large number of interviewers in our analytic sample compared to the number of sample members. We have 138 interviewers, and the mean number of cases per interviewer is about 3.7 for both acceptances and declinations. Analytically, we expect that interviewer effects would be conveyed primarily via the interviewer's actions, actions that are usually unobserved but that we are able to measure. In 135 of the calls in the full analytic sample of 514 cases a third party answers the telephone and calls the sample member to the telephone.

Operationalization of disfluencies:

Our operationalization of disfluencies was adapted from that of Bortfeld et al. (2001) who provide a detailed description of their method. Some of their procedures were more complex and detailed than those we had resources to implement, and so we made some adaptations and simplifications. Coding was done using string functions in Stata, supplemented by review of cases that did not match the coding rules; because we relied on machine coding, we do not estimate reliability.

Bortfeld et al. developed a complex system for identifying the location of fillers that included "phrase-internal fillers" and "between-phrase" fillers. We use a simpler system that builds on our analysis of actions to distinguish three locations for disfluencies:

1. Beginning of turn (i.e., before the first action in the turn),
2. At the beginning of a second or higher-order action in a turn. Fillers that come between actions are allocated to the later action, and
3. Within an action (including at the end of the last action in a turn).

Our counting rules were adopted or adapted from Bortfeld et al. (2001, 131–133). They code these for turns; we code for actions.

1. When one disfluency followed another (e.g., "um um" or "um uh" or "um I- ah") each was counted as a disfluency,
2. A filler was counted as beginning an action if it immediately preceded the action (whether or not the filler was itself immediately preceded by a "nonpropositional element"), and
3. Bortfeld et al. ignored "nonpropositional elements" in determining whether a filler began a turn. Our data included complex strings that combined fillers and nonpropositional elements before actions in complex ways (e.g., "and uh um"). We proceeded in these ways:
   a. In computing total initial fillers, we counted just the fillers that immediately preceded an action and were not interrupted by other elements (e.g., "um uh ACTION"),
   b. In computing total initial disfluencies, nonpropositional elements that immediately

preceded the fillers (see a) were also counted as disfluencies (e.g., "okay and um uh ACTION"),

c. Broken-off talk immediately preceding an action is counted as a initial disfluency, and it is counted when computing total disfluencies (e.g., "um uh broken-off ACTION"), and
d. Midstream (mid-action) disfluencies include fillers and broken-off talk (e.g., "ACTION-begins um broken-off uh ACTION-continues")

Bortfeld et al. included the following in their word counts: "fillers, word fragments, and other words implicated in repeats and restarts." We did not include fillers or broken-off talk in our word count. We included nonpropositional elements when they were not counted as fillers. Thus, we examined two word counts: One that excluded fillers and broken-off talk, and one that excluded fillers, broken-off talk, and nonpropositional elements that were part of a string of fillers. We report analyses with the second.

When we consider all actions together, we combine actions that are first with those later in a turn.

9. References

Arnold, J.E., C.L.H. Kam, and M.K. Tanenhaus. 2007. "If You Say Thee uh You Are Describing Something Hard: The On-Line Attribution of Disfluency During Reference Comprehension." *Journal of Experimental Psychology: Learning, Memory, and Cognition* 33(5): 914–930. DOI: https://doi.org/10.1037/0278-7393.33.5.914.

Arnold, J.E., M.K. Tanenhaus, R.J. Altmann, and M. Fagnano. 2004. "The Old and Thee, uh, New: Disfluency and Reference Resolution." *Psychological Science* 15(9): 578–582. DOI: https://doi.org/10.1111/j.0956-7976.2004.00723.x.

Barr, D.J. and M. Seyfeddinipur. 2010. "The role of fillers in listener attributions for speaker disfluency." *Language and Cognitive Processes* 25(4): 441–455. DOI: https://doi.org/10.1080/01690960903047122.

Boomer, D.S. 1965. "Hesitation and Grammatical Encoding." *Language and Speech* 8(3): 148–158. DOI: https://doi.org/10.1177%2F002383096500800302.

Bortfeld, H., S.D. Leon, J.E. Bloom, M.F. Schober, and S.E. Brennan. 2001. "Disfluency Rates in Conversation: Effects of Age, Relationship, Topic, Role, and Gender." *Language and Speech* 44(2): 123–149. DOI: https://doi.org/10.1177%2F00238309010440020101.

Brennan, S.E. and M.F. Schober. 2001. "How listeners compensate for disfluencies in spontaneous speech." *Journal of Memory and Language* 44: 274–296. DOI: https://doi.org/10.1006/jmla.2000.2753.

Chamberlain, G. 1980. "Analysis of Covariance with Qualitative Data." *The Review of Economic Studies* 47(1): 225–238. DOI: https://doi.org/10.2307/2297110.

Christenfeld, N. 1995. "Does it hurt to say um?" *Journal of Nonverbal Behavior* 19(3): 171–186. DOI: https://doi.org/10.1007/BF02175503.

Clark, H.H. and J.E. Fox Tree. 2002. "Using uh and um in spontaneous speaking." *Cognition* 84(1): 73–111. DOI: https://doi.org/10.1016/S0010-0277(02)00017-3.

Conrad, F.G., J.S. Broome, J.R. Benkí, F. Kreuter, R.M. Groves, D. Vannette, and C. McClain. 2013. "Interviewer speech and the success of survey invitations." *Journal*

of the Royal Statistical Society: Series A (Statistics in Society) 176(1): 191–210. DOI: https://doi.org/10.1111/j.1467-985X.2012.01064.x.

Dijkstra, W. and J. Smit. 2002. "Persuading Reluctant Recipients in Telephone Surveys." In *Survey Nonresponse*, edited by R.M. Groves, D.A. Dillman, J.L. Eltinge, and R.A. Little, 121–134. New York: John Wiley & Sons.

Draisma, S. and W. Dijkstra. 2004. "Response Latency and (Para)linguistic Expression as Indicators of Response Error." In *Methods for Testing and Evaluating Survey Questionnaires*, edited by S. Presser, J.M. Rothgeb, M.P. Couper, J.T. Lessler, E. Martin, J. Martin, and E. Singer, 131–148. New York: Springer-Verlag.

Draisma, S., Y. Ongena, and W. Dijkstra. 2005. "Qualified Answers and Other Doubt Expressions as Indicators of Cognitive Problems in a Health Survey." American Association for Public Opinion Research Conference, Miami Beach, FL, May 2005. Available at: http://www.asasrms.org/Proceedings/y2005f.html (accessed June 2019).

Fox Tree, J.E. 2002. "Interpreting Pauses and Ums at Turn Exchanges." *Discourse Processes* 34(1): 37–55. DOI: https://doi.org/10.1207/S15326950DP3401_2.

Fox Tree, J.E. 2007. "Folk notions of um and uh, you know, and like." *Text & Talk – An Interdisciplinary Journal of Language, Discourse Communication Studies* 23(3): 297–314. DOI: https://doi.org/10.1515/TEXT.2007.012.

Garbarski, D., N.C. Schaeffer, and J. Dykema. 2011. "Are Interactional Behaviors Exhibited When the Self-Reported Health Question Is Asked Associated with Health Status?" *Social Science Research* 40(4): 1025–1036. DOI: https://doi.org/10.1016/j.ssresearch.2011.04.002.

Groves, R.M. and M.P. Couper. 1996. "Contact-Level Influences on Cooperation in Face-to-Face Surveys." *Journal of Official Statistics* 12(1): 63–83. Available at: https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/contact-level-influences-on-cooperation-in-face-to-face-surveys.pdf (accessed June 2019).

Groves, R.M. and K. McGonagle. 2001. "A Theory-Guided Interviewer Training Protocol Regarding Survey Participation." *Journal of Official Statistics* 17(2): 249–266. Available at: https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/a-theory-guided-interviewer-training-protocol-regardingsurvey-participation.pdf (accessed June 2019).

Groves, R.M., B.C. O'Hare, D. Gould-Smith, J.R. Benkí, and P. Maher. 2008. "Telephone Interviewer Voice Characteristics and the Survey Participation Decision." In *Advances in Telephone Survey Methodology*, edited by J.M. Lepkowski, C. Tucker, J.M. Brick, E.D. De Leeuw, L. Japec, P.J. Lavrakas, M.W. Link, and R.L. Sangster, 385–400. New Jersey: John Wiley & Sons.

Hauser, R.M. 2005. "Survey Response in the Long Run: The Wisconsin Longitudinal Study." *Field Methods* 17(1): 3–29. DOI: https://doi.org/10.1177%2F1525822X04272452.

Houtkoop-Steenstra, H. and H. van den Bergh. 2002. "Effects of Introductions in Large-Scale Telephone Survey Interviews." In *Standardization and Tacit Knowledge: Interaction and Practice in the Survey Interview,* edited by D.W. Maynard, H. Houtkoop-Steenstra, N.C. Schaeffer, and J. van der Zouwen, 205–218. New York: Wiley.

Mathiowetz, N.A. 1999. "Respondent Uncertainty as Indicator of Response Quality." *International Journal of Public Opinion Research* 11(3): 289–296. DOI: https://doi.org/10.1093/ijpor/11.3.289.

Maynard, D.W., J. Freese, and N.C. Schaeffer. 2010. "Calling for Participation: Requests, Blocking Moves, and Rational (Inter)action in Survey Introductions." *American Sociological Review* 75(5): 791–814. DOI: https://doi.org/10.1177%2F0003122410379582.

Maynard, D.W. and N.C. Schaeffer. 1997. "Keeping the Gate: Declinations of the Request to Participate in a Telephone Survey Interview." *Sociological Methods & Research* 26(1): 34–79. DOI: https://doi.org/10.1177%2F0049124197026001002.

Maynard, D.W., H. Houtkoop-Steenstra, N.C. Schaffer, and J. van der Zouwen. 2002. *Standardization and Tacit Knowledge: Interaction and Practice in the Survey Interview*, edited by D.W. Maynard, H. Houtkoop-Steenstra, N.C. Schaeffer, and J. van der Zouwen. New York: Wiley.

Morton-Williams, J. 1993. *Interviewer Approaches*. Aldershot, England: Dartmouth Publishing.

Oksenberg, L. and C.F. Cannell. 1988. "Effects of Interviewer Vocal Characteristics on Nonresponse." In *Telephone Survey Methodology*, edited by R.M. Groves, P.P. Biemer, L.E. Lyberg, J.T. Massey, W.L. Nicholls II, and J. Waksberg, 257–272. New York: John Wiley & Sons.

Schaeffer, N.C. forthcoming. "Interaction before and during the survey interview: Insights from conversation analysis." *International Journal of Social Research Methodology*.

Schaeffer, N.C. and J. Dykema. 2011. "Response 1 to Fowler's Chapter: Coding the Behavior of Interviewers and Respondents to Evaluate Survey Questions." In *Question Evaluation Methods: Contributing to the Science of Data Quality*, edited by J. Madans, K. Miller, A. Maitland, and G. Willis, 23–39. Hoboken, NJ: John Wiley & Sons, Inc.

Schaeffer, N.C., D. Garbarski, J. Freese, and D.W. Maynard. 2013. "An Interactional Model of the Call for Participation in the Survey Interview: Actions and Reactions in the Survey Recruitment Call." *Public Opinion Quarterly* 77(1): 323–351. DOI: https://doi.org/10.1093/poq/nft006.

Schaeffer, N.C. and D.W. Maynard. 2002. "Occasions for Intervention: Interactional Resources for Comprehension in Standardized Survey Interviews." In *Standardization and Tacit Knowledge: Interaction and Practice in the Survey Interview*, edited by D.W. Maynard, H. Houtkoop-Steenstra, N.C. Schaeffer, and J. van der Zouwen, 261–280. New York: Wiley.

Schaeffer, N.C., B.H. Min, T. Purnell, D. Garbarski, and J. Dykema. 2018. "Greeting and Response: Predicting Participation from the Call Opening?" *Journal of Survey Statistics and Methodology* 1(1): 122–148. DOI: https://doi.org/10.1093/jssam/smx014.

Schegloff, E. 2010. "Some Other "Uh(m)"s." *Discourse Processes* 47(2): 130–174. DOI: https://doi.org/10.1080/01638530903223380.

Schober, M.F. and J.E. Bloom. 2004. "Discourse Cues that Respondents Have Misunderstood Survey Questions." *Discourse Processes* 38(3): 287–308. DOI: https://doi.org/10.1207/s15326950dp3803_1.

Schober, M.F., F.G. Conrad, W. Dijkstra, and Y.P. Ongena. 2012. "Disfluencies and Gaze Aversion in Unreliable Responses to Survey Questions." *Journal of Official Statistics*

28(4): 555–582. Available at: https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/disfluencies-and-gaze-aversion-in-unreliable-responses-to-survey-questions.pdf (accessed June 2019).

Sharf, D.J. and M.E. Lehman. 1984. "Relationship between the speech characteristics and effectiveness of telephone interviewers." *Journal of Phonetics* 12(3): 219–228. Abstract: https://psycnet.apa.org/record/1985-22253-001 (accessed June 2019).

Shriberg, E. 1996. "Disfluencies in Switchboard." Proceedings, International Conference on Spoken Language Processing (ICSLP '96), Vol. Addendum, 11–14. Philadelphia, PA, October 3–6, 1996. Available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.16.5822&rep=rep1&type=pdf (accessed June 2019).

Smith, V.L. and H.H. Clark. 1993. "On the Course of Answering Questions." *Journal of Memory and Language* 32(1): 25–38. DOI: https://doi.org/10.1006/jmla.1993.1002.

Sturgis, P. and P. Campanelli. 1998. "The Scope for Reducing Refusals in Household Surveys: An Investigation Based on Transcripts of Tape-Recorded Doorstep Interactions." *Journal of the Market Research Society* 40(2): 121–139. Available at: https://search.proquest.com/docview/214805239/fulltextPDF/7DBB4A59542349A9PQ/1?accountid=465 (accessed June 2019).

Van der Vaart, W., Y. Ongena, A. Hoogendoorn, and W. Dijkstra. 2006. "Do Interviewers' Voice Characteristics Influence Cooperation Rates in Telephone Surveys?" *International Journal of Public Opinion Research* 18(4): 488–499. DOI: https://doi.org/10.1093/ijpor/edh117.

Received September 2018
Revised June 2019
Accepted October 2019

sciendo

Journal of Official Statistics, Vol. 36, No. 3, 2020, pp. 561–588, http://dx.doi.org/10.2478/JOS-2020-0029

Measurement of Interviewer Workload within the Survey and an Exploration of Workload Effects on Interviewers' Field Efforts and Performance

Celine Wuyts[1] *and Geert Loosveldt*[1]

Interviewer characteristics are usually assumed fixed over the fieldwork period. The number of sample units that require the interviewers' attention, however, can vary strongly over the fieldwork period. Different workload levels produce different constraints on the time interviewers have available to contact, recruit and interview each target respondent, and may also induce different motivational effects on interviewers' behavior as they perform their different tasks. In this article we show that fine-grained, time-varying operationalizations of project-specific workload can be useful to explain differences in interviewers' field efforts and achieved response outcomes over the fieldwork period. We derive project-specific workload for each interviewer on each day of fieldwork in two rounds of the European Social Survey in Belgium from contact history and assignment paradata. Project-specific workload is measured as (1) the number of sample units which have been and remain assigned on any day t (assigned case workload), and (2) the number of sample units for which interviewer activity has started and not yet ceased on any day t (active case workload). Capturing temporal variation in interviewers' workloads in a direct way, the time-varying operationalizations, are better predictors than are the interviewer-level operationalizations of typical (active or potential) workload that are derived from them, as well as the traditional total-count workload operationalization.

Key words: Nonresponse; interviewer effort; interviewer effects; time-varying interviewer characteristics; paradata.

1. Introduction

Interviewer workload is widely believed to be a constraining factor on the effort survey interviewers are able and willing to apply in the face-to-face recruitment of respondents, but the available empirical evidence in support of these beliefs is sparse. This article addresses an existing gap in the literature on interviewer workload as a source of survey interviewers' differential performance in the contact and recruitment task. The interviewers' performance in the contact and recruitment task, that is, the extent to which the demands of this task are adequately met, can be evaluated from a data quality perspective in terms of the interviewers' contribution to nonresponse error, which in turn depends on the interviewers' effort in applying the contact procedures (e.g., number and timing of contact attempts) (Loosveldt et al. 2004).

[1] Centre for Sociological Research, Catholic University of Leuven, Parkstraat 45 - bus 3601, 3000 Leuven, Belgium. Emails: celine.wuyts@kuleuven.be and geert.loosveldt@kuleuven.be

We focus specifically on interviewer workload in one particular survey project and will refer to this workload component as 'project-specific workload'. We do not elaborate on work resulting from activities and responsibilities other than involvement in the survey under study (e.g., involvement in other survey projects and other jobs) even though all work components may combine to constrain time availability and alter interviewers' incentives and motivations. Project-specific workloads can be controlled, if not fully determined, by the field supervisors, within the constraints imposed by the survey design and the available interviewer workforce.

The total amount of work involved in recruiting and interviewing respondents in any particular survey is driven by survey design features such as the mode, the contact procedures and the sampling design, and characteristics of the target population such as its geographical distribution and accessibility. The capacity of the interviewer workforce as a whole to deliver the desired results is determined by the number of interviewers available, the interviewers' temporal and (in the case of face-to-face surveys) geographical distribution, and the interviewers' experience and skill. The size and composition of the available interviewer workforce, relative to the size and composition of the (gross) sample size, places constraints on the way the work is allocated.

Project-specific interviewer workloads are conventionally quantified in terms of numbers of sample units, that is, the unit of measurement is the sample unit, and result from an intricate allocation process. Sample units may be (initially) allocated all at once at the start of the fieldwork, or in several batches over the fieldwork period. After the initial allocation, sample units may also be reallocated at some point. This may be done as a deliberate strategy to improve response rates as well as in response to interviewers dropping out of the interviewer workforce. Reallocation as a response enhancement strategy is premised on the recognition that even if (repeated) attempts to make contact and obtain cooperation by one interviewer are unsuccessful, additional attempts by a different interviewer may still yield a completed interview (e.g., Calderwood et al. 2016). Interviewers are usually assigned more sample units, both initially and in view of response enhancement, if they are expected to be more successful based on past performance, and if they reside in areas with high target population density relative to interviewer availability.

Different workload levels produce different constraints on the time interviewers have available to contact, recruit and interview each target respondent, and induce interviewers to adjust their behavior accordingly. Interviewers have limited available time and energy to allocate over the sample units assigned to them. Larger workloads imply that each case in an interviewer's workload on average receives a smaller share of the interviewer's total available time, and in particular time available during the most productive days (weekends) and on the most productive hours (evenings) (Botman and Thornberry 1992; Groves and Couper 1998, 274). The attempts that are made to contact and recruit each target respondent are likely fewer in number and less optimally timed if workloads are large (Japec 2008). Given that (face-to-face) interviewers typically have full discretion in dividing their time and energy among the sample units assigned to them, and especially under piece-rate payment schemes, whereby interviewers are paid per completed interview, larger workloads are likely to induce interviewers to focus their attention to easily reached and cooperative target respondents. Interviewers may make fewer

round-trips to relatively sparsely populated areas, disregard when target respondents are likely to be at home, and push for a quick decision at the doorstep, rather than carefully planning visits, tailoring and maintaining the interaction in a single or over multiple visits. Geographical clustering of sample units and scale efficiencies in travel time may encourage interviewers to make more round-trips to distant geographical areas when workloads for those areas are large, and induce them to give up on these areas entirely when workloads are small.

The idea of interviewers taking shortcuts when they experience tasks as burdensome, for example because of a heavy workload, was advanced by Japec (2005; 2008). These shortcuts are generally intended to reduce the amount of time and effort invested for each individual case below the level corresponding to the researchers' data quality targets. Japec (2005; 2008) therefore uses the term 'interviewer satisficing'. Interviewer satisficing behaviors in contact and recruitment put additional pressure on the trade-off between achieving response rates that are acceptable from a quality perspective and the cost of extending and/or intensifying fieldwork operations.

Workload-related design choices and allocation procedures may therefore affect the survey data collection process and the quality of the collected data. Investigations into the nature and impacts of interviewer workload should be highly relevant to survey practitioners but are surprisingly sparsely documented. Most of the available literature on explaining differential interviewer performance, whether in the contact and recruitment task or in the interview administration task, focuses on sociodemographics and interviewer characteristics related to experience, skill and confidence (West and Blom 2016), rather than motivation and effort or their determinants. The limited available evidence on the workload-performance relationship in the contact and recruitment task of survey interviewers is not only sparse but the usual approach to measuring workload has also not been altogether convincing and invites further research.

In the current article, we examine the association between interviewers' project-specific workload on the one hand and expended efforts and achieved response rates on the other. Contact history and assignment paradata for two rounds of the European Social Survey in Belgium allow the construction of more fine-grained measurements of project-specific workload over the fieldwork period than traditionally have been used. Project-specific workload and other interviewer characteristics (see e.g., West and Blom 2016 for a recent overview) are usually considered as fixed over the fieldwork period. As will be discussed in the following section, valuable insights into the underlying mechanisms may be gained by observing project-specific workload as a time-varying interviewer characteristic.

The following section presents some important considerations in the measurement of project-specific workload, summarizes the available literature on the link between interviewers' project-specific workloads and response rates and elaborates on the limitations of the traditional approach to workload measurement, before introducing an alternative perspective and associated alternative operationalizations. In particular, we advance two approaches to measure interviewer workload on each day in the fieldwork: (1) the number of sample units which have been and remain assigned on any day t (assigned case workload), and (2) the number of sample units for which interviewer activity has started and not yet ceased on any day t (active case workload).

2. Measurement of Project-Specific Workload

A common approach is to measure project-specific workload as the total number of sample units worked on by an interviewer for the survey project (e.g., Blom 2012; Beullens et al. 2016). Beullens et al. (2016), for example, observe that interviewers in Round 7 of the European Social Survey worked on average between 12 (Czech Republic) and 53 (Switzerland) sample units over the course of the fieldwork. This is, however, only one of multiple possible approaches.

There are two key measurement issues to consider: (1) which sample units to include, and (2) whether and how temporal variation over the fieldwork period is taken into account. With regard to the first question, two approaches appear reasonable. The first approach would be to include sample units if they are assigned to the interviewer and thus could potentially be pursued by the interviewer. The second approach would be to include sample units only if they actually are actively being pursued by the interviewer.

Both the number of sample units that have been and remain assigned and the number of sample units that are actually being pursued may vary over time. Many sample units naturally cease to entail any additional work at some point, because an interview has been completed, ineligibility has been established, or the sample unit is abandoned as unproductive. The remaining number of sample units that require any attention therefore (more or less gradually) shrinks over time, at least until additional sample units are assigned. More erratic fluctuations may occur especially when sample units are assigned in batches or reallocated over the fieldwork period. Whether and how these changes are taken into account in measuring project-specific workload constitutes an important measurement decision.

Only a handful of empirical studies have reported estimates of correlations between interviewers' project-specific workloads and response rates, usually with workload as one of many interviewer characteristics under consideration. None have taken into account temporal variation or investigated the effect of interviewer workload on nonresponse error in a comprehensive way.

An early mention of workload as an interviewer characteristic was made by Singer et al. (1983). Singer et al. (1983) observed a negative correlation between interviewer workload (operationalized as the total number of cases assigned to each interviewer) and response rates in a random digit dialing, RDD telephone survey. The observed screening rate dropped from 91% for interviewers with fewer than 30 cases to 84% for interviewers who were assigned a total of 88 or more cases. The cooperation rate dropped from 78% to 60%. They offer workload fatigue or burden as one possible explanation, but also acknowledge that the result may be an artefact of the case assignment procedure. Some interviewers were assigned large numbers of cases at the end of fieldwork and these cases may have been more difficult and/or may not have been pursued as intensively due to lack of time.

Nicoletti and Buck (2004) reported that interviewer workload (operationalized as the total number of households assigned to each interviewer) was negatively associated with the probability of contact in two of four household panel surveys studied and with the probability of cooperation, conditional on contact, in three of the surveys. For the German Socio-Economic Panel survey, they observe the expected negative association between

interviewer workload and the probability of cooperation but a *positive* association for the probability of contact. A positive association between interviewer workload (operationalized as the total number of previous-wave respondents assigned at the start of the fieldwork) and contact probability in a panel survey (at least for reasonably sized workloads of fewer than 124 cases) has also been reported by Watson and Wooden (2009). While claiming that large workloads may in fact be beneficial for response, they do admit that the observed pattern may be an artefact of the case assignment procedure. Larger numbers of cases had commonly been assigned to the best interviewers.

Blom et al. (2011), on the other hand, found no evidence in support of the hypothesized negative effect of interviewer workload (operationalized as the total number of cases worked on over the course of the fieldwork) on contact and cooperation rates in seven countries in the first round of the European Social Survey. Blom (2012), examining a partially different set of seven countries, likewise did not observe the expected association, except for one country. She also suggests that in most countries larger numbers of cases may have been assigned to the best interviewers, counterbalancing a potential negative effect of workload on response rates.

In all of these studies, project-specific workload has been operationalized by total-count measures at the interviewer level, that is, the total number of sample units assigned to each interviewer or the total number of sample units worked on by each interviewer. Total-count measures at the interviewer level are usually readily available or easy to derive but have their limitations. The main limitation of using a total-count measure to explain interviewers' differential task performance, expressed amongst others by Blom (2012), is the risk of bias due to reverse causality. Concerns about interviewers' performance in the survey project driving their total case workload are warranted. In addition, by adopting a total-count measure, one makes abstraction of any temporal variation in the number of sample units that require the interviewers' attention over the fieldwork period. Note that because interviewers rarely leave sample units that have been assigned to them completely untouched over the entire fieldwork period, the question of which sample units to include should not be of great importance for this type of operationalization. Any discrepancies between the number of sample units assigned and the number of sample units actively worked on will be minor, and negligible in practice.

As previously discussed, temporal variation results from particularities of the allocation process and the natural progression of fieldwork. At any moment in time, we would expect interviewers to adjust their behavior in response to the number of sample units that require their attention at that time. We therefore propose that operationalizations of project-specific workload that take temporal variation over the fieldwork into account are more appropriate to explain interviewers' fieldwork effort and task performance.

A useful strategy would thus be to measure project-specific workload at different points in time over the fieldwork period. Time-varying operationalizations of project-specific workload capture temporal variation in a direct way. We would also argue that by observing variability in project-specific workload over time in addition to variability in project-specific workload across interviewers, the risk of bias due to reverse causality between workload and task performance may be partially mitigated. At a particular point in time, it seems much more likely that the interviewers' workload affects their expended effort and achieved response rates at that time than the other way around.

The question of which sample units to include at each time point remains to be answered. The inclusion of sample units hinges on the identification of the moments in time at which sample units can be presumed to enter and leave the interviewers' workloads. We previously presented two reasonable approaches. The first approach would consider a sample unit as being in an interviewer's workload as long as the interviewer could potentially pursue the unit in question. Each sample unit is assumed to enter the interviewer's workload when the interviewer is assigned the sample unit and leaves the interviewer's workload when the sample unit is returned to the field office. Applying this approach, 'assigned' case workload at time t is measured as the number of sample units that have been and remain assigned at that time. The second approach would consider a sample unit as being in an interviewer's workload as long as the interviewer is actively pursuing the unit in question. Each sample unit is assumed to enter the interviewer's workload when the interviewer first attempts to contact the sample unit and leaves the interviewer's workload after the last attempt. Applying this second approach, 'active' case workload at time t is measured as the number of sample units for which interviewer activity has started and not yet ceased at that time. Figure 1 illustrates the application of the two approaches for a hypothetical interviewer with a total case workload of four sample units. Sample unit 1 leaves both active and assigned case workload after the completed interview. Sample units 2, 3 and 4 leave assigned case workload after being returned to the field office (at day 9), but leave active case workload after the last unsuccessful attempt (at day 7 for sample units 2 and 3, at day 8 for sample unit 4). Note that for these time-varying operationalizations, the choice of which sample units to include is actually highly relevant. Discrepancies between the number of sample units that have been and remain assigned and the number of sample units actively worked on may be considerable.

The first operationalization of *assigned* case workload is conceptually superior to the operationalization of *active* case workload in that it better captures the amount of work interviewers are expected to do, rather than just the amount of work interviewers are currently engaged with in the field. Cases that are assigned but not (yet) actively engaged with may carry comparable weight because such cases still involve some planning and

Time	1	2	3	4	5	6	7	8	9	10
SU1	A		NR	I						
SU2	A	NR	NR				NR		R	
SU3						A	NR		R	
SU4						A	NR	NR	R	
Assigned case workload	2	2	2	2	1	3	3	3	3	0
Active case workload	0	1	2	2	1	1	3	1	0	0

Note: A = sample unit is assigned; R = sample unit is returned to the field office; NR = unsuccessful contact attempt (no interview); I = interview.

Fig. 1. Illustration of two approaches to measuring project-specific workload over time.

administrative work. In addition, the *active* case workload operationalization is endogenous to interviewers' field efforts. Active case workload will be lower if assigned sample units are being shelved or abandoned, and higher when more sample units are being attempted. As this operationalization itself depends heavily on actual interviewer activity, it may not be particularly well suited to assess workload effects on field efforts and outcomes. Estimated "effects", especially with respect to interviewers' field efforts, should be interpreted with caution. Because *assigned* case workload depends primarily on the survey agency's case assignment practices, and is less sensitive to actual interviewer activity, causal interpretations of estimated workload effects should be less questionable for this operationalization.

From the two time-varying operationalizations of project-specific workload, we can also derive operationalizations at the interviewer level. A (temporally weighted) average (active/assigned) case workload better approximates workload on any one typical fieldwork day than the traditional total-count measures at the interviewer level.

3. Study Objectives

The main objective of this study is to develop and assess alternative project-specific workload operationalizations at two levels of aggregation. We adopt new operationalizations expressing *time-varying* interviewer characteristics (at the level of interviewer days) and operationalizations derived thereof and expressing *fixed* interviewer characteristics (at the interviewer level).

Our conception of project-specific workload measured at different points in time over the fieldwork period is inspired by, and analogous to, the conception of within-survey experience measured at each interview over the fieldwork period. The introduction of *time-varying* interviewer characteristics to explain changes in interviewer behavior and performance can be attributed to Olson and Peytchev (2007), who proposed using interview order to capture interviewers' increasing familiarity with the survey instrument over the fieldwork period. Whereas interview order has become the standard operationalization of interviewers' within-survey experience (e.g., Loosveldt and Beullens 2013; Kirchner and Olson 2017), most other interviewer characteristics studied in the survey research literature remain defined and operationalized at the interviewer level. Many interviewer characteristics can indeed be assumed sufficiently stable over the fieldwork period. This assumption is plausible for interviewer characteristics such as age, gender, race and personality traits, and likely for many attitudes as well, but not for the amount of work interviewers are expected to do.

The secondary objective is to apply these different operationalizations in an exploration of whether the commonly held beliefs about large interviewer workloads negatively affecting fieldwork outcomes in face-to-face surveys are legitimate. Since face-to-face contact and recruitment demand a large share of the total time and effort expended by the interviewers, we may expect a negative effect of workload on the number of personal visits made to each individual case in their workload because of limited time availability. Both contact and cooperation rates may to some extent be affected by large workloads implying that relatively fewer visits are made during the more productive hours, and cooperation rates may be further reduced by less effortful doorstep interactions. We therefore will

evaluate the validity of the following two assertions, using the alternative workload operationalizations.

Hypothesis 1: When interviewers carry larger workloads, their contact effort per case (in quantitative terms) is reduced.

Hypothesis 2: When interviewers carry larger workloads, their contact and cooperation rates are reduced.

4. Data and Methods

We use data from Round 6 and Round 7 of the European Social Survey (ESS) in Belgium (European Social Survey 2014, 2016). The units in the Belgian gross samples, individual persons drawn from the National Register, were assigned to individual interviewers in several batches over the fieldwork period, and relatively large numbers of non-responding sample units (both initial non-contacts and initial refusals) were reassigned to other interviewers. Such a fieldwork strategy complicates the operationalization of project-specific workload over the fieldwork period, compared to a fieldwork strategy with all units allocated only once at the start of fieldwork. The ESS in Belgium therefore represents an attractive case for the current study.

In Round 6 (ESS6-BE) 3,267 sample units were issued (of which 622 initial nonrespondents were reassigned and worked on by another interviewer), 155 interviewers were employed and 1,869 interviews were administered (response rate 58.7%) between September 2012 and the end of December 2012. In Round 7 (ESS7-BE 2014-2015) 3,204 sample units were issued (of which 1,041 initial nonrespondents were reassigned and worked on by another interviewer), 151 interviewers were employed and 1,769 interviews were administered (response rate 57.0%) between September 2014 and the end of January 2015. The same survey agency was contracted and the two groups of interviewers partially overlap. The interviewers were paid per completed interview, with the piece rate itself adjusted if an interviewer exceeded a threshold response rate, administered interviews and collected paradata of sufficiently high quality, and completed his or her assignments on time.

We construct several measurements of the main interviewer characteristic of interest, project-specific workload, from the publicly available ESS contact history data ("contact forms data") and additional interviewer assignment data. The ESS contact history data set contains a detailed record of when sample units were attempted, by which interviewer, by which mode and with which outcome. The interviewer assignment data set on which we can draw for these two survey rounds of the ESS in Belgium contains a record of when sample units were assigned to which interviewer.

The following two subsections offer a detailed description of the operationalizations at the level of interviewer-days (workload as a time-varying interviewer characteristic) and at the interviewer level (workload as a fixed interviewer characteristic), respectively.

4.1. Operationalizing Project-Specific Workload at the Level of Interviewer-Days

We derive daily project-specific workload measures along the lines of the two approaches described above (active case workload and assigned case workload), and by taking into account reassigned sample units as well as sample units initially assigned at the start or over

the course of the fieldwork ($N = 3{,}889$ in Round 6 and $N = 4{,}245$ in Round 7). The number of sample units for which interviewer activity has started and not yet ceased (active case workload) can be derived for each interviewer and on each fieldwork day from the contact history data. Each sample unit enters an interviewer's *active* case workload on the date of the first recorded attempt and leaves the interviewer's active case workload on the date of the last recorded attempt by that interviewer. The number of sample units that have been and remain assigned (assigned case workload) could similarly be derived for each interviewer and each fieldwork day from detailed assignment data. Each sample unit enters an interviewer's *assigned* case workload on the recorded date of assignment, unless the date of assignment is missing ($N = 2$ in Round 6, $N = 6$ in Round 7) or follows rather than precedes the first contact attempt ($N = 83$ in Round 6, $N = 97$ in Round 7), in which case the date of the first recorded attempt is taken. Ideally, the interviewer assignment data set would include, for each sample unit that is assigned to a particular interviewer, not only the date of assignment but also the date of return to the field office. Because data on case returns to the field office are not available in this study, each sample unit is assumed to leave the interviewer's assigned case workload on the date of the last recorded attempt in the assignment set (i.e., the group of sample units assigned on the same date and in the same geographical area). Note that assignments of sample units to interviewers for which no attempt is recorded in the contact history data are therefore necessarily excluded here (4% and 7% of assignments in Round 6 and Round 7, respectively). Respondents leave the interviewer's workload on the date of the completed interview. Sample units identified as deceased also leave the interviewer's workload. Because under the Belgian ESS fieldwork strategy each sample unit is assigned to a single interviewer at any one point in time, we should not observe any sample units simultaneously counted in multiple interviewers' workloads. We observe only 0.1% and 2.1% of sample units simultaneously counted in multiple interviewers' assigned case workloads in Rounds 6 and 7, respectively. Aggregating the number of sample units that have entered, and have not yet left, the (active or assigned) workload of the interviewer at time t yields the interviewer's (active or assigned) project-specific workload at that time t.

Using these two time-varying operationalizations of project-specific workload, we first descriptively examine the degree to which temporal variation in project-specific workload over the fieldwork period actually occurs. Figure 2 provides an illustration of an interviewer's project-specific workload trajectory over the fieldwork period according to the two approaches. The interviewer presented was assigned a total of 64 sample units. The first panel tracks the number of sample units the interviewer started, and has not yet stopped, pursuing on each day over the fieldwork (active case workload). The second panel tracks the number of sample units the interviewer was assigned and remain to be worked (assigned case workload). This visualization shows strong fluctuations in project-specific workload over the fieldwork period, and supports the proposed relevance of workload operationalizations that take temporal variation into account. The visualization also demonstrates that the two time-varying operationalizations produce measurements that are strongly related but far from identical. As would be expected, assigned case workload is not only higher on average (as indicated by the dotted line) but also changes in broader steps over time.

We observe project-specific workload (for interviewer j on day t) for 155 interviewers over 106 fieldwork days in Round 6 and for 151 interviewers over 145 fieldwork days in

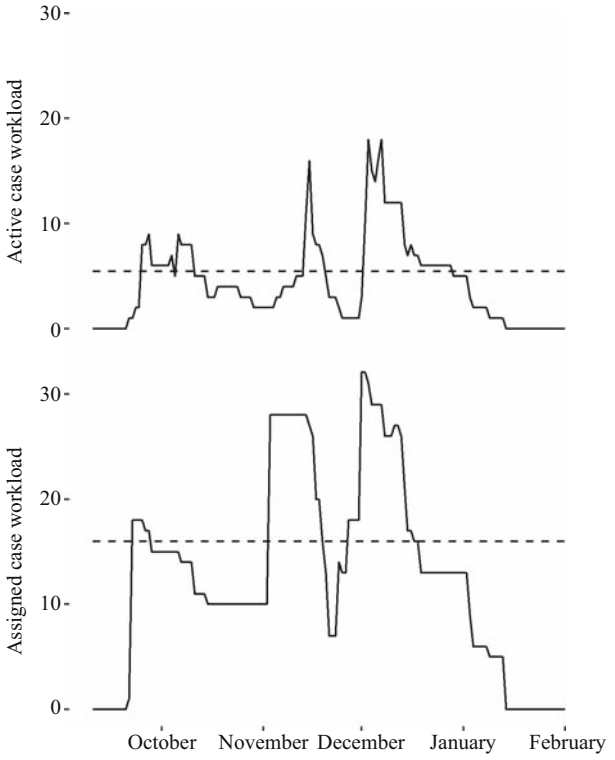

Fig. 2. Illustration of an interviewer's project-specific workload over the fieldwork period.

Round 7. Table 1 presents descriptive statistics for the two project-specific workload operationalizations at the interviewer-day level. The upper panel presents the mean and standard deviation across all interviewer-days. The lower panel presents the mean and standard deviation across the interviewer-days that will be included in the analysis. Given

Table 1. Descriptive statistics of project-specific workload at the interviewer-day level.

	ESS6-BE		ESS7-BE	
	Mean	SD	Mean	SD
All interviewer-days				
Active case workload	2.51	3.63	2.22	3.60
Assigned case workload	6.55	7.28	6.08	7.84
N	16,430		21,895	
Interviewer-days with strictly positive active case workload				
Active case workload	5.35	3.58	5.26	3.82
Assigned case workload	11.9	5.54	12.31	6.96
N	7,702		9,259	

that no interviewer activity can be expected when interviewers have no cases to work on, the analysis only makes sense for interviewer-days on which workload is strictly positive. The boundaries of the relevant analytical sample thus also depend on the chosen operationalization. In order to make valid comparisons across operationalizations, we drop interviewer-days with zero workload according to either operationalization, which in practice corresponds to dropping interviewer-days with zero *active* case workload because assigned case workload is always at least as large as active case workload.

Including all interviewer-days, the two measures exhibit disproportionally high variability and are strongly correlated ($r = .69$ in Round 6, $r = .71$ in Round 7). The variances and correlation are driven up, and the averages are driven down, by the large numbers of interviewer-days at which workload is zero. Active case workload is zero on 53% and 58% of interviewer days in Round 6 and Round 7, respectively. Assigned case workload is zero on 46% and 51% of interviewer days in Round 6 and Round 7, respectively. The number of days with zero workload varies considerably across interviewers. The interviewers had strictly positive active case workloads on between 1 and 131 fieldwork days in Round 6 ($M = 61.32$, SD $= 30.57$) and between 4 and 100 fieldwork days in Round 7 ($M = 49.69$, SD $= 21.61$). This is to a large extent due to the selective engagement of the geographically better located and the more effective interviewers among those available for the second phase of the fieldwork, when many non-responding sample units were reallocated. Within the relevant analytical samples ($N = 7{,}702$ in Round 6 and $N = 9{,}259$ in Round 7), the bivariate correlation between the two workload measures is much reduced ($r = .49$ in Round 6, $r = .53$ in Round 7), but remains positive and significant.

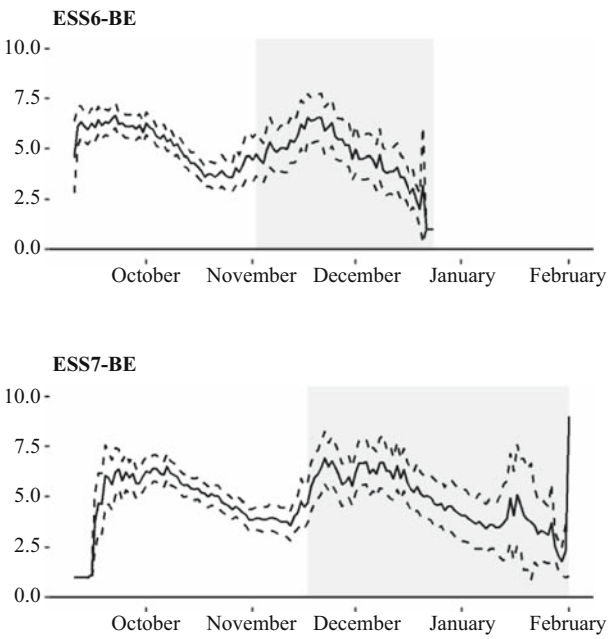

Fig. 3. Average active case workload (and 95% confidence interval) for active interviewers over the fieldwork period.
Note: The pattern for assigned case workload is highly similar.

The intensity of project-specific workload variation over the fieldwork period is clearly demonstrated by the workload pattern observed among active interviewers (Figure 3). The cyclic pattern in the average project-specific workload over the fieldwork period is very similar for the two rounds. The average workload remains fairly stable over the first month, then gradually decreases over the remainder of the first phase of the fieldwork. At the start of the second fieldwork phase, characterized by intensive reallocations, the average workload rebounds to a level similar to that at the start of the fieldwork, before gradually decreasing again.

4.2. *Aggregating Project-Specific Workload to the Interviewer Level*

In order to compare the predictive power of project-specific workload as a time-varying interviewer characteristic to project-specific workload as a fixed interviewer characteristic, we also construct the corresponding interviewer-level summary measures of the time-varying measurements. Taking the average of the time-varying measurements for each interviewer essentially corresponds to weighting (sets of) sample units by the amount of time they are in the interviewers' workloads. Whereas the traditional total-count measure captures case workload over the entire fieldwork period as a whole, these new measures of interviewer workload at the interviewer level capture the active/assigned case workload on a typical fieldwork day. Overall, interviewers worked on average on a total of 25 and 28 sample units over the fieldwork period in Round 6 and Round 7, respectively. On a typical fieldwork day (i.e., according to the temporally weighted average) in either round, the interviewers had on average about five sample units in their 'active' case workload and about 11 in their 'assigned' case workload (Table 2).

Among the three operationalizations at the interviewer level, the two measuring *typical* workload are most strongly correlated ($r = .59$ in Round 6, $r = .64$ in Round 7). Total case workload is more strongly related to typical assigned case workload ($r = .50$ in Round 6, $r = .57$ in Round 7) than to typical *active* case workload ($r = .22$ in Round 6, $r = .36$ in Round 7).

4.3. *Daily Field Efforts and Response Rates*

We define the following interviewer field effort and performance measures at the level of interviewer-days. The measures are derived from the ESS contact history data.

Table 2. Descriptive statistics of project-specific workload at the interviewer level.

	ESS6-BE		ESS7-BE	
	Mean	SD	Mean	SD
Total case workload	25.09	13.49	28.11	17.56
Active case workload on a typical fieldwork day	5.20	1.85	5.05	1.94
Assigned case workload on a typical fieldwork day	11.45	3.26	11.54	3.95
N	155		151	

As previously noted, the analysis is restricted to interviewer-days with strictly positive active case workloads ($N = 7{,}702$ in Round 6 and $N = 9{,}259$ in Round 7). Some of the dependent variables are defined only on a smaller set of interviewer-days, as indicated below.

Contact activity observed for interviewer j on day t is a binary indicator that is equal to 1 if any personal visits are made by that interviewer on that day of fieldwork. Activity is observed on 35.34% ($N = 2{,}722$) of interviewer-days in Round 6 and 31.87% ($N = 2{,}951$) of interviewer days in Round 7.

The **contact effort** observed for (active) interviewer j on day t is the number of personal visits made by that interviewer, relative to his case workload (in tens of cases) on that day. We evaluate the relative number of visits only for interviewer-days on which there was any interviewer activity ($N = 2{,}722$ in Round 6 and $N = 2{,}951$ in Round 7). On the average (active) interviewer-day, 3.57 visits (SD = 3.03) were made in Round 6 and 3.61 visits (SD = 3.27) were made in Round 7. Taking the interviewers' daily case workloads into account, the contact effort on the average (active) interviewer-day was 3.02 visits per ten assigned cases in workload in both rounds (SD = 2.30 in Round 6, SD = 2.49 in Round 7).

The **contact rate** observed for interviewer j on day t is defined as the relative number of cases visited by interviewer j on day t with which contact with the target respondent was successfully made. The contact rate can thus only be evaluated for days on which at least one visit was made ($N = 2{,}722$ in Round 6 and $N = 2{,}951$ in Round 7). On the average (active) interviewer-day, contact was made with 55.44% (SD = 38.75%) of visited target respondents in Round 6 and for 52.25% (SD = 39.92%) of visited target respondents in Round 7.

The **conditional cooperation rate** observed for interviewer j on day t is defined as the relative number of target respondents contacted by interviewer j on day t from which an interview was successfully completed. The cooperation rate can thus only be evaluated for days on which at least one successful contact was made ($N = 2{,}177$ in Round 6 and $N = 2{,}220$ in Round 7). On the average (active) interviewer-day, an interview was completed for 52.86% (SD = 44.62%) of contacted target respondents in Round 6 and for 47.29% (SD = 44.86%) of contacted target respondents in Round 7.

4.4. Modelling Approach

For each outcome variable y_{jt} observed for interviewer j on day t, multilevel regression models (logistic for contact activity, linear for contact efforts, contact rates and cooperation rates) are estimated with interviewer-days nested within interviewers and one of the five measures of project-specific workload (at the level of interviewer-days or at the interviewer level), and its quadratic term, as the explanatory variables. The different project-specific workload operationalizations tested are briefly recapitulated in Table 3.

Model 0 includes a random intercept and fieldwork day control variables but none of the interviewer workload variables. This base model is used to estimate the share of the variability in daily field efforts and response outcome rates which may be attributed to

Table 3. Operationalizations of project-specific workload.

	At the level of interviewer days	At the interviewer level
Traditional operationalizations	–	Total active case workload = total number of sample units worked on over the fieldwork period as a whole
New operationalizations	Active case workload on day t = number of sample units for which interviewer activity has started and not yet ceased	Active case workload on a typical fieldwork day = *average* number of sample units for which interviewer activity has started and not yet ceased
	Assigned case workload on day t = number of sample units that have been and remain assigned	Assigned case workload on a typical fieldwork day = *average* number of sample units that have been and remain assigned

systematic differences between the interviewers, and to evaluate the model fit of subsequent models.

$$y_{jt} = \gamma_{00} + \theta t + \Gamma \boldsymbol{Day\ of\ the\ week}_t + u_{0j} + \varepsilon_{jt} \qquad (0)$$
$$\theta(t) = \theta_1 t + \theta_2 t^2 + \theta_3 t^3 + \theta_4 t^4$$

γ_{00} represents the overall mean intercept, u_{0j} is the interviewer-specific deviation from the overall mean (with zero mean and variance σ^2_{u0}) and ε_{jt} the interviewer-day residual (with zero mean and variance σ^2_{ε}). $\theta(t)$ is a quartic function of fieldwork day t and $\boldsymbol{Day\ of\ the\ week}_t$ is a vector of six dummy variables indicating the day of the week (with Sunday as reference category). A quartic function of time suits the two-phased fieldwork and yields adequate model fits for all four outcome variables. For predicting daily contact rates and daily cooperation rates a linear function of fieldwork day and a cubic function of fieldwork day, respectively, would have sufficed. For these two field performance measures, adding the higher order polynomial terms does not improve model fit. Adding the higher order polynomial terms (up to four) does significantly improve model fit for predicting the probability of any contact activity and the contact efforts made.

Models 1 and 2 additionally include, respectively, a fixed (interviewer-level) measure of project-specific workload (and its quadratic term) and a time-varying (interviewer-day-level) measure of project-specific workload (and its quadratic term):

$$y_{jt} = \gamma_{00} + \beta_1 Ln(Workload_j) + \beta_2 Ln(Workload_j)^2 + \theta t \qquad (1)$$
$$+\Gamma \boldsymbol{Day\ of\ the\ week}_t + u_{0j} + \varepsilon_{jt}$$

$$y_{jt} = \gamma_{00} + \beta'_1 Ln(Workload_{jt}) + \beta'_2 Ln(Workload_{jt})^2 + \theta t \qquad (2)$$
$$+\Gamma \boldsymbol{Day\ of\ the\ week}_t + u_{0j} + \varepsilon_{jt}$$

where $Workload_j$ is the total case workload (Model 1A), the active case workload on a typical fieldwork day (Model 1B) or the assigned case workload on a typical fieldwork day (Model 1C) for interviewer j, and $Workload_{jt}$ is the active case workload (Model 2B) or the assigned case workload (Model 2C) for interviewer j on day t.

4.5. Robustness Checks

The standard multilevel model assumption of compound symmetry, that is, the different workload measurements are equally correlated, irrespective of their distance in time, is usually not realistic for repeated measures, but the fixed parameter estimates tend to be robust to small misspecifications of the random part of the model (Hox et al. 2018). The empirical autocorrelation functions for the within-interviewer residuals indicate that the residual covariance matrix across measurements (at least for the measurements of contact activity, contact efforts and contact rates) may be better characterized by a (first-order) autoregressive structure. Although explicitly specifying a more complex (autoregressive) within-interviewer correlation structure for these outcome variables does improve the model fit, the workload parameter estimates are not altered in a meaningful way (results not tabulated).

Controlling for fixed interviewer characteristics that may affect both assigned workloads and contact efforts and response outcome rates, prior experience (number of years working as a survey interviewer, in 5 categories) and having another job (binary indicator), does not meaningfully alter the workload parameter estimates either (results not tabulated).

As previously mentioned, the analytic sample contains only interviewer-days with strictly positive *active* case workload. Alternatively, the analytic sample could have been delineated on the basis of strictly positive *assigned* case workloads. Given that the analyses for contact effort, contact rates and cooperation rates further restrict the analytic sample to interviewer-days with any interviewer activity, the results for these outcomes are unaffected by this analytic choice. For the first outcome measure, the binary indicator of any interviewer activity, the analytic sample would have been somewhat larger if interviewer-days with zero active case workload (but non-zero assigned case workload) were also included. Using this extended analytic sample, the model fit relative to the base model (Model 0) is more pronounced for all workload operationalizations (the models with workload operationalized as a fixed interviewer characteristic as well as the models with workload operationalized as a time-varying interviewer characteristic), but the workload parameter estimates suggest a similar relationship between workload and interviewer activity over the relevant range (results not presented). As may be expected given the additional interviewer-days with zero active case workload, the estimated regression curve is somewhat steeper at the very lowest workload levels when workload is operationalized as the daily *active* case workload but is close to unaltered when workload is operationalized as the daily *assigned* case workload. The reported results are based on the more conservative delineation of the analytic sample.

5. Results

We start by estimating the share of the variability in daily field efforts and response outcome rates which may be attributed to systematic differences between the interviewers

(interviewer-level variance component) and to within-interviewer differences across fieldwork days (interviewer day residual variance component). Table 4 presents the variance components and intraclass correlation coefficients (ICCs) estimated from the base model (Model 0) for the four outcome variables. The ICCs suggest that only a very small share (daily contact activity and cooperation rates) to a moderately large share (daily contact efforts and contact rates) of the observed variability is due to systematic differences *between* interviewers. Much residual variance remains to be explained at the level of interviewer days.

Table 5 shows an overview of the differences in the Akaike information criterion (AIC) for the three models with project-specific workload operationalized as a fixed interviewer characteristic and the two models with project-specific workload operationalized as a time-varying interviewer characteristic, each compared to the base model (Model 0). We first consider the three fixed (interviewer-level) workload operationalizations. These operationalizations have very limited predictive power, and none consistently performs well across the outcome variables and for both survey rounds. For predicting daily contact activity, a marginal improvement in model fit compared to the base model (Model 0) is achieved by including any fixed workload operationalization in Round 7, but no improvement is achieved by including either fixed operationalization in Round 6. For predicting daily contact efforts, some improvement in model fit is achieved in both rounds by including total case workload or typical assigned case workload as fixed workload operationalizations. Typical assigned case workload appears to outperform total case workload as well as typical active case workload in terms of predictive power for this measure of field efforts. For predicting daily contact and cooperation rates, some improvement in model fit is achieved in both rounds by including typical active case workload and total case workload, respectively. Model fits for both outcome rates gain

Table 4. Estimated variance components for fieldwork effort and outcome variables.

	ESS6-BE	ESS7-BE
Daily contact activity		
Interviewer-level variance	0.1361***	0.1729***
ICC	0.0397	0.0499
Daily contact effort		
Interviewer-level variance	0.5415***	0.4897***
Residual variance	4.6477	5.3025
ICC	0.1044	0.0845
Daily contact rate		
Interviewer-level variance	0.0157***	0.0154***
Residual variance	0.1318	0.1409
ICC	0.1068	0.0986
Daily cooperation rate		
Interviewer-level variance	0.0041*	0.0039*
Residual variance	0.1907	0.1927
ICC	0.0209	0.0200

Note: Statistical significance of the random intercept is tested by a likelihood ratio test. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table 5. Model fit statistics (ΔAIC) relative to base model (Model 0) for alternative operationalizations of project-specific workload.

	Daily contact activity (ref. = no personal visits made)		Daily contact effort		Daily contact rate		Daily cooperation rate	
	ESS6-BE	ESS7-BE	ESS6-BE	ESS7-BE	ESS6-BE	ESS7-BE	ESS6-BE	ESS7-BE
Fixed interviewer characteristic models								
Model 1A (Total case workload)	–	−5.70	−13.10	−25.14	−3.25	–	−6.73	−3.11
Model 1B (Active case workload on a typical fieldwork day)	–	−3.10	–	–	−15.14	−12.16	–	−4.39
Model 1C (Assigned case workload on a typical fieldwork day)	–	−2.38	−68.25	−86.88	−6.75	–	−4.40	–
Time-varying interviewer characteristic models								
Model 2B (Daily active case workload)	−646.79	−706.80	−325.12	−162.76	−52.48	–	−108.93	−68.47
Model 2C (Daily assigned case workload)	−109.41	−207.04	−162.15	−369.49	−19.61	–	−12.61	−7.17
N	7,702	9,259	2,722	2,951	2,722	2,951	2,177	2,220

Note: Values for models with no statistically significant improvement in model fit compared to the base model (Model 0), as indicated by the likelihood ratio test ($p < .05$) are suppressed.

slightly from including any of the other fixed workload operationalization in one of the two survey rounds, but not in the other.

The predictive power of the two time-varying workload operationalizations is much stronger. At least for predicting daily contact activity, contact efforts and cooperation rates, the models with the time-varying workload operationalizations yield considerable improvements in model fit compared to the base model (Model 0) in both rounds. Only for predicting daily contact rates, the time-varying workload operationalizations are not much more convincing than the fixed (interviewer-level) workload operationalizations. Some improvement in model fit for this outcome variable is achieved by including either time-varying workload operationalization in Round 6, but no improvement is achieved by either operationalization in Round 7. Of the two time-varying operationalizations, daily *active* case workload frequently exceeds daily *assigned* case workload in terms of predictive power.

The inferred relationships between project-specific workload, operationalized in a time-varying manner (daily active case workload and daily assigned case workload), and the four outcome variables are discussed in the following subsections. Because the inclusion of the quadratic term complicates interpreting the parameter estimates directly, we present prediction plots for each outcome variable (Figures 4 to 7) for one hypothetical fieldwork

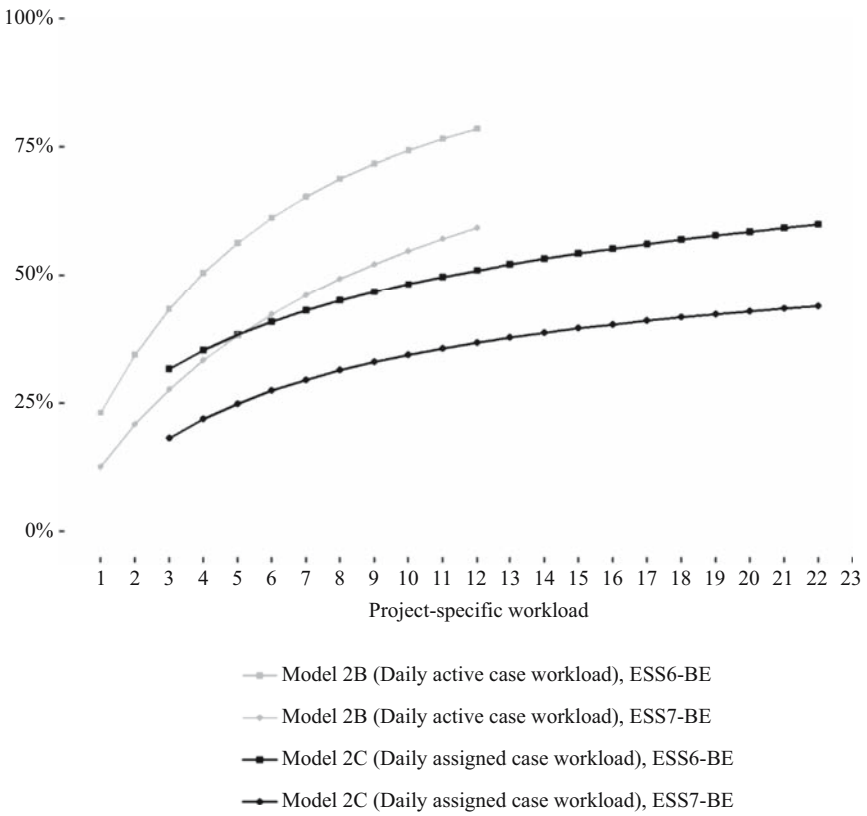

Fig. 4. Predicted probability of contact activity as a function of time-varying project-specific workload, evaluated on a Monday, the 50th fieldwork day.

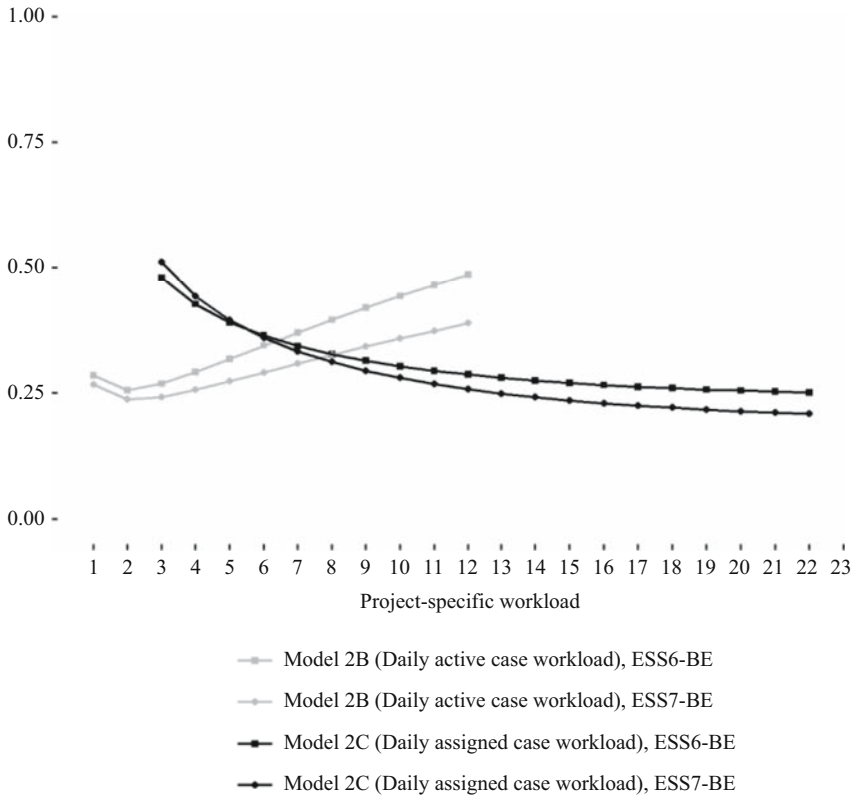

Fig. 5. Predicted contact effort per case in workload as a function of time-varying project-specific workload, evaluated on a Monday, the 50th fieldwork day.

day. These prediction plots demonstrate the dominant direction and curvature of the regression curve. The choice of fieldwork day presented (Monday, the 50th fieldwork day) is arbitrary but, in the absence of any interaction effects between fieldwork day and project-specific workload, only affects the fieldwork-day-specific intercept, not the shape of the curve. In order to avoid extrapolations beyond the normal workload ranges, the predicted field efforts and response outcome rates are presented over the range defined by the 5% and 95% workload quantiles (3 to 22 cases in assigned case workload, 1 to 12 cases in active case workload).

5.1. Contact Activity

The probability of contact activity on a given day tends to increase with project-specific workload (Figure 4). The results for the two survey rounds are very similar. The predicted probability of any interviewer activity on a given day is about 10 percentage points higher when an interviewer has an assigned case workload of 10 cases rather than 5, and again 5 to 6 percentage points higher when an interviewer has an assigned case workload of 15 cases rather than 10. The relationship is stronger for daily *active* case workload than for daily *assigned* case workload but, as previously noted, the strength of the association may

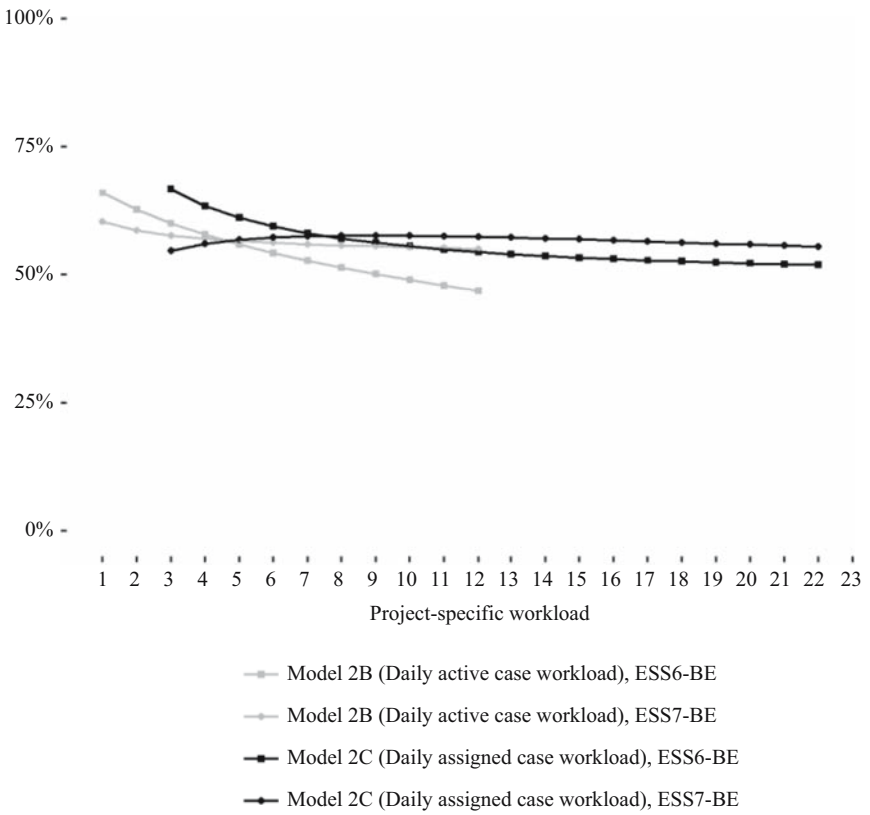

Fig. 6. Predicted daily contact rate as a function of time-varying project-specific workload, evaluated on a Monday, the 50th fieldwork day.

be inflated by active case workloads increasing as more sample units are actually being attempted. The predicted probability of any interviewer activity on a given day is 17 to 18 percentage points higher when an interviewer has an active case workload of 10 cases rather than 5.

5.2. *Contact Effort*

Contact efforts tend to decrease with daily *assigned* case workload (Figure 5). The relationship between project-specific workload and contact effort is in the opposite direction for the daily *active* case workload operationalization. As for predicting contact activity, the positive association for daily active case workload may well be an artefact of this particular workload operationalization. In both survey rounds, the negative relationship between project-specific workload (assigned case workload operationalization) and contact effort is strongest at very low workload levels but more gradual for higher workload levels. Beyond an assigned case workload of about ten cases the negative relationship flattens out almost completely and contact efforts per case remain fairly constant across increasingly large workloads.

5.3. Contact Rate

Contact rates are inconsistently, if at all, related to project-specific workload (Figure 6). The parameter estimates of the daily *assigned* case workload model suggest a weak negative association in Round 6 but a weak positive association in Round 7. The estimates of the daily *active* case workload model suggest a weak, approximately linear negative association in Round 6, but no association in Round 7. Altogether these results indicate that daily workload and contact rates are not really related.

5.4. Cooperation Rate

Cooperation rates tend to decrease with project-specific workload (Figure 7). The results for the two survey rounds are again fairly similar. Over the relevant range of daily workload levels, the negative relationship is approximately linear, and much steeper for daily *active* case workload than for daily *assigned* case workload. The predicted cooperation rate on a given day is 5 to 6 percentage points lower when an interviewer has an assigned case workload of 10 cases rather than 5, and again 2 to 3 percentage points lower when an interviewer has an assigned case workload of 15 cases rather than 10. The predicted cooperation rate on a given day is 10 to 14 percentage points lower when an

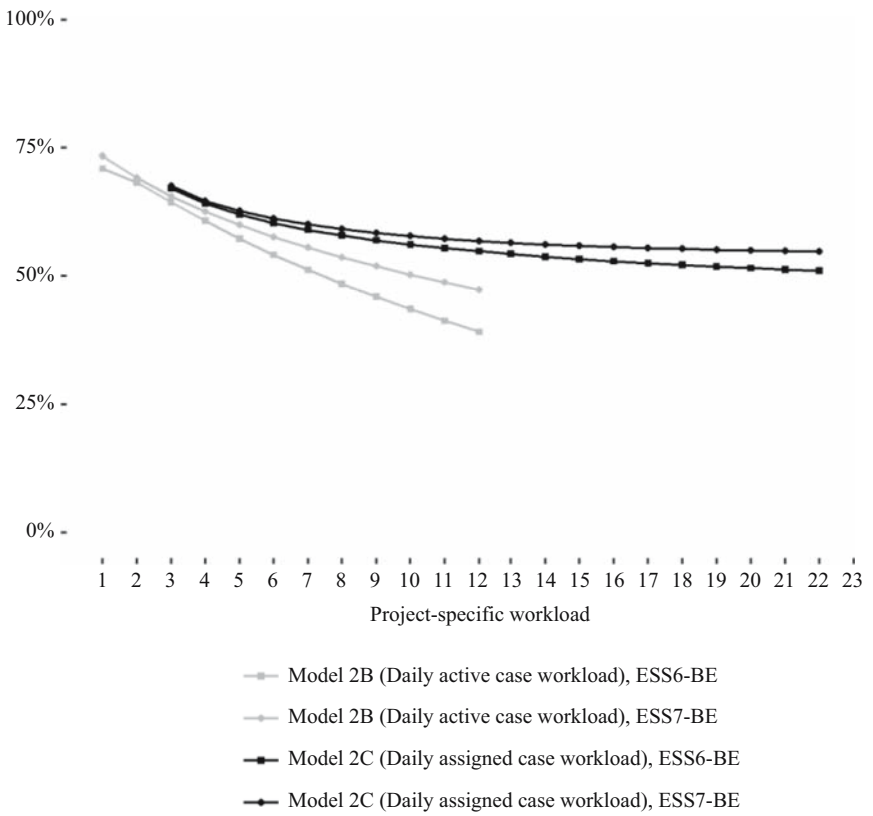

Fig. 7. Predicted daily cooperation rate as a function of time-varying project-specific workload, evaluated on a Monday, the 50th fieldwork day.

interviewer has an *active* case workload of 10 cases rather than 5. Although the negative relationship between project-specific workload (assigned case workload operationalization) and cooperation rates is strongest at very low workload levels and much more gradual for higher workload levels, even these apparently small differences are far from negligible in view of commonly observed cooperation rates.

6. Conclusions and Discussion

Interviewer characteristics are usually assumed fixed over the fieldwork period. In this article we show that since interviewers' project-specific workloads can vary strongly over the fieldwork period, a more fine-grained, time-varying operationalization can be useful to explain differences in interviewers' field efforts and performance. We proposed two approaches to measure project-specific workload for interviewers on each day in the fieldwork: (1) the number of sample units which have been and remain assigned on any day t (assigned case workload), and (2) the number of sample units for which interviewer activity has started and not yet ceased on any day t (active case workload). We have used these operationalizations in an examination of workload correlates in two rounds of the European Social Survey in Belgium (results not presented).

The results indicate that none of the operationalizations at the interviewer level, whether the traditional total-count workload measure or the proposed operationalizations of *typical* (active or assigned) case workload, consistently help to explain the observed systematic differences in efforts and response outcome rates between interviewers in the contact and recruitment task. The remaining variability in task efforts and response outcome rates within interviewers, across fieldwork days, is not completely random. This variability can partially be explained by differences in interviewers' workloads at this level, across fieldwork days. Capturing temporal workload variation directly, the time-varying operationalizations, daily active case workload and daily assigned case workload, fairly consistently outperform the interviewer-level *typical* workload operationalizations that are derived from them, as well as the traditional total-count workload operationalization. The results are remarkably similar for the two survey rounds studied, supporting their external validity.

Although the results show that *active* case workload frequently has a better predictive power than *assigned* case workload when predicting interviewers' efforts and performance in the contact and recruitment task, the positive associations with contact activity and efforts observed for the active case workload operationalization may be biased by reverse causality. An interviewer's field efforts to a large extent drive the number of active cases he or she has. Active case workload itself would thus by construction be positively related to contact activity and efforts. *Assigned* case workload is much less sensitive to variations in interviewers' field efforts over the fieldwork period and can thus be regarded as the conceptually more adequate operationalization of project-specific workload. Unfortunately, the assigned case workload operationalization requires additional case assignment paradata where contact history paradata suffices for the *active* case workload operationalization. Case assignment paradata may be even less readily available than contact history paradata. Without access to additional case assignment data, it may be possible to derive an approximate assigned case workload measure from contact history

data by identifying groups of sample units with overlapping recorded interviewer activity, suggestive of simultaneous assignment. Our attempt to derive such an approximation has not altogether been successful for these two survey rounds of the European Social Survey in Belgium.

Further research on the impact of interviewer workload in particular, and research aimed at unpacking the fieldwork process more generally, would benefit from contact history paradata being supplemented with case assignment and conversion paradata. Timestamped assignments to interviewers, returns to field offices, and transitions in incentive schemes, recruitment modes and other features of the data collection that may be dynamically altered over the course of the fieldwork, would allow investigations into aspects of the fieldwork process that are not captured by contact history paradata and currently remain obscure in many settings. Interviewer workload patterns over the fieldwork period is one such aspect. Other aspects of the fieldwork process into which valuable insights may be gained are the progression of sample units being initially issued, the application of different reassignment and conversion strategies, and the extent to which sample units are (temporarily) shelved by field supervisors because they are costly to pursue or interviewers are (temporarily) unavailable, or by interviewers because they are remote or located in disadvantaged areas.

Although research on the fieldwork process would benefit from case assignment and conversion paradata supplementing contact history paradata, simply collecting more paradata is a poor strategy if adequate paradata quality cannot be maintained. The proposed approach of measuring workload more precisely presumes higher accuracy of the underlying (contact history and case assignment) paradata than is required for the traditional-total count measure. Comprehensive quality assessments of this type of paradata are rarely published (West and Sinibaldi 2013), but some studies have demonstrated that (unsuccessful) contact attempts, especially attempts that may not be unambiguously considered proper attempts by the interviewers (e.g., 'drive-by' visits, Biemer et al. 2013; Wagner et al. 2017), are commonly underreported. We observed small numbers of cases for which the recorded date of the first contact attempt preceded the recorded date of assignment, indicative of entry errors in the contact history paradata and/or the assignment paradata. Especially when contact histories are not recorded by the interviewers in a timely manner, we may worry about their accuracy. The quality of case assignment paradata may likewise suffer from underreporting and inaccurate recording by field managers. As for other types of paradata, quality assurance and control procedures should safeguard the quality of contact history and case assignment paradata.

The results related to the secondary study objective, to apply the different workload operationalizations in a further exploration of the interviewer workload-performance relationship, are tentative, but emphasize some interesting mechanisms that should invite further research. The performance of face-to-face survey interviewers in the contact and recruitment task does appear to be affected by the number of sample units that require their attention over time. However, these workload effects are not unidirectional and straightforward.

The hypothesis that when interviewers carry larger workloads their contact effort per case is reduced (Hypothesis 1) is only partially supported. The initial evidence suggests that when interviewers carry heavier workloads they tend to be *more* likely to make any

personal visits at all. This positive result suggests that interviewers are more strongly motivated to make a round trip of visits when they are carrying larger workloads than when they are carrying small workloads. When sample units are geographically clustered, the fixed cost of travel to the cluster may only be worth the investment if many can be visited on the same round trip. The result substantiates the comments from several of the interviewers in the study (collected via a post-fieldwork interviewer questionnaire) that assignment sizes are *too small* rather than too large. To which extent interviewers may experience larger workloads as a greater potential income (Wuyts and Loosveldt 2016), rather than as a burden, is an important question to answer if the mechanisms through which interviewer workload affect fieldwork outcomes are to be well understood. Whereas larger workloads were related to a higher probability of any interviewer activity, they were also related to smaller numbers of personal visits made relative to the number of cases in workload. Interviewers may be induced to spread their contact efforts over more working days, rather than reduce their contact efforts, when they have more cases to work on. The negative association is most pronounced at lower workload levels, and almost completely flattens out for moderate to large workloads. The observed compression of the negative effect of workload on contact effort challenges the presumed dominant role of time availability constraints in the contact and recruitment task.

The hypothesis that when interviewers carry larger workloads their contact and cooperation rates are reduced (Hypothesis 2) is also only partially supported. Overall, the link between interviewer workload and contact efforts is stronger than the link between workload and outcome rates. This may be explained by efforts being completely within the survey interviewers' control. That contact rates were not consistently related to project-specific workload suggests that contact schedules are no less productive when interviewers carry larger workloads than when they have only small numbers of cases to work on. It is possible that time constraints do not severely limit the number of visits during the more productive hours. Alternatively, the number of these visits may be considerably limited but interviewers may take likely accessible at home patterns and contact histories of different target respondents into account when planning their visits, for example by dedicating evening visits to target respondents that are of working age and/or previously could not be contacted during the day. In this sense, large workloads may encourage interviewers to optimize their contact schedule within the given time constraints, offsetting the direct negative effects of these time constraints. That cooperation rates, but not contact rates, were consistently negatively related to interviewer workload suggests that when interviewers carry larger workloads they interact with contacted target respondents in a different, less (immediately) successful way at the doorstep, possibly more easily accepting reluctance or postponement. The negative association, however, is most pronounced at lower workload levels, and flattens out for moderate to large workloads.

A common limitation of any operationalization of interviewer workload in terms of case counts (whether in the aggregate or time-varying) is that assignments are assumed equivalent in the amount of work they imply. For many reasons the amount of work may vary across sample units, and the amount of work for a single sample unit need not even be constant over time. A first obvious reason is that not all sample units are equally accessible and cooperative. A second reason in the context of face-to-face survey data collection is that travel distance to and among assignment clusters (a function of the geographical

distribution of sample units relative to the geographical distribution of interviewers) may vary considerably and is also likely to weigh heavily on interviewers' experienced workloads. Marginal travel costs may also shift as different (clusters of) sample units move in and out of interviewers' workloads. The recent study of interviewer travel behavior by Wagner and Olson (2018) is worth mentioning in this context. They examine administrative travel data at the level of interviewer days and observe that interviewer travel between clusters is a relevant factor to both field outcomes and fieldwork costs.

A plausible implication is that the amount of work for one sample unit is larger when the number of assigned sample units is small (and mostly dispersed, hard-to-contact and uncooperative sample units remain). This would weaken our conclusions for the workload effects at low workload levels. We previously noted that the observed negative association of assigned case workload with the number of personal visits made relative to the number of cases in workload, and with the daily cooperation rate, tends to flatten out for moderate to large workload levels. The observed negative associations at low workload levels may to some extent be driven by the particular composition of the remaining workload.

A second limitation of this study is the limited range of observed interviewer workloads. We observe an average assigned case workload of 12 sample units across interviewer-days, and an assigned case workload of only up to 22 cases for 95% of interviewer-days. A plausible explanation for why we observe mostly moderate workload levels lies in the batch allocation strategy that was used in the two survey rounds studied, against the backdrop of the explicit workload restrictions advocated by the standards and specifications of the European Social Survey. In the European Social Survey, a maximum total of 48 sample units is in principle allowed to be assigned per interviewer over the course of the fieldwork, including any reassignments as well as initial assignments (Stoop et al. 2010). If workload only has strong harmful effects on interviewers' fieldwork efforts and outcomes at higher levels, these cannot be observed. Excessive workloads that may negatively affect interviewers' performance may thus have been systematically avoided by design (Blom et al. 2011).

One should also keep in mind that the question about reasonable interviewer workloads is not only relevant to the fieldwork process and nonresponse error, but also (and possibly even more so) for measurement error. The extent to which interviewers' individual systematic effects on responses to survey questions result in inflated standard errors of survey estimates depends on the total number of respondents interviewed by each interviewer (Kish 1962; Beullens et al. 2016).

In summary, when interviewers carry heavier project-specific workloads, we observe (1) a higher probability of making any visits at all, but, up to a certain point, a reduced number of visits if any are made, and (2) a similar probability of attempted target respondents being successfully contacted, but, again up to a point, a reduced probability of contacted target respondents actually participating.

Overall, our results warrant no serious concern (at moderate workload levels) but some caution about possible harmful workload effects on nonresponse error. Given that interviewers' project-specific workload is a factor that can be controlled by survey designers and field supervisors, further studies on the risk of workload effects are called for. Larger workload levels than studied here deserve particular attention. On the other hand, the observed association with the likelihood of making any personal visits at all

encourages further investigation into minimum efficient workloads and the possible benefits and risks of continuous replenishment of workloads to a manageable level. There may be motivational mechanisms driving task performance up with increasing workload (Delasay et al. 2019). In particular at low workload levels, increasing workloads may be experienced by engaged and responsive interviewers as a greater challenge and a more worthwhile investment of time and effort. An inverted-U pattern may emerge from different workload mechanisms (with effects in different directions) dominating over low- versus high-workload ranges. Such a pattern has been commonly hypothesized for task performance as a function of subjective workload measures and stress (Westman and Eden 1996; Muse et al. 2013) and has been observed using objective workload measures in field studies in the restaurant and grocery industry (Tan and Netessine 2014; Bruggen 2015).

Our results for contact efforts and contact rates do not unambiguously support the commonly held belief that larger workloads allow interviewers less time and effort to expend for individual sample units, and therefore necessarily harm response rates. Interviewers are likely to adjust their contact activities on the basis of a careful (though not necessarily deliberate) balancing exercise that depends on preferences about work time organization and anticipated benefits and costs of additional efforts, as well as overall time availability. The result for cooperation rates highlights that not only interviewers' contact activities, but also interviewers' behavior at the doorstep can be affected.

7. References

Beullens, K., G. Loosveldt, K. Denies, and C. Vandenplas. 2016. "Quality Matrix for the European Social Survey, Round 7." Leuven, Belgium: Centre for Sociological Research, KU Leuven.

Biemer, P.P., P. Chen, and K. Wang. 2013. "Using Level-of-Effort Paradata in Non-Response Adjustments with Application to Field Surveys." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 176(1): 147–168. DOI: https://doi.org/10.1111/j.1467-985X.2012.01058.x.

Blom, A.G. 2012. "Explaining Cross-Country Differences in Survey Contact Rates: Application of Decomposition Methods: Cross-Country Differences in Survey Contact Rates." *Journal of the Royal Statistical Society, Statistics in Society* 175(1): 217–242. DOI: https://doi.org/10.1111/j.1467-985X.2011.01006.x.

Blom, A.G., E.D. de Leeuw, and J.J. Hox. 2011. "Interviewer Effects on Nonresponse in the European Social Survey." *Journal of Official Statistics* 27 (2): 359–377. Available at: https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/interviewer-effects-on-nonresponse-in-the-european-social-survey.pdf (accessed April 2020).

Botman, S.L. and O.T. Thornberry. 1992. "Survey Design Features Correlates of Nonresponse." In *Proceedings of the Survey Research Methods Section:* American Statistical Association, August 9–13, 1992. 309–314. Alexandria, VA: American Statistical Association. Available at: http://www.asasrms.org/Proceedings/papers/1992_048.pdf (accessed April 2016).

Bruggen, A. 2015. "An Empirical Investigation of the Relationship between Workload and Performance." *Management Decision* 53(10): 2377–2389. DOI: https://doi.org/10.1108/MD-02-2015-0063.

Calderwood, L., I. Plewis, S. Ketende, and T. Mostafa. 2016. "Evaluating the Immediate and Longer Term Impact of a Refusal Conversion Strategy in a Large Scale Longitudinal Study." *Survey Research Methods* 10(3): 225–236. DOI: https://doi.org/10.18148/srm/2016.v10i3.6275.

Delasay, M., A. Ingolfsson, B. Kolfal, and K. Schultz. 2019. "Load Effect on Service Times." *European Journal of Operational Research* 279(3): 673–686. DOI: https://doi.org/10.1016/j.ejor.2018.12.028.

European Social Survey (2014). *ESS Round 6 Data from Contact forms, edition 2.0*. Bergen, Norway. Norwegian Social Science Data Services, Norway–Data Archive and distributor of ESS data.

European Social Survey (2016). *ESS Round 7 Data from Contact forms, edition 2.1*. Bergen, Norway. Norwegian Social Science Data Services, Norway–Data Archive and distributor of ESS data.

Groves, R.M. and M. Couper. 1998. "How Survey Design Features Affect Participation." In *Nonresponse in Household Interview Surveys,* 269–293. New York, NY: John Wiley & Sons.

Hox, J.J., M. Moerbeek, and R. van de Schoot. 2018. *Multilevel Analysis: Techniques and Applications*. Third edition. Quantitative Methodology Series. New York, NY: Routledge.

Kirchner, A. and K. Olson. 2017. "Examining Changes of Interview Length over the Course of the Field Period." *Journal of Survey Statistics and Methodology* 5: 84–108. DOI: https://doi.org/10.1093/jssam/smw031.

Japec, L. 2005. "The Concept of Interviewer Burden." Presented at the International Workshop on Household Survey Nonresponse, Tällberg, Sweden.

Japec, L. 2008. "Interviewer Error and Interviewer Burden." In *Advances in Telephone Survey Methodology*, edited by J.M. Lepkowski, C. Tucker, J.M. Brick, E.D. de Leeuw, L. Japec, P.J. Lavrakas, M.W. Link, and R.L. Sangster, 187–211. New York, NY: John Wiley & Sons.

Kish, L. 1962. "Studies of Interviewer Variance for Attitudinal Variables." *Journal of the American Statistical Association* 57(297): 92–115. DOI: https://doi.org/10.2307/2282442.

Loosveldt, G. and K. Beullens. 2013. "'How Long Will It Take?' An Analysis of Interview Length in the Fifth Round of the European Social Survey." *Survey Research Methods* 7(2): 69–78. DOI: https://doi.org/10.18148/srm/2013.v7i2.5086.

Loosveldt, G., A. Carton, and J. Billiet. 2004. "Assessment of Survey Data Quality: A Pragmatic Approach Focused on Interviewer Tasks." *International Journal of Market Research* 46(1): 65–82. DOI: https://doi.org/10.1177/147078530404600101.

Muse, L.A., S.G. Harris, and H.S. Feild. 2003. "Has the Inverted-U Theory of Stress and Job Performance Had a Fair Test?" *Human Performance* 16(4): 349–364. DOI: https://doi.org/10.1207/S15327043HUP1604_2.

Nicoletti, C. and N.N. Buck. 2004. "Explaining Interviewee Contact and Co-Operation in the British and German Household Panels." 2004–2006. Working Papers of the Institute for Social and Economic Research. Colchester, United Kingdom: University of Essex.

Olson, K. and A. Peytchev. 2007. "Effect of Interviewer Experience on Interview Pace and Interviewer Attitudes." *Public Opinion Quarterly* 71(2): 273–286. DOI: https://doi.org/10.1093/poq/nfm007.

Singer, E., M.R. Frankel, and M.B. Glassman. 1983. "The Effect of Interviewer Characteristics and Expectations on Response." *Public Opinion Quarterly* 47(1): 68–83. DOI: https://doi.org/10.1086/268767.

Stoop, I., J. Billiet, A. Koch, and R. Fitzgerald. 2010. *Improving Survey Response: Lessons Learned from the European Social Survey*. Hoboken, NJ: John Wiley & Sons.

Tan, T.F. and S. Netessine. 2014. "When Does the Devil Make Work? An Empirical Study of the Impact of Workload on Worker Productivity." *Management Science* 60(6): 1574–1593. DOI: https://doi.org/10.1287/mnsc.2014.1950.

Wagner, J., K. Olson, and M. Edgar. 2017. "The Utility of GPS Data in Assessing Interviewer Travel Behavior and Errors in Level-of-Effort Paradata." *Survey Research Methods* 11(3): 219–233. DOI: https://doi.org/10.18148/srm/2017.v11i3.6794.

Wagner, J. and K. Olson. 2018. "An Analysis of Interviewer Travel and Field Outcomes in Two Field Surveys." *Journal of Official Statistics* 34(1): 211–237. DOI: https://doi.org/10.1515/jos-2018-0010.

Watson, N. and M. Wooden. 2009. "Identifying Factors Affecting Longitudinal Survey Response." In *Methodology of Longitudinal Surveys*, edited by P. Lynn. Chichester, United Kingdom: John Wiley & Sons.

West, B.T. and A.G. Blom. 2016. "Explaining Interviewer Effects: A Research Synthesis." *Journal of Survey Statistics and Methodology* 5(2): 175–211. DOI: https://doi.org/10.1093/jssam/smw024.

West, B.T. and J. Sinibaldi. 2013. "The Quality of Paradata: A Literature Review." In *Improving Surveys with Paradata: Analytic Uses of Process Information*, edited by F. Kreuter. Hoboken, NJ: John Wiley & Sons.

Westman, M. and D. Eden. 1996. "The Inverted-U Relationship between Stress and Performance: A Field Study." *Work & Stress* 10(2): 165–173. DOI: https://doi.org/10.1080/02678379608256795.

Wuyts, C. and G. Loosveldt. 2016. "Workload-Related Interviewer Characteristics and Unit Nonresponse in ESS Belgium." Presented at the International Workshop on Household Survey Nonresponse, Oslo, Norway.

Received September 2018
Revised June 2019
Accepted September 2019

sciendo

Journal of Official Statistics, Vol. 36, No. 3, 2020, pp. 589–607, http://dx.doi.org/10.2478/JOS-2020-0030

Assessing Interviewer Performance in Approaching Reissued Initial Nonrespondents

Laurie Peeters[1], *David De Coninck*[1], *Celine Wuyts*[1], *and Geert Loosveldt*[1]

Nonresponse is a repeatedly reported concern in survey research. In this article, we investigate the technique of reissuing nonrespondents to another interviewer and attempting to convert them into respondents, using data of Rounds 7 and 8 of the European Social Survey (ESS) in Belgium. The results show no marked differences between respondents interviewed by the more and the less successful interviewers, indicating that the latter are not more successful in persuading more reluctant respondents to participate. Sample units that were unsuccessfully approached in the initial phase by an interviewer with a high response rate are more difficult to convert during the reissue phase. Sample units that were unsuccessfully approached in the initial phase by an interviewer with a low response rate are easier to convert during the reissue phase.

Key words: Nonresponse; European social survey; reissuing.

1. Introduction

Much of the literature on nonresponse in survey research states that response rates are declining, and more efforts (more contact attempts, tailored advance and reminder letters and brochures, incentives for respondents, bonus arrangements for interviewers) should be made to stimulate participation and keep response rates up to standard (Beullens et al. 2018).

Concerns about response rates are strongly driven by the potential selectivity of nonresponse and the risk of nonresponse bias, but low response rates do not necessarily imply selectivity and bias. Empirical studies have reported that correlations between nonresponse rates and nonresponse bias are weaker than expected (Brick and Tourangeau 2017; Groves 2006; Wright 2015). In reality, additional fieldwork efforts are often focused on increasing response rates rather than reducing bias, with the underlying expectation that an increase in response rates will also result in less nonresponse bias. Reissuing initial nonrespondents requires significant resources and efforts, which might not necessarily pay off in terms of lowering nonresponse bias even if an increase in response rate is achieved. Some studies suggest that reissuing does reduce bias in the sample (Lynn and Clarke 2002), while others find no significant change (Curtin et al. 2000; Groves and Couper 2012; Stoop 2004). In recent years, in response to decreasing response rates and the increasing cost of surveys, adaptive survey design were tested. The results of the

[1] Catholic University of Leuven, Centre for Sociological Research, Parkstraat 45, 3000 Leuven, Belgium. Emails: laurieesteepeeters@gmail.com, david.deconinck@kuleuven.be, celine.wuyts@kuleuven.be and geert.loosveldt@kuleuven.be

implementation of these designs suggest that putting differential fieldwork efforts into different groups of the population will lead to less biased survey results, at lower costs (Chun et al. 2018; Schouten et al. 2017). Given the need and interest for efficient allocation of the limited resources available for fieldwork, it is necessary to further investigate the impact of reissuing on nonresponse rates and bias. The current article adds to the literature in this field by investigating the effectiveness of reissuing as a common fieldwork practice. While the general concept of reissuing is well documented, the organization of the practice and its effect on the composition of the respondent group has not been documented in great detail. We specifically explore whether the reissuing procedure can be optimized by using information about the response rates achieved by interviewers during the initial fieldwork phase (i.e., before the start of reissuing). In addition, we also investigate whether the risk of nonresponse bias is influenced by the reissuing procedure.

2. The Reissuing Procedure and the Role of the Interviewer

The term "reissuing" is used to describe the process of reattempting to contact sample units that initially did not participate in a survey. Although one can question whether high response rates should be a primary objective in survey practice (Beullens and Loosveldt 2012), the main objective of a reissue procedure is to increase these rates. Instead of accepting the targeted sample unit's initial response outcome, the survey organization may choose to make further attempts to convert these cases into interviews. These additional attempts can be made by the same interviewer, or a different one (Burton et al. 2006; Tarnai and Moore 2008). Reissuing can cover any or all sources of nonresponse (refusal, noncontact, and others, possibly including ineligibility). It can thus be considered a generalization of refusal conversion, which is aimed solely at converting those cases that refused to participate during the initial fieldwork phase.

Reissuing initial nonrespondents presumes that target respondents do not thoroughly deliberate the decision to participate or not, and a refusal decision may therefore be overturned (Groves and Couper 2012). A different interviewer applying a different contact schedule or a different doorstep approach may be able to convert an initial nonresponse outcome into a completed interview.

Interviewers play a crucial role in the success of both the initial fieldwork phase and any reissue activities. Previous research into interviewer effects on nonresponse clearly demonstrates significant differences between interviewers in terms of contact and cooperation rates (Blom et al. 2011; Durrant et al. 2010; O'Muircheartaigh and Campanelli 1999; Pickery and Loosveldt 2002), indicating that some interviewers are more successful than others in contacting targeted sample units and persuading them to cooperate. These interviewer effects also create differential nonresponse error across interviewers (West and Olson 2010), so it is doubtful that nonresponse within interviewers is completely random. This brings us to the question of which interviewers might be more successful at getting reissued nonrespondents to participate, and thus should preferably be selected for reissue activities?

Interviewers may be selected for reissuing on the basis of their gender, age, or ethnic background in order to maximize the likelihood of converting nonrespondents with similar characteristics (Gideon 2012). Typically, however, the better trained, more

experienced interviewers or those who achieved a high response rate in the first phase of the fieldwork or in previous surveys are selected for the reissue phase. Previous studies of the effectiveness of reissue procedures suggest that when highly performing interviewers from the initial phase are deployed in reissue activities, conversion rates are significantly affected in a positive way (Beullens et al. 2009; Stoop et al. 2014). These studies have not examined whether respondents interviewed by "high-performance" interviewers are different from respondents interviewed by their less successful colleagues in terms of socio-demographic or socio-political characteristics. One can assume that interviewers with lower response rates follow the line of least resistance and focus on the "low hanging fruit" (Beullens et al. 2009), mostly interviewing people with a higher response propensity (e.g., those spending more time at home or having greater interest in the topic of the survey). On the other hand, interviewers with higher response rates may be able to track, contact, and convince more reluctant targets and those that are harder to reach.

In this study, we investigate the process of reissuing and its effect on the survey response rate and the composition of the net sample in Round 7 and Round 8 of the European Social Survey (ESS) in Belgium. We do this by identifying the interviewers engaged in the initial fieldwork efforts and reissue phase, comparing their response rates in each phase of the fieldwork, and investigating socio-demographic and socio-political differences between the groups of respondents that are interviewed in each part of the fieldwork. We formulate the following hypotheses:

Hypothesis 1: Compared to interviewers who achieve low response rates in the initial phase, interviewers who achieve high response rates in the initial phase interview respondents with a different profile.

Hypothesis 2: Interviewers who achieve high response rates in the initial phase are also more successful, in terms of achieved response rates, in the reissue phase.

Hypothesis 3A: Nonrespondents approached in the initial phase by an interviewer with a high response rate are more difficult to convert during the reissue procedure.

Hypothesis 3B: Nonrespondents approached in the initial phase by an interviewer with a low response rate might have completed an interview during the first phase if they had been approached by an interviewer with a high response rate.

3. Data

Data from Round 7 and Round 8 of the European Social Survey (ESS) in Belgium are used to test the hypotheses. The same organization was responsible for fieldwork in both rounds, and there are several interviewers who were deployed in both rounds. Fieldwork for Round 7 started on 15 September 2014 and ended approximately 20 weeks later, on 1 February 2015. For Round 8, the fieldwork also took about 20 weeks (14 September 2016 until 31 January 2017). The total number of issued sample units for each round is 3,204 (Barbier et al. 2016; Wuyts et al. 2018). See the documentation reports of ESS7 (European Social Survey 2015) and ESS8 (European Social Survey 2017) for more information on the survey designs. Both rounds are separately analyzed so that we can compare and validate the results.

The reissue procedure is delineated by a new interviewer having been assigned to a case after the first interviewer had completed all contact attempts in the initial contact phase without obtaining an interview (Gideon 2012). In the current article, all types of initial nonresponse (noncontacts, refusals, other nonresponse, and ineligibility) are included. It was specified in the contract with the fieldwork organization that these sample units would be re-allocated to a different interviewer, who would apply the contact procedure all over again, using the information available from previous contacts. The interviewers used in the refusal conversion procedure would be selected based on their response performance in the running ESS project and previous projects. Within the constraints of these contract specifications, the actual selection of the interviewers and the sample units for reissuing was at the discretion of the survey agency.

For Round 7, the 151 interviewers that began fieldwork completed a total of 1,506 interviews out of the 3,204 potential respondents (initial response rate = 47%). Out of the 1,698 individuals that did not complete an interview, 1,040 (61%) were reissued and re-attempted face-to-face by a new interviewer. Round 8 started off with 139 interviewers who completed 1,475 interviews in the initial phase (initial response rate = 46%). Half of the initial nonrespondents (861 cases) were reissued in the traditional, face-to-face mode. Some 389 additional cases were assigned to the survey agency's call center and assigned to a face-to-face interviewer only after agreeing to participate. In order to ensure comparability, this latter part of the reissuing procedure in Round 8 is not included in the analysis of the reissue phase. Reissuing was finalized at these numbers, because an acceptable response rate was achieved and to increase it further would have required additional means that were not available.

Most cases selected for reissues in either round had refused to participate in the initial phase. Initial noncontacts also make up a large proportion of reissues, whilst the other two sample unit groups (initial "others" and ineligibles) are much smaller (see Table 8 in Appendix, Section 7).

The distribution of response outcomes after the initial and the reissue phase for Rounds 7 and 8 indicates that a total of 263 additional interviews in Round 7 (25% of the reissued cases), and 225 interviews in Round 8 (26% of the reissued cases) were conducted during the reissue phase, most of which were initial refusals. The response rate for reissues was thus about half as large as the response rate of the initial phase (25% and 47% in Round 7, 26% and 46% in Round 8). At the end of fieldwork activities in Round 7, 1769 interviews had been obtained (response rate = 57%) and at the end of Round 8, 1766 interviews were realized (response rate = 56.8%). For more information on the response outcomes after the initial and reissue phase, see Table 9 and Table 10 in Appendix. For both rounds, the objective of attaining an acceptable response rate was realized by virtue of the extensive reissuing of nonrespondents (adding 10-11 percentage points). In the next section, we investigate which interviewers were selected and were most successful in the reissue procedure.

4. Interviewers' Response Rates and the Impact on Respondent Profiles During the Initial Fieldwork Phase

In Round 7, the average response rate of an interviewer in the initial phase was 48.0% (SD = 19.3). Only 10 interviewers (6.6%) had a response rate below or equal to 20%,

Table 1. Interviewer variance and the ICCs for the multilevel logistic null models with the dependent variable response/nonresponse in the initial phase and the interviewer at the second level.

	ESS7	ESS8
Interviewer variance	0.3529*	0.3547*
ICC	0.0969	0.0973

*$p < 0.001$

40 interviewers (26.5%) had a rate between 20% and 40%, 62 interviewers (41.1%) had a rate between 40% and 60%, and 39 interviewers (25.8%) achieved a response rate more than 60%. In Round 8, the average response rate of an interviewer was 45.3% ($SD = 17.4$). Only 11 interviewers (7.9%) had a response rate below or equal to 20%, 40 interviewers (28.8%) had a rate between 20% and 40%, 57 interviewers (41.0%) had a rate between 40% and 60%, and 31 interviewers (22.3%) achieved a response rate over 60%.

In order to formally test these differences, a multilevel logistic regression model with sample units nested within interviewers was estimated to predict the probability of obtaining an interview in the initial phase for both rounds. The interviewer variance for the null model, including only the random intercept effect and no explanatory variables, and the intraclass correlation coefficients (ICCs) are presented in Table 1. The ICCs for the two null models suggest that 9.7% of the variability in the obtained participation of sample units in the initial phase is explained by between-interviewer differences. These estimates provide evidence that some interviewers are more successful than others.

4.1. Hypothesis 1: Compared to interviewers who achieve low response rates in the initial phase, interviewers who achieve high response rates in the initial phase interview respondents with a different profile

The question is whether interviewers with higher response rates interview respondents with different socio-demographic or socio-political profiles than interviewers with lower response rates. To answer this question and test the first hypothesis, interviewers were divided into two groups: those with an individual response rate in the initial phase higher than the overall response rate of the survey and those with an individual response rate in the initial phase lower than the overall response rate. The overall response rates for Rounds 7 and 8 are 57.0% and 56.8%, respectively. These cut-offs are also used in the following sections to differentiate interviewers with low response rates and those with high response rates. We expect interviewers with a low response rate to mainly interview respondents who are more easily persuaded to participate, whereas interviewers with a high response rate in the initial phase are expected to be better at contacting and persuading more reluctant or hard-to-reach-respondents.

In Table 2, we compare respondent profiles of interviewers with a low response rate to those from interviewers with a high response rate. There are no significant differences in respondent profiles between the two groups of interviewers in either round regarding respondents' age, gender, work situation, and the presence of children in the household. The expectation that interviewers with high response rates would interview more hard-to-reach-respondents due to being employed or having children in the household is thus not

Table 2. Background characteristics of respondents interviewed in the initial phase, by interviewers' initial response rate.

	ESS7				ESS8			
	Interviewers with a low response rate		Interviewers with a high response rate		Interviewers with a low response rate		Interviewers with a high response rate	
	#	%	#	%	#	%	#	%
Age ($\chi^2 = 0.60$; df = 3; p = 0.8971)					($\chi^2 = 1.81$; df = 3; p = 0.6128)			
14–30	197	22.4	150	23.9	186	23.0	167	25.0
31–45	199	22.6	135	21.5	186	23.0	153	22.9
46–60	227	25.8	159	25.4	206	25.5	176	26.4
+60	256	29.1	183	29.2	230	28.5	171	25.6
Gender ($\chi^2 = 0.06$; df = 1; p = 0.8072)					($\chi^2 = 0.00$; df = 1; p = 0.9931)			
Male	440	50.1	318	50.7	406	50.3	335	50.2
Female	439	49.9	309	49.3	402	49.8	332	49.8
Income ($\chi^2 = 4.24$; df = 2; p = 0.12)					($\chi^2 = 6.64$; df = 2; p = 0.0361)			
Low income	169	21.0	123	21.4	164	21.2	114	16.5
Average income	382	47.4	298	51.9	391	50.6	361	52.1
High income	255	31.6	153	26.7	218	28.2	153	31.5
Political interest ($\chi^2 = 5.96$; df = 1; p = 0.0147)					($\chi^2 = 0.61$; df = 1; p = 0.4343)			
Very/quite interested	454	51.7	283	45.1	398	49.3	314	47.1
Hardly/not at all interested	425	48.4	344	54.9	410	50.7	353	52.9
Trust in politicians ($\chi^2 = 4.04$; df = 2; p = 0.1329)					($\chi^2 = 15.04$; df = 2; p = 0.0005)			
Low trust	306	35.0	238	38.1	320	39.7	219	33.0
Average trust	423	48.3	305	48.8	374	46.4	375	56.5
High trust	146	16.7	82	13.1	112	13.9	70	10.5
Trust in the European Parliament ($\chi^2 = 7.13$; df = 2; p = 0.0283)					($\chi^2 = 0.3227$; df = 2; p = 0.851)			
Low trust	213	24.6	184	29.6	239	29.8	190	28.8
Average trust	389	44.9	282	45.3	365	45.5	310	47.0
High trust	264	30.5	156	25.1	198	24.7	160	24.2

Table 2. Continued

	ESS7				ESS8			
	Interviewers with a low response rate		Interviewers with a high response rate		Interviewers with a low response rate		Interviewers with a high response rate	
	#	%	#	%	#	%	#	%
Paid work ($\chi^2 = 0.72$; df = 1; p = 0.3975)					($\chi^2 = 0.29$; df = 1; p = 0.5932)			
Paid work in the last 7 days	442	50.3	330	52.6	417	51.6	334	50.1
No paid work in the last 7 days	437	49.7	297	47.4	391	48.4	333	49.9
Children ($\chi^2 = 0.41$; df = 1; p = 0.5213)					($\chi^2 = 0.00$; df = 1; p = 0.9844)			
Children at home	357	40.7	243	38.9	319	39.5	263	39.4
No children at home	521	59.3	382	61.1	489	60.5	404	60.6
Minority group identity ($\chi^2 = 4.05$; df = 1; p = 0.0444)					($\chi^2 = 8.51$; df = 1; p = 0.0035)			
Belonging to a minority group	43	4.9	17	2.7	25	3.1	43	6.5
Not belonging to a minority group	832	95.1	609	97.3	778	96.9	622	93.5
Total	879		627		808		667	

confirmed. In Round 8, there is a significant difference concerning respondents' income: interviewers with a high initial response rate are more likely to interview respondents with a higher income than interviewers with a low initial response rate. It must be noted that there is some item nonresponse for the income variable, but this is never larger than 10%. Although the results are mixed, there are some indications that interviewers with a high response rate interview a greater number of respondents with less interest and trust in politics. In line with expectations, there is a tendency for interviewers with a high response rate to be more successful in interviewing respondents that are hardly or not at all interested in politics (only significant in Round 7), respondents with average confidence in politicians (only for Round 8) and respondents with little confidence in the European Parliament (only for Round 7). In Round 8, interviewers with a high response rate also interviewed more respondents belonging to a minority group, but the opposite pattern is observed for Round 7.

When analysing respondents' characteristics for interviewers with a low response rate as opposed to interviewers with a high response rate, no strong evidence is found to support the hypothesis that interviewers who achieve high response rates in the initial phase also interview respondents with a different socio-demographic or socio-political profile compared to interviewers with low response rates in the initial phase.

5. The Impact of the Interviewers on the Success of the Reissue Phase

After the initial phase, initial nonrespondents were assigned to another interviewer to re-attempt contact and persuade them to participate. In Table 3, respondent characteristics of both phases are compared. In both rounds, significantly more respondents with a low interest in politics were interviewed in the reissue phase than in the initial phase. Other significant differences between the initial and the reissue phase arise, but are not consistent across the two rounds. In Round 7, a smaller percentage of respondents with a high level of trust in the European Parliament and a higher percentage of respondents from a minority group were interviewed in the reissue phase than in the initial phase. In Round 8, a higher percentage of young people, people with a lower income, and people with an average trust in politicians were interviewed. Although the specifics are inconsistent, these results suggest that different types of respondents are interviewed in the reissue phase.

Typically, only a subset of the interviewers working in the initial phase of fieldwork are selected to participate in the reissue phase. Of the 46 interviewers with a high response rate in the initial phase of Round 7, 27 (58.7%) were engaged in the reissue phase. Most of the interviewers in the reissue phase of Round 7 (53 out of 62 or 85%) had a respectable to very good response rate (over 40%) in the initial contact phase. In Round 8, a similar pattern emerges: of the 39 interviewers with a high response rate in the initial phase, 27 (69.2%) were employed in the reissue phase, and 63% of the interviewers used for reissuing had an initial response rate higher than 40%. These results are in line with expectations based on previous literature and experience with survey agencies that suggest that the selection of the interviewers is driven by their recorded performance (in addition to availability and travel costs), and that response rates that were previously

Table 3. Background characteristics of respondents interviewed in the initial and reissue phase.

	ESS7				ESS8			
	Initial phase		Reissue phase		Initial phase		Reissue phase	
	#	%	#	%	#	%	#	%
Age ($\chi^2 = 2.38$; df = 3; p = 0.4968)					($\chi^2 = 11.08$; df = 3; p =0.0113)			
14–30	347	23.0	63	24.0	353	23.9	66	29.3
31–45	334	22.2	60	22.8	339	23.0	64	28.4
46–60	386	25.6	75	28.5	382	25.9	54	24.0
+60	439	29.2	65	24.7	401	27.2	41	18.2
Gender ($\chi^2 = 3.16$; df = 1; p = 0.0756)					($\chi^2 = 0.45$; df = 1; p = 0.5032)			
Male	749	49.7	147	55.9	741	50.2	119	52.9
Female	757	50.3	116	44.1	734	49.8	106	47.1
Income ($\chi^2 = 2.75$; df = 2; p = 0.2524)					($\chi^2 = 12.56$; df = 2; p = 0.0019)			
Low income	292	21.2	50	20.9	278	19.8	53	24.8
Average income	680	49.3	130	54.4	752	53.7	128	59.8
High income	408	29.6	59	24.7	371	26.5	33	15.4
Political interest ($\chi^2 = 6.35$; df = 1; p = 0.0118)					($\chi^2 = 13.98$; df = 1; p = 0.0002)			
Very/quite interested	737	48.9	106	40.3	712	48.3	78	34.7
Hardly/not at all interested	769	51.1	157	59.7	763	51.7	147	65.3
Trust in politicians ($\chi^2 = 1.99$; df = 2; p = 0.3706)					($\chi^2 = 8.22$; df = 2; p = 0.0164)			
Low trust	544	36.3	97	37.2	539	36.7	68	30.5
Average trust	728	48.5	133	51.0	749	51.0	136	61.0
High trust	228	15.2	31	11.9	182	12.4	19	8.5
Trust in the European Parliament ($\chi^2 = 18.85$; df = 2; p = 0.0001)					($\chi^2 = 5.26$; df = 2; p = 0.0721)			
Low trust	397	26.7	83	31.9	429	29.3	58	26.0
Average trust	671	45.1	137	52.7	675	46.1	121	54.3
High trust	420	28.2	40	15.4	358	24.5	44	19.7

Table 3. Continued

	ESS7				ESS8			
	Initial phase		Reissue phase		Initial phase		Reissue phase	
	#	%	#	%	#	%	#	%
Paid work ($\chi^2 = 0.50$; df = 1; p = 0.4783)					($\chi^2 = 0.23$; df = 1; p = 0.6312)			
Paid work in the last 7 days	772	51.3	128	48.7	751	50.9	119	52.9
No paid work in the last 7 days	734	48.7	135	51.3	724	49.1	106	47.1
Children ($\chi^2 = 2.44$; df = 1; p = 0.1182)					($\chi^2 = 0.44$; df = 1; p = 0.5079)			
Children at home	600	39.9	91	34.6	582	39.5	83	36.9
No children at home	903	60.1	172	65.4	893	60.5	142	63.1
Minority group identity ($\chi^2 = 19.66$; df = 1; p = 0.0000)					($\chi^2 = 0.72$; df = 1; p = 0.3972)			
Belonging to a minority group	60	4.0	28	10.7	68	4.6	7	3.1
Not belonging to a minority group	1441	96.0	234	89.3	1400	95.4	217	96.9
Total	1506		263		1475		225	

achieved (e.g., in an earlier phase of fieldwork) may serve as a useful criterion in the selection.

Three additional hypotheses about the impact of the interviewers' response rates in the initial phase on the results in the reissue phase are tested.

5.1. Hypothesis 2: Interviewers who achieve high response rates in the initial phase are also more successful, in terms of achieved response rates, in the reissue phase

Table 4 presents the mean response rates achieved in the reissue phase for interviewers with a low initial response rate and those with a high one. These results support the hypothesis: interviewers with a high response rate in the initial phase achieve a higher response rate in the reissue phase, but this difference at the interviewer level is not statistically significant. In the reissue phase, the interviewers fail to realize response rates that are comparable to those in the initial phase. This indicates that sample units in the reissue phase are more difficult to convert. It is clear that the sample units that were still not interviewed after the reissue phase may be considered as "high hanging fruit", but the characteristics of this remaining group of nonrespondents cannot be observed.

To elaborate and refine the results at the sample unit level, a multilevel logistic regression model with a random intercept is estimated to predict the probability of an interview in the reissue phase based on the interviewers' response rates in the initial phase. For both rounds, the results indicate that the interviewers' initial response rate is indeed a positive predictor of target sample units' participation in the reissue phase. Interviewers who achieved high response rates in the initial phase had a greater likelihood of obtaining participation from the sample units assigned to them in the reissue phase. This is in line with the results in Table 4.

5.2. Hypothesis 3A: Nonrespondents approached in the initial phase by an interviewer with a high response rate are more difficult to convert during the reissue procedure

Although strong and systematic differences between the respondent groups interviewed in the initial phase by interviewers with a high response rate and by interviewers with a low response rate could not be identified, one assumes that sample units unsuccessfully approached by interviewers with a high response rate in the initial phase will be more difficult to convert during the reissue phase than those approached by interviewers with a low response rate in the initial phase.

In Round 7, there is a higher percentage of reissued sample units participating in the survey if they were approached by an interviewer with a lower response rate in the initial

Table 4. Mean response rates in the reissue phase for interviewers with low and high initial response rates.

	Low initial response rate	High initial response rate
ESS7 ($t = -1.58$; $df = 54.42$; $p = 0.06$)	19.35%	28.08%
ESS8 ($t = -0.68$; $df = 171.74$; $p = 0.25$)	26.16%	29.13%

phase (26.1%) than if they were approached by an interviewer with a higher response rate in the initial phase (21.4%), but the difference (4.7 percentage points) is not statistically significant ($X^2 = 1.42$; $df = 1$; $p = 0.2313$). In Round 8, the opposite pattern is observed, as there appears to be a higher but not statistically significant percentage of participating reissued sample units if they were approached by an interviewer with a higher initial response rate (27.2%) rather than by an interviewer with a lower initial response rate (25.9%) ($X^2 = 0.04$; $df = 1$; $p = 0.8382$). However, when using the continuous response rates in a multilevel logistic model, taking into consideration the random effect of the interviewers, the initial interviewers' response rate becomes statistically significant (see Table 5). This suggests that there is an effect of the initial interviewers' response rate based on the continuous response rates instead of on the binary separation of low and high response rates. This estimate is negative, suggesting that nonrespondents who were approached by an interviewer with a higher response rate are more difficult to convert in the reissue phase, as suggested in the hypothesis.

5.3. *Hypothesis 3B: Nonrespondents approached in the initial phase by an interviewer with a low response rate might have completed an interview during the first phase if they had been approached by an interviewer with a high response rate*

Based on hypothesis 3B, we expect that there is an interaction effect on the conversion rate between the response rate of the interviewer in the first phase (interviewer 1) and the initial response rate of the interviewer in the reissue phase (interviewer 2). Such an interaction effect would imply that nonrespondent sample units approached in the initial phase by an interviewer with a low response rate might very well have completed an interview during the first phase if they had been approached by an interviewer with a high response rate.

The results in Table 5 and Table 6 show that in both rounds, interviewers with high initial response rates successfully converted about one third of the reissued cases when these had been approached in the initial phase by an interviewer with a low response rate (Round 7: 35.7%; Round 8: 32.8%), which is significantly more than interviewers with low initial response rates (Round 7: 17.9% ($t = -5.9$; $df = 764.1$; $p < 0.01$); Round 8: 17.7% ($t = -2.3$; $df = 288.8$; $p = 0.03$). In Round 7, the difference in successfully interviewed reissued cases between the group of interviewers with high initial response rates and those with low initial response rates is larger for the sample

Table 5. Distribution of the conversion rate by response rate of interviewer 1 and interviewer 2 for Round 7.

Response rate interviewer 1	Response rate interviewer 2	Conversion				Total	
		No		Yes			
		#	%	#	%	#	%
Low	Low	376	82.1	82	17.9	458	53.3
	High	258	64.3	143	35.7	401	46.7
High	Low	79	82.3	17	17.7	96	55.8
	High	56	73.7	20	26.3	76	44.2

Table 6. Distribution of the conversion rate by response rate of interviewer 1 and interviewer 2 for Round 8.

Response rate interviewer 1	Response rate interviewer 2	Conversion				Total	
		No		Yes			
		#	%	#	%	#	%
Low	Low	390	76.2	122	23.8	512	74.0
	High	121	67.2	59	32.8	180	26.0
High	Low	50	78.1	14	21.9	64	50.4
	High	40	63.5	23	36.5	63	49.6

units interviewed in the initial phase by an interviewer with a low response rate. In Round 8, the expected difference is present for the group of sample units that was approached in the initial phase by an interviewer with a low response rate. In the group with an interviewer with a high initial response rate, the difference in conversion rates is even larger.

However, the interaction between the two initial response rates is not statistically significant; both either tested with the binary or with the continuous response rate variables. This implies that there is no difference in the second interviewer's conversion rate of the sample units based on the initial interviewer's response rate. However, when only including the main effects, the initial response rates of the interviewers become significant (Table 7). These results apply to both Round 7 and Round 8. As expected, there is a negative main effect of the first interviewer's initial response rate on conversion in the reissue phase: the higher the response rate of the interviewer in the initial phase, the lower the conversion rate in the reissue phase. There is, also as expected, a positive main effect of the second interviewer's initial response rate: the higher the second interviewer's initial response rate, the higher the conversion rate. These coefficients are small due to the fact that a change of one percentage point in the response rate is small, but larger percentage differences will have a greater impact on the conversion of the sample unit.

The results of the logistic regression analysis (Table 7) also indicate that in both rounds, conversion rates are significantly higher for soft refusals (refusals for which the initial interviewer indicates that the sample unit might still be persuaded to participate) and the "other" nonresponse category. In Round 7, the conversion rate is significantly lower for female respondents, as well as for the 31–45 and the over 60 age groups when compared with the youngest category.

6. Conclusion and Discussion

Our results, based on data from two rounds of the European Social Survey in Belgium, suggest that in both the initial and the reissue phase, some interviewers are more successful than others in contacting sample units and convincing them to participate in the survey. Nevertheless, there is no strong evidence that interviewers with a higher response rate interview more reluctant sample units in the initial phase than interviewers with a lower response rate.

Table 7. Logistic regression models predicting conversion in the reissue phase (odds ratios).

	Conversion of reissues	
	ESS7	ESS8
Constant	0.13***	0.23***
Initial contact outcome		
Soft refusal	2.04**	2.16*
Noncontact	1.32	1.25
Other	3.31**	2.58*
Age of case		
31–45 years old	0.55**	0.88
46–60 years old	0.70	0.90
+60 years old	0.64*	0.93
Gender of case		
Female	0.70*	0.89
Initial response rates of interviewers		
Initial response rates interviewer 1	0.98***	0.98***
Initial response rates interviewer 2	1.03***	1.02***

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$

Reference categories: "Hard refusal" (for initial contact outcome), "15–30 years old" (for age), and "Male" (for gender).

Interviewers with a higher response rate in the initial phase are also more successful in the reissue phase, and there is evidence that nonrespondents approached by an interviewer with a high response rate in the initial phase are more difficult to convert during the reissue phase.

Another question is whether the reissue phase is worth the investment. The most important effect is increasing the response rate, with an increase of the statistical power as a result. The higher response rate may also increase people's confidence in the results. It can be noted that the response rate can also be increased by using higher-skilled and successful interviewers from the beginning and eliminating interviewers with low response rates. For future ESS rounds, we recommend looking at the skills of interviewers based on participation in the fieldwork of previous ESS rounds and similar projects if possible and use this information to select the best possible interviewers for both the initial and the reissue phase. The impact of reissuing on the risk of nonresponse bias seems limited, with only marginal changes found in the socio-demographic and socio-political composition of the respondent groups in the initial and reissue phase. There are some indications that different types of respondents are interviewed during the reissue phase, but the differences are small and inconsistent for the two rounds examined. As the effect of re-issuing on the sample composition is not particularly large, the impact on nonresponse bias is likely not large either.

It can be said that the sample units that have still not been interviewed after the reissue phase are very difficult to contact and persuade to take part in an interview. However, based on the available content variables, it was not possible to further characterize this group of ultimate nonrespondents.

The information about the implementation of the refusal conversion procedure during the fieldwork is limited. According to the contract specifications with the fieldwork

organization, we can assume that the fieldwork organization selected the best performing interviewers that were available in an area. For future research, more transparency of the survey agency regarding the selection of the interviewers (e.g., their initial response rate, years of experience, socio-demographic characteristics, type of region – urban or rural – in which they work) and sample units (e.g., initial response outcome, sociodemographic characteristics) for the reissue phase would be desirable. Especially more information about the selection of the sampling units that are being reissued can contribute to a more thorough evaluation of the procedure. We can assume that the fieldwork organization applies a pragmatic approach. It is important to know which considerations (e.g., geographical proximity of sample units, availability of interviewers with, type of nonresponse) play a role in this process, and how different considerations are weighed against each other.

Another limitation of this study lies in the differential nature of the contact procedure in the reissue phase: in Round 8, 389 reissued cases were assigned to call center reissuing and went on to a face-to-face interview only after agreeing to participate, while all reissues in Round 7 were done via the face-to-face mode. Because of this discrepancy, these 389 cases were only included in the analysis of the initial nonresponse. This means that part of the picture on reissue activities in Round 8 was not taken into account. However, it is possible that reissued cases in the ESS8 may be more difficult sample units than in ESS7. Future studies that evaluate the effectiveness of reissues in increasing response rates and reducing bias (especially in repeated surveys such as the European Social Survey) should consider that survey organizations try to optimize their scarce means every step of the way to maximize the response rate and reduce bias. Research into the impact of the contact procedure (e.g., telephone versus face-to-face) during the reissuing phase may contribute to this.

7. Appendix

Table 8. Distribution of initial response outcomes of cases selected for reissuing.

		Noncontact	Refusal	Other	Ineligible	Total
Reissued cases ESS7	#	263	740	33	4	1040
	%	25.3	71.2	3.2	0.4	100
Reissued cases ESS8	#	298	530	29	4	861
	%	34.6	61.6	3.4	0.5	100

Table 9. Distribution of initial (after the initial phase) and final (after the reissue phase) response outcomes, Round 7 of the ESS (Belgium).

		Reissue outcome													
		Interview		Noncontact		Refusal		Other		Ineligible		Not selected		Total	
		#	%	#	%	#	%	#	%	#	%	#	%	#	%
Initial outcome	Interview	.	.	.	.	.	.	.	.	.	.	1506	69.6	1506	47.0
	Noncontact	79	30.0	54	75.0	105	17.2	15	21.1	10	45.5	100	4.6	363	11.3
	Refusal	168	63.9	12	16.7	498	81.4	52	73.2	10	45.5	225	10.4	965	30.1
	Other	16	6.1	3	4.2	9	1.5	4	5.6	1	4.5	253	11.7	286	8.9
	Ineligible	0	0.0	3	4.2	0	0.0	0	0.0	1	4.5	80	3.7	84	2.6
	Total	263	100	72	100	612	100	71	100	22	100	2164	100	3204	100

Table 10. Distribution of initial (after the initial phase) and final (after the reissue phase) response outcomes, Round 8 of the ESS (Belgium).

		Reissue outcome															
		Interview		Noncontact		Refusal		Other		Ineligible		Not selected		Call centre		Total	
		#	%	#	%	#	%	#	%	#	%	#	%	#	%	#	%
Initial outcome	Interview	.	.	.	.	.	.	.	.	.	.	1475	75.5	.	.	1475	46.0
	Noncontact	88	39.1	90	88.2	85	19.1	21	30.4	14	66.7	98	5.0	63	16.2	459	14.3
	Refusal	122	54.2	10	9.8	348	78.4	44	63.8	6	28.6	94	4.8	315	81.0	939	29.3
	Other	14	6.2	2	2.0	10	2.3	3	4.3	0	0.0	216	11.1	11	2.8	256	8.0
	Ineligible	1	0.4	0	0.0	1	0.2	1	1.4	1	4.8	71	3.6	0	0	75	2.3
	Total	225	100	102	100	444	100	69	100	21	100	1954	100	389	100	3204	100

8. References

Barbier, S., C. Wuyts, P. Italiano, and G. Loosveldt. 2016. *European Social Survey Round 7 Belgium*: Process evaluation for the data collection. Leuven: KU Leuven.

Beullens, K., J. Billiet, and G. Loosveldt. 2009. *Selection strategies for refusal conversion of four countries in the European Social Survey*, 3rd round. Leuven: KU Leuven.

Beullens, K. and G. Loosveldt. 2012. "Should high response rates really be a primary objective?" *Survey Practice* 5(3): 1–5. DOI: https://doi.org/10.29115/SP-2012-0019.

Beullens, K., G. Loosveldt, C. Vandenplas C., and I. Stoop. 2018. "Response rates in the European Social Survey: Increasing, decreasing, or a matter of fieldwork efforts? Survey methods: Insights from the field." Available at: https://surveyinsights.org/?p = 9673 (accessed May 2020).

Blom, A., E. de Leeuw, and J. Hox. 2011. "Interviewer effects on nonresponse in the European Social Survey." *Journal of Official Statistics* 27(2): 359–377. Available at: https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/interviewer-effects-on-nonresponse-in-the-european-social-survey.pdf (accessed May 2020).

Brick, J.M. and R. Tourangeau. 2017. "Responsive survey designs for reducing nonresponse bias." *Journal of Official Statistics* 33(3): 735–752. DOI: https://doi.org/10.1515/jos-2017-0034.

Burton, J., H. Laurie, and P. Lynn. 2006. "The long-term effectiveness of refusal conversion procedures on longitudinal surveys." *Journal of the Royal Statistical Society. Series A (Statistics in Society)* 169(3): 459–478. DOI: https://doi.org/10.1111/j.1467-985X.2006.00415.x.

Chun, A.Y., S.G. Heeringa, and B. Schouten. 2018. "Responsive and adaptive design for survey optimization." *Journal of Official Statistics* 34(3): 581–597. DOI: https://doi.org/10.2478/jos-2018-0028.

Curtin, R., S. Presser, and E. Singer. 2000. "The effects of response rate changes on the index of consumer sentiment." *Public Opinion Quarterly* 64(4): 413–428. DOI: https://doi.org/10.1086/318638.

Durrant, G., R. Groves, L. Staetsky, and F. Steele. 2010. "Effects of interviewer attitudes and behaviors on refusal in household surveys". *Public Opinion Quarterly* 74(1): 1–36. DOI: https://doi.org/10.1093/poq/nfp098.

European Social Survey. 2015. ESS Round 7 (2014/2015) Technical Report. London: ESS ERIC.

European Social Survey. 2017. ESS Round 8 (2016/2017) Technical Report. London: ESS ERIC.

Gideon, L. 2012. *Handbook of survey methodology for the social sciences*. New York: Springer.

Groves, R. 2006. "Nonresponse rates and nonresponse bias in household surveys." *Public Opinion Quarterly* 70(5): 646–675. DOI: https://doi.org/10.1093/poq/nfl033.

Groves, R.M. and M.P. Couper. 2012. *Nonresponse in household interview surveys*. Hoboken, New Jersey: John Wiley & Sons.

Lynn, P. and P. Clarke. 2002. "Separating refusal bias and non-contact bias: Evidence from UK national surveys." *Journal of the Royal Statistical Society* 51(3): 319–333. DOI: https://doi.org/10.1111/1467-9884.00321.

O'Muircheartaigh, C. and P. Campanelli. 1999. "A multilevel exploration of the role of interviewers in survey non-response." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 162(3): 437–446. DOI: https://doi.org/10.1111/1467-985X.00147.

Pickery, J. and G. Loosveldt. 2002. "A multilevel multinomial analysis of interviewer effects on various components of unit nonresponse." *Quality and Quantity* 36(4): 427–437. DOI: https://doi.org/10.1023/A:1020905911108.

Schouten, B., A. Peytchev, and J. Wagner. 2017. "Adaptive Survey Design." Series on Statistics Handbooks, Chapman and Hall/CRC.

Stoop, I. 2004. "Surveying nonrespondents." *Field Methods* 16(1): 23–54. DOI: https://doi.org/10.1177/1525822X03259479.

Stoop, I., A. Koch, V. Halbherr, R. Fitzgerald, and S. Widdop. 2014. *Field procedures in the European Social Survey Round 7: Enhancing response rates*. The Hague: European Social Survey, SCP. Available at: http://www.europeansocialsurvey.org (accessed June 2020).

Tarnai, J. and D.L. Moore. 2008. "Measuring and improving telephone interviewer performance and productivity." In *Advances in telephone survey methodology*, edited by E. de Leeuw, L. Japec, P.J. Lavrakas, M.W. Link, and R.L. Sangster. pp. 359–384, Hoboken, New Jersey: John Wiley & Sons.

West, B.T. and K. Olson. 2010. "How much of interviewer variance is really nonresponse error variance?" *Public Opinion Quarterly* 74(5): 1004–1026. DOI: https://doi.org/10.1093/poq/nfq061.

Wright, G. 2015. "An empirical examination of the relationship between nonresponse rate and nonresponse bias." *Statistical Journal of the IAOS* 31(2): 305–315. DOI: https://doi.org/10.3233/sji-140844.

Wuyts, C., L. Jacobs, D. de Coninck, P. Italiano, and G. Loosveldt. 2018. *European Social Survey Round 8 Belgium: Process evaluation for data collection*. Leuven: KU Leuven.

Received September 2018
Revised April 2019
Accepted April 2020

sciendo

Journal of Official Statistics, Vol. 36, No. 3, 2020, pp. 609–629, http://dx.doi.org/10.2478/JOS-2020-0031

Implementing Adaptive Survey Design with an Application to the Dutch Health Survey

Kees van Berkel[1], *Suzanne van der Doef*[1], *and Barry Schouten*[2]

Adaptive survey design has attracted great interest in recent years, but the number of case studies describing actual implementation is still thin. Reasons for this may be the gap between survey methodology and data collection, practical complications in differentiating effort across sample units and lack of flexibility of survey case management systems. Currently, adaptive survey design is a standard option in redesigns of person and household surveys at Statistics Netherlands and it has been implemented for the Dutch Health survey in 2018. In this article, the implementation of static adaptive survey designs is described and motivated with a focus on practical feasibility.

1. Introduction

Adaptive survey design assumes that differentiation of effort over relevant population subgroups is either effective in improving survey quality or efficient in reducing survey costs. The designs have received a lot of interest over the last decade in response to budget pressure due to gradual but persistent declines of response rates, for example Chun et al. (2018). Despite the interest in the designs, the number of implemented case studies described in the literature is still relatively small. This article discusses implementation of adaptive survey designs at Statistics Netherlands as it has been initiated in 2016.

National Statistical Offices have the task of publishing reliable and coherent statistical information that responds to the needs of society. In order to maintain a good balance between quality, efficiency and cost-effectiveness, continuous evaluation and improvement of processes and working methods are necessary. In 2016, four data collection policy decisions were made at Statistics Netherlands in order to arrive at a more efficient data collection strategy: incentives were used to increase overall response rates, a second supplier of telephone numbers was deployed so that more telephone observation is possible, follow-up sample sizes for CATI and CAPI were fixed in order to stabilise interviewer workload, and adaptive survey design became a standard design choice in sequential mixed-mode surveys. These four changes were implemented to varying degrees for a large number of surveys since 2017. Here, the fourth change is discussed.

Adaptive and responsive survey designs exist since the early days of surveys, but until about 15 years ago had never been named explicitly nor been the subject of structured and formalised research. Groves and Heeringa (2006), Wagner (2008) and Luiten and

[1] Statistics Netherlands, Division of Data Services, Research and Innovation, P.O. Box 4481, 6401 CZ, Heerlen, the Netherlands.
[2] Statistics Netherlands, Division of Data Services, Research and Innovation, P.O. Box 24500, 2490 HA, Den Haag, the Netherlands.

Schouten (2013) give early accounts of such designs. So why did it take so long to implement the designs in practice? At least three reasons can be given. First, design changes may have an impact on the survey results. Second, the designs require very flexible case management systems and monitoring. Third, practical and logistical constraints hamper a translation of optimal designs to implemented designs. In this article, a case study is presented that has been implemented and in which practical considerations played an important role.

Adaptive survey designs have four main elements: quality and cost objectives, design features, stratification of the target population, and an optimisation and implementation strategy. See Schouten et al. (2017) for a general overview and Tourangeau et al. (2017) for a discussion. Here, the focus is on the fourth element, the optimisation and implementation strategy, within the context of mixed-mode surveys.

Optimisation approaches range from trial-and-error to case prioritisation to advanced mathematical programming. Two approaches are confronted with each other in the case study: case prioritisation, for example Peytchev et al. (2010), Wagner (2013) and Wagner and Hubbard (2013), and mathematical optimisation, for example see Schouten et al. (2013) and Kaputa and Thompson (2017). For practical reasons, it was decided that case prioritisation should be implemented, but realised results are compared to expected results of optimised designs.

This article reads as follows: Section 2 describes the methodology behind the adaptive survey design. Section 3 discusses the application to the Dutch Health Survey. Section 4 ends with a discussion of results, limitations and future activities.

2. Methodology

In this section, the four main elements of adaptive survey design, quality and cost criteria, design features, stratification and optimisation, are discussed. This is done from an operational perspective.

2.1. Quality Indicators

The adaptive survey design is focussed on optimising balance of response through the coefficient of variation (CV) of response propensities for relevant population subgroups. See Schouten et al. (2009), De Heij et al. (2015), and Moore et al. (2018).

The CV is based on the desire to limit the risk of nonresponse bias over a range of variables. A random response model is adopted and it is assumed that each population unit has a response probability and response of the unit is independent of other population units. The response probabilities are unknown and are replaced by estimated response propensities based on a set of relevant auxiliary variables.

Consider a sample survey with a target population of N people. Let persons be labelled by k in the target population, $k = 1, 2, \ldots, N$. For the survey, a single random sample without replacement with size n is drawn from the target population. Let a_k be the inclusion indicator for person k. This means that a_k is equal to 1 if person k is in the sample, and 0 if person k is not in the sample. The expected value of a_k is equal to the probability that person k is selected in the sample, $E(a_k) = n/N$. Each person k in the target population is assumed to have a response probability ρ_k, which is only known to person k. If person k

is selected in the sample, this person is subjected to a Bernoulli experiment that results in response with probability ρ_k and in nonresponse with probability $1-\rho_k$. Let r_k be the response indicator for person k, belonging to the corresponding Bernoulli experiment. So r_k is equal to 1 if person k responds and 0 if person k does not respond. The expected value of r_k is equal to the probability that person k responds, $E(r_k) = \rho_k$. The number of respondents r in the sample survey is a random variable $r = \sum_{k=1}^{N} a_k r_k$ with expected value $n\bar{\rho}$, where $\bar{\rho}$ denotes the mean response probability in the population.

The aim of the survey is the estimation of population means for several target variables. An estimator of the population mean $\bar{y}$ of variable y is the response mean,

$$\bar{y} = \frac{1}{r}\sum_{k=1}^{N} a_k r_k Y_k.$$

The response mean $\bar{y}$ is in general a biased estimator for the population mean $\bar{Y}$. If $\rho_k = \bar{\rho}$ for all k, then $\bar{y}$ is unbiased, but this is generally not true. Bethlehem (1988) shows that

$$B(\bar{y}) = E(\bar{y}) - \bar{Y} \approx \frac{1}{\bar{\rho}N}\sum_{k=1}^{N}(\rho_k - \bar{\rho})Y_k = \frac{1}{\bar{\rho}}cov(\rho, Y).$$

Here $cov(\rho, Y)$ is the population covariance between the response probabilities and the values of the target variable. Thus, there is no bias if there is no correlation between response propensity and the target variable. Introduce Pearson's correlation coefficient:

$$R(\rho, Y) = \frac{cov(\rho, Y)}{S_\rho S_Y},$$

where S_ρ is the standard deviation of the response probabilities and S_Y is the standard deviation of the values of the target variable. Then the bias approximation formula can be written as

$$B(\bar{y}) \approx \frac{R(\rho, Y)\, S_\rho\, S_Y}{\bar{\rho}}.$$

From this expression it follows:

1. $B(\bar{y}) = 0$ if there is no linear relationship between ρ and Y.
2. The stronger the linear relationship between ρ and Y, the larger $B(\bar{y})$.
3. $B(\bar{y}) = 0$ if there is no variation of response rates or no variation in the values of the target variable.
4. The smaller the variation of response rates, the smaller $B(\bar{y})$.
5. The smaller the variation in the values of the target variable, the smaller $B(\bar{y})$.
6. The greater the mean response rate, the smaller $B(\bar{y})$.

Since the absolute value of Pearson's correlation coefficient does not exceed 1, an upper limit for the bias can be given:

$$|B(\bar{y})| \leq \frac{S_\rho S_y}{\bar{\rho}} = CV(\rho)\, S_Y.$$

Here, $CV(\rho)$ denotes the coefficient of variation of the response probabilities. A lower CV for response propensities defined by auxiliary variables implies smaller nonresponse

biases on these variables. When auxiliary variables are associated with survey variables, then a lower *CV* also implies smaller nonresponse biases on these variables before weighting adjustment. However, a lower *CV* does not necessarily imply smaller nonresponse biases on these variables after weighting adjustment. Nevertheless, there are three reasons to still pursue a lower *CV*. The first reason is that a more balanced response leads to less variation in adjustment weights and, as a consequence, to a more efficient sampling design. The second reason is that Schouten et al. (2016) have shown that empirically nonresponse bias on survey variables is on average still smaller for more balanced response, even after weighting adjustment using the same auxiliary variables. This finding conforms to the intuition that a more balanced data collection is a sign of a more effective design, in general. The final reason is that balancing response forces survey designers to come up with strategies to raise response rates of the strata that are harder to contact and to get participation. In the remainder of this article, an attempt is made to minimise $CV(\rho)$ by interfering in the process of data collection.

2.2. *Design Features*

The focus in this article is on the mix of survey modes. It is assumed that a sequential mixed-mode design is used with CAWI (Computer-Assisted Web Interviewing) as the starting mode. Follow-up of CAWI nonresponse is done through interviewer modes. Here, it is assumed that the follow-up is done by CAPI (Computer-Assisted Personal Interviewing). The design feature to adapt is the CAPI follow-up.

In the sequential mode strategy, all sampled people are first asked by letter to participate in the survey by completing a questionnaire on the internet. People who have not responded to this request after no more than two reminders are visited at home to conduct an interview. The observation strategy of the face-to-face interviews is adjusted as follows. To reduce the variation of response rates, more CAPI is used for groups that respond badly via the internet than for groups that respond well. However, the entire sample starts with CAWI. The identification of these target groups is carried out using cluster analysis.

It is assumed in this article that the answers obtained are the same in different observation modes, that is, mode-specific measurement bias is absent and can be ignored. This is a simplification, as such biases are conjectured to exist and should then be incorporated in the design decisions within the adaptive survey design. Such inclusion of measurement biases is not straightforward as they are confounded with selection biases, unless experimental designs are used to disentangle the biases. The discussion section returns to this complication.

2.3. *Stratification of the Target Population*

Determining target groups is also called segmentation or clustering of the target population. The target groups are composed by means of response propensities of people per mode. This may mean that two target groups have approximately the same response rate at CAWI, but that their CAPI response rates differ. It is also possible that the total response rates of two target groups are approximately the same, but that their response rates differ per mode.

Clustering is performed with a classification tree algorithm. People are divided into groups based on personal characteristics. The algorithm divides the groups that differ most in response behaviour first. To ensure that reliable response rates per mode can be estimated for each target group, it is important that the target groups are not too small. To prevent this, a minimum size per target group can be set.

2.4. Optimisation

Two approaches to optimisation are explored: case prioritisation and mathematical optimisation, including expected yield of the face-to-face follow-up.

2.4.1. The Optimisation Problem

Let G be the set of groups used to determine the target groups. Each target group is the union of one or more groups from G. For each, $g \in G$, let $N(g)$ denote the population size of group g. For a simple random sample of size n, it is assumed that the size of the sample in group g equals $n(g) = n \cdot N(g)/N$.

Furthermore, for each group $g \in G$ it is assumed that all people have the same CAWI response probability $p_w(g)$, the same probability $p_e(g)$ of being eligible for face-to-face follow-up and the same CAPI response probability $p_p(g)$ in the face-to-face approached sample of group g. Let $f_p(g)$ be the CAPI sampling fraction in group g, that is the proportion of people to be approached face-to-face in the CAWI nonrespondents who are eligible for face-to-face follow-up in group g. The total response probability in group g equals

$$p(g) = p_w(g) + p_e(g) f_p(g) p_p(g).$$

This allows the mean response probability and the population variance of the response probabilities to be estimated:

$$\bar{\rho} = \frac{1}{N} \sum_{g \in G} N(g) p(g) \text{ and } S_\rho^2 = \frac{1}{N} \sum_{g \in G} N(g)(p(g) - \bar{\rho})^2.$$

The following problem needs to be solved.

$$\text{Minimise } CV(\rho) = S_\rho / \bar{\rho} \text{ under a specified number of constraints.}$$

Different types of constraints can be used:

- *Budget*. This can be done at different levels, such as an available budget for the total observation or per observation mode.
- *Capacity*. An upper limit can be specified for the sample size to be approached face-to-face. This can be at national or regional level.
- *Precision*. This concerns requirements for the number of respondents or the number of respondents per subpopulation.
- *Response rates*. For example, a minimum response rate, or minimum response rates per mode or per subpopulation.
- *Ratio of the CAWI/CAPI modes in the response*. For example, a minimum percentage of CAPI response in the total response, or minimal CAPI sampling fractions per target group.

One CAPI sampling fraction is used per target group. This leads to the extra constraint:

$$\text{For each target group } d \text{ and all groups } g_1,\ g_2 \subset d:\ f_p(g_1) = f_p(g_2) \text{ applies.}$$

The decision variables for which the minimum can be found, are the CAWI sample size n and the CAPI sampling fractions $f_p(d)$ per target group d.

The optimisation problem requires a search for the numbers of people to be approached by target group and observation mode. The lower the CAWI response propensity of a target group, the more face-to- face observation is applied. This may lead to a smaller variation of response rates, and the ratio of the target groups in the response may be more similar to the ratio of the target groups in the population. This may, however, be at the expense of the overall response rate.

2.4.2. Optimisation Approaches

Two approaches are elaborated: case prioritisation and mathematical optimisation.

Case prioritisation is based on the rationale that the weakest performing population subgroups need the most attention and need to be allocated first. Response propensities at the end of a data collection phase, in this article CAWI, are estimated and sorted in increasing order. The sample units or sample strata with the lowest propensities are re-approached until budget is depleted and/or other constraints are met. Case prioritisation does not guarantee that the coefficient of variation is actually decreased, since expected conditional response propensities in subsequent data collection phases are not included. Such conditional propensities may have an opposite order of size and may even deteriorate balance. Such opposite ranking is, however, unusual in practice.

Mathematical programming accounts for expected yield in follow-up data collection phases, as it includes follow-up response propensities. As such, it guarantees improvement under the condition that response propensities are estimated accurately. Here, the minimisation problem is solved with the Auglag function of the Alabama R package. This R package uses the "Augmented Lagrangian Adaptive Barrier Minimisation Algorithm for optimising smooth nonlinear objective functions with constraints". The optimisation problem of Subsection 2.4 is smooth and nonlinear, because the partial derivatives of the objective function, the coefficient of variation of the response probabilities exist, and the objective function is nonlinear. The problem is also solved with the solver in Excel. This solver uses the GRG nonlinear solver method to solve the nonlinear problem and this algorithm uses the generalised reduced gradient method. Because it is a nonlinear problem and the algorithm can end up in a local minimum, different random starting points were used and the best solution was selected.

3. Application of Adaptive Survey Design to the Dutch Health Survey

3.1. The Dutch Health Survey

The aim of the Dutch Health Survey is to provide as complete an overview as possible of developments in health, medical contacts, lifestyle and preventive behaviour of the population in the Netherlands. The target population consists of all people living in the Netherlands who do not belong to the institutional population. The sample is a stratified

two stage sample in which people with equal probabilities are selected. This sampling design is approximately the same as the simple random sampling design. The observation starts with CAWI and the re-approach mode is CAPI. As a response increasing measure, iPads are raffled among the sampled people.

3.2. Stratification

The classification tree algorithm is implemented in R with the rpart package. Demographic and regional characteristics have been used that are known to have a different response distribution than the population. Examples are ethnicity, ethnicity of parents, age, income, urbanity of the municipality, urbanity of the neighbourhood, living in the four largest cities, educational level, type of household, number of people in the household, place in the household, number of children, marital status, wealth, gender, and home ownership. For more details on the characteristics used, see Section 5, Appendix. The algorithm determines which characteristics are used to split the groups and in which order. For categorical variables, the algorithm also determines where to split. This ensures that, for example, for a variable such as age, a classification can be made that best matches the response behaviour.

The results of the classification tree algorithm are the characteristics used for the Health Survey to record the target groups: ethnicity (NL resident, western migrant, non-western migrant), age (in years), income (in quintiles) and urbanity of the municipality in which the person lives (very strongly urban, strongly urban, moderately urban, few urban, and non-urban). The algorithm ensures that the characteristics are merged into larger groups. Ethnicity is divided into two groups, namely western (NL residents and western migrants) and non-western (non-western migrants). Age is divided into four categories: 0–11, 12–24, 25–64, and 65+. The income used is the standardised household income and the classification is into two categories, with the low income category consisting of the lowest 20% and the high income category consisting of the remaining 80%. Urbanity is reduced to two categories, namely very strongly urban and all others. Figure 1 shows the classification tree. The tree is read from top to bottom. In each node a division is made

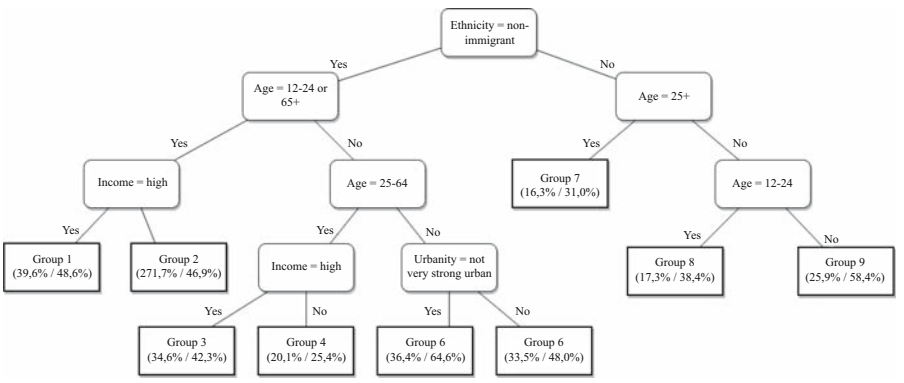

Fig. 1. Classification tree based on results of the Health Survey 2016.

target group	age	income	urbanity
1	12–24, 65+	high	-
2	12–24, 65+	low	-
3	25–64	high	-
4	25–64	low	-
5	0–11	-	2–5
6	0–11	-	1

Fig. 2. Partition of NL residents and western migrants into target groups.

based on a characteristic and the group is split. At the bottom of the tree, the ultimate target groups can be found, together with the response rates of CAWI and CAPI.

The first six target groups partition the NL residents and western migrants. Figure 2 contains an overview of these target groups. Here, urbanity = 1 means very strongly urban and urbanity = 2–5 means the union of the remaining categories. A dash means that there is no restriction for the variable in question.

The non-western migrants are divided into three target groups by age: 25+ in target group 7; 12–24 in target group 8 and 0–11 in target group 9.

3.3. The Dutch Health Survey Optimisation Problem

The set G of groups used to determine the target groups consists of 32 groups: ethnicity(2) × age(4) × income(2) × urbanity(2). The coefficient of variation of response probabilities $CV(\rho) = S_\rho/\bar{\rho}$ is estimated as described in Subsection 2.4.1. Minimising $CV(\rho)$ is carried out under the constraints:

- $n \leq n_{max}$ {CAWI sample size does not exceed n_{max}},
- $n \cdot \bar{\rho} \geq R$ {expected response size is at least R},
- $\sum_{g \in G} p_e(g) f_p(g) n(g) \leq C$ {total CAPI sample size is at most C },
- For each target group d and all groups $g_1, g_2 \subset d : f_p(g_1) = f_p(g_2)$ applies {one CAPI sampling fraction per target group}.

Here n_{max}, R and C are constants to be filled in. The parameters with which the minimum can be found are the CAWI sample size n and the CAPI sampling fractions for face-to-face observation $f_p(d)$ in the target groups d. Note that it follows from the first two constraints that $\bar{\rho} \geq R/n_{max}$.

In the case of the Health Survey 2018, the target groups with corresponding response rates per mode and probabilities of re-approachable CAWI nonresponse have been determined with data from the results of the Health Survey in January-June of 2017. The maximum CAWI sample size n_{max} has been set to 18,000 people. To be quite sure that 9,500 responses are achieved, the expected response size R has been set to 9,631 people. The maximum CAPI sample size C is 8,039 addresses, based on the available CAPI budget. The mean response rate must therefore be at least 9,631 / 18,000 = 53.5%.

3.4. Mathematical Optimisation

First, the mathematical optimisation approach is explored, as this approach may be used as a benchmark to the case prioritisation approach.

The minimisation problem is solved with the solver in R, with different random starting values for the CAWI sample size n and the CAPI sampling fractions f_p per target group, because the problem is nonlinear, allowing the algorithm to stop in a local minimum. The optimal solution is the solution with the lowest coefficient of variation. In 100 hours the algorithm found 11 solutions. The coefficients of variation of the different solutions are between 0.1123 and 0.1217, except for one solution that had a coefficient of variation of 0.21. This solution is not considered further. It cannot be guaranteed that 0.1123 is the overall minimum of the coefficient of variation.

The remaining 10 solutions have almost the same coefficients of variation, but with different sampling fractions for face-to-face observation per target group. Table 1 shows the minimum, maximum and mean CAPI sampling fractions per target group of the 10 solutions. It is assumed that in each target group, 3% of the CAWI sample is not eligible for follow-up. It is striking that in each solution, the target groups 4 and 7 are entirely re-approached face-to-face. Also in target group 8, the CAPI sampling fraction is always relatively high.

Table 2 contains two solutions with approximately the same coefficient of variation, but with different CAPI sampling fractions per target group. The differences are largest for target groups 8, 9 and 3.

Not all solutions use the maximum allowable number of people to be approached face-to-face. For the solution with the least use of CAPI, 7,406 people are approached face-to-face with a coefficient of variation of 0.1123. The solution with the most use of CAPI, 8,039 people are approached face-to-face with a coefficient of variation of 0.1217. This solution was ultimately chosen because the variation coefficients hardly differ from each other, but the use of face-to-face observation is fully utilised.

With the adaptive survey design, the mean response rate decreases compared to a sequential CAWI → CAPI design, in which all CAWI nonrespondents eligible for follow-up are visited at home. However, both the standard deviation and the variation coefficient of the response probabilities are smaller for adaptive survey design. The CAWI part in the response increases: in the new design, approximately 64% of the response will be realised with CAWI, compared to 58% in the current design.

Table 3 shows the results of the chosen solution. The column n CAWI contains the CAWI sample size, the column r CAWI the expected number of CAWI respondents and p

Table 1. Minimum, maximum and mean CAPI sampling fractions per target group.

Stratum	Min	Max	Mean
	%		
1	49	57	53
2	66	88	80
3	66	76	70
4	100	100	100
5	42	47	44
6	50	77	68
7	100	100	100
8	81	100	93
9	63	91	71

Table 2. CAPI sampling fractions per target group for two different solutions.

Stratum	Solution 1	Solution 2
	%	
1	49	57
2	77	83
3	67	76
4	100	100
5	42	47
6	77	75
7	100	100
8	81	100
9	91	72

CAWI shows the expected response rate for CAWI. Column n *elig* shows the number of CAWI nonrespondents eligible for face- to-face follow-up. Columns n CAPI, f CAPI, r CAPI and p CAPI represent the CAPI sample size, the CAPI sampling fraction n CAPI / n elig, the expected number of CAPI respondents and the expected CAPI response rate. The columns r tot and p tot indicate the total number of expected responses and the total response rates per target group. These response rates have been estimated with results of the Health Survey, January-June 2017, with an adjustment to the CAWI response rates due to the raffle of iPads among the sampled people.

Table 4 shows quality measures, in which the situations without and with adaptive survey design are compared. This table shows that the use of adaptive survey design causes the overall response rate to decrease, but the variation of the response rates is improving and the ultimate quality measure $CV(\rho)$ is also improving. This is an indication for less bias due to selective nonresponse.

3.5. Case Prioritisation

Case prioritisation employs the same nine strata and sorts the strata after the CAWI phase by estimated response propensities. One practical complication is added. The Netherlands is divided into ten interviewer regions, each of which contains about one-tenth of the population. Each interviewer region employs 10 to 15 interviewers. Since 2016, CAPI sample numbers per month, survey and interview region have been fixed in advance. The advantage of this is that the required CAPI interview capacity can be planned easier. A disadvantage is that a possible decrease of CAWI response can no longer be compensated by increasing the number of face-to-face re-approaches. In regions where relatively many people belong to target groups that need to be re-approached via CAPI, the fixed number of face-to-face interviews is relatively large.

At the end of the CAWI observation, a sample is drawn from CAWI nonrespondents eligible for follow- up, with agreed sizes per interview region. Prior to this, a priority is defined: in order of the realised CAWI response rate in the sample portion concerned, the target group with the lowest response rate receives the highest priority and the target group with the highest response rate receives the lowest priority. The CAPI potential is then sorted per region by priority, where all elements from one target group are given the same

Table 3. Results of adaptive survey design.

Stratum	n CAWI	r CAWI	p CAWI	n elig	n CAPI	f CAPI	r CAPI	p CAPI	r tot	p tot
			%			%		%		%
1	4,542	1,936	42.6	2,470	1,472	59.6	714	48.5	2,650	58.3
2	785	194	24.8	567	508	89.5	244	48.0	438	55.8
3	7,361	2,765	37.6	4,375	3,480	79.5	1,473	42.3	4,239	57.6
4	727	168	23.1	538	538	100.0	137	25.5	305	41.9
5	1,472	580	39.4	848	410	48.4	265	64.6	845	57.4
6	332	121	36.5	201	149	74.3	72	48.0	193	58.1
7	1,274	245	19.3	991	991	100.0	304	30.7	550	43.1
8	411	83	20.3	315	315	100.0	122	38.6	205	49.9
9	363	105	28.9	247	176	71.4	103	58.3	208	57.3
Total	17,268	6,198	35.9	10,551	8,039	76.2	3,433	42.7	9,631	55.8

Table 4. Quality indicators for adaptive survey design.

Adaptive survey design	$\bar{\rho}$	S_ρ	$CV(\rho)$
		%	
No	63.6	10.1	15.8
Yes	55.8	6.8	12.2

priority. The CAPI potential with the highest priority is selected, then the CAPI potential with the second highest priority is selected and so on until the fixed size per region is reached. If one target group has to be partially selected in order to reach exactly the fixed size, a systematic sample is drawn from this. For that purpose, the CAPI potential with the priority in question of this target group is first sorted by postal code, house number, house letter, addition and designation. The other target groups with the lowest priorities are not observed face-to-face at all in the region concerned.

Because a new prioritisation is made every month, the above approach implies that the CAPI sampling fractions may differ per target group and region on a monthly basis. Fluctuations depend on the CAWI response. It is possible that the fixed CAPI size in one region has already been reached at a certain priority, while in the other region people with a lower priority still have to be selected to reach the fixed CAPI size. The fixed CAPI size per region has been determined in such a way that it is expected that in each region the same target groups will be approached face-to-face.

Statistics Netherlands ultimately decided for a case prioritisation approach based on its practical simplicity and ease in handling interviewer region workloads. A full mathematical optimisation would yield allocation probabilities that, in expectation, lead to the right workloads. However, in practice we are dealing with sample variation across different months, since contact and participation of sample units cannot be controlled. Therefore, per month a subsampling is performed that exactly matches the available workloads.

Table 5 presents the realised allocations, based on six months of the Dutch Health Survey, from January-June 2018. In this period, the realised mean response rate $\bar{\rho}$ equals 57.0%, the standard deviation of response probabilities S_ρ equals 7.6% and the realised coefficient of variation of response probabilities $CV(\rho)$ equals 13.3%. Compared with Table 4, these values are between the corresponding values of the proposed adaptive survey design and the design without adaptation.

Comparison of Tables 3 and 5 shows that most of the realised CAPI sampling fractions differ from the proposed CAPI sampling fractions. This is not only caused by the method of CAPI selection, but also by differences between estimates and realisations of CAWI sample sizes, CAWI response rates and proportions of CAWI nonrespondents eligible for face-to-face follow-up per target group. The largest differences between f CAPI in Tables 3 and 5 can be found in ascending order in target groups 6, 1, 9, 3 and 5. In target groups 7 and 8 f CAPI equals 100% as estimated. Overall, the realisation of f CAPI is 0.9 percent points larger than estimated.

In all target groups, the realised CAWI response rate is higher than expected, except in target group 7. The largest differences between p CAWI in Tables 3 and 5 can be found in ascending order in target groups 6 and 2. The overall CAWI response rate is 2.7 percent points larger than estimated.

Table 5. Realisations of adaptive survey design, Health Survey January–June 2018.

Stratum	n CAWI	r CAWI	p CAWI	n elig	n CAPI	f CAPI	r CAPI	p CAPI	r tot	p tot
			%			%		%		%
1	2,441	1,079	44.2	1,250	335	26.8	157	46.9	1,236	50.6
2	305	101	33.1	200	179	89.5	82	45.8	183	60.0
3	3,725	1,490	40.0	2,187	2,166	99.0	868	40.1	2,358	63.3
4	500	145	29.0	347	328	94.5	138	42.1	283	56.6
5	758	342	45.1	413	140	33.9	88	62.9	430	56.7
6	171	78	45.6	93	31	33.3	16	51.6	94	55.0
7	672	114	17.0	547	547	100.0	173	31.6	287	42.7
8	226	57	25.2	166	166	100.0	74	44.6	131	58.0
9	200	63	31.5	135	128	94.8	66	51.6	129	64.5
Total	8,998	3,469	38.6	5,338	4,020	75.3	1,662	41.3	5,131	57.0

In target groups 4, 8, 6 and 7, the realised CAPI response rate is higher than estimated. In the other target groups less CAPI response is achieved than estimated. The largest differences between p CAPI in Tables 3 and 5 can be found in target groups 4 and 9. The overall CAPI response rate is 1.4 percent points larger than estimated.

In target groups 4, 8, 9, 3 and 2, the total realised response rate is higher than estimated. In the other target groups less response is achieved than estimated. The largest differences between p tot in Tables 3 and 5 can be found in target groups 4, 8 and 1. The overall response rate is 1.2 percent points larger than estimated.

3.6. Method Effects for the Dutch Health Survey

To get an idea of the effect of the adaptive survey design on the results of the Health Survey, simulations were carried out using bootstrapping. To this end, samples were drawn with replacement from the sample of the past year with the correct numbers for CAWI and matching numbers per target group for CAPI. The response data and the survey answers were then linked to these samples. For each sample, the corresponding response was weighted using the weighting model of the Health Survey. Thereafter, estimates were made for the most important target variables and these were compared with the regular estimates.

For the bootstrapping, 1,000 samples with replacement were drawn from the 2016 sample. Each sample had the right CAWI size and the right CAPI size per target group. The numbers of responses may accidentally differ from one sample to another. The sample numbers were taken from the sampling design with adaptive survey design for the Health Survey 2018. After weighting the response per sample, target variables were estimated for both the entire population and subpopulations. These estimates were compared with the results of the Health Survey 2016. One of the assumptions to use CAPI in an adaptive survey design is that the respondents' answers do not depend on the mode in which they respond. This is a strong assumption that is not always true in practice.

The target variable smoking status is known to have mode effects. The proportion of smokers among CAWI respondents is smaller than among CAPI respondents. Thus, if relatively more are observed via CAWI and fewer via CAPI, the number of smokers is expected to decrease. With adaptive survey design, more non-western migrants are approached face-to-face and fewer NL residents or western migrants. Therefore, it is expected that the proportion of smokers among non-western migrants will increase and that the proportion of smokers among NL residents and western migrants will decrease.

The results of the bootstrapping are in line with this, see Figures 3 and 4. Figure 3 shows the smoking status for non-western migrants. The estimate and the corresponding 95% confidence interval of the Health Survey 2016 are shown with the black and dashed lines. The histogram represents the results for this variable in the 1,000 samples of the bootstrapping. Figure 4 shows the smoking status of NL residents and western migrants. For NL residents and western migrants, the proportion of smokers seems to decrease when adaptive survey design is used and for non-western migrants, the proportion of smokers seems to increase compared to the measurement from 2016.

Questions about alcohol, drug use and sexual health are asked in the face-to-face approach via Computer Assisted Self Interviewing. Therefore, fewer mode effects are

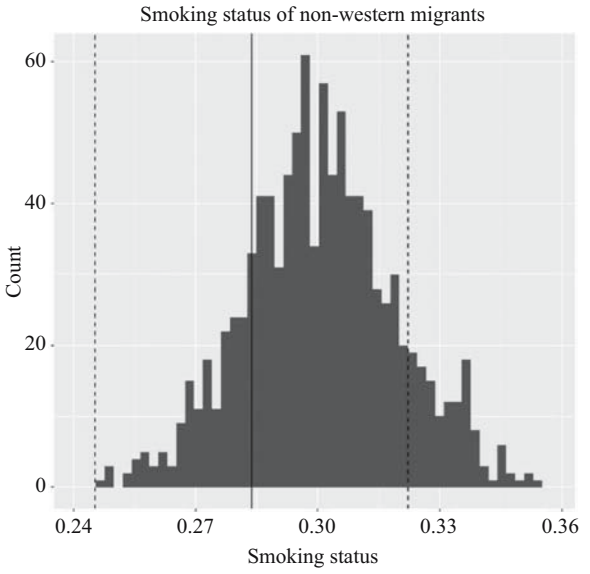

Fig. 3. Smoking status of non-western migrants increases with adaptive design.

expected for these variables. Using the sample data from the bootstrapping, estimates were made for the eleven core variables: contact with general practitioner, contact with dentist, use of non-prescribed medication, experienced health, diabetes, mental health problems, disabilities, informal care, smoking, obesity and drug use, see Table 6. Columns 2016 and SE 2016 show the estimates and standard errors for the core variables of the Health Survey 2016. The last two columns contain the estimates from the bootstrapping samples.

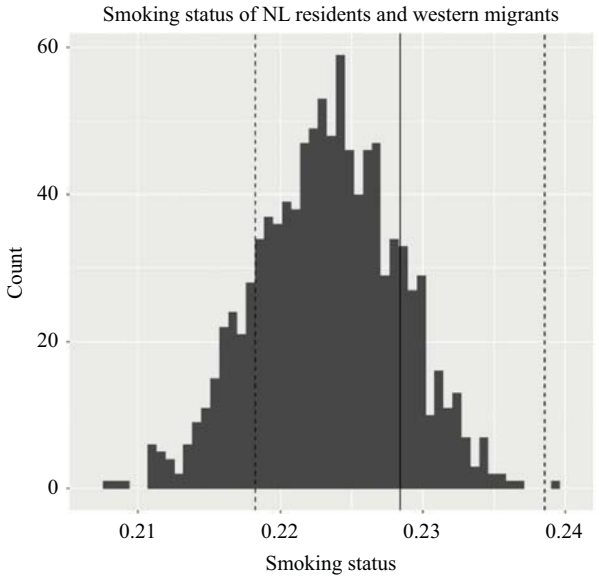

Fig. 4. Smoking status of NL residents and western migrants decreases with adaptive design.

Table 6. Estimates and standard errors for the eleven core variables for the Health Survey 2016, and for the bootstrapping samples.

	2016	SE 2016	Bootstrap	SE Bootstrap
	%			
General practitioner contact	70.9	0.6	71.2	0.8
Dentist contact	79.0	0.5	79.4	0.7
Use of non-prescribed medication	39.7	0.6	39.1	0.8
Experienced health	76.3	0.5	76.1	0.7
Diabetes	5.8	0.3	5.7	0.4
Psychologically unhealthy (MHI-5 score)	11.8	0.4	12.4	0.6
At least 1 OESO-restriction	12.2	0.4	12.0	0.5
Informal care	13.8	0.4	13.8	0.6
Smoking status	23.4	0.5	23.2	0.7
Obesity	13.6	0.4	13.8	0.4

On the basis of the bootstrapping, it is expected that most of the survey results with adaptive survey design do not differ much from those without adaptation. The greatest shifts can be seen in use of non- prescribed medication and psychologically unhealthy. Figure 5 shows the estimates of use of non- prescribed medication. The black and dashed lines show the estimate and 95% confidence interval for the Health Survey 2016. The number of people taking non-prescribed medicines is expected to decrease compared to the 2016 estimate. Figure 6 contains estimates for the variable psychologically unhealthy. It is likely that the introduction of adaptive survey design will increase the number of people who are mentally unhealthy.

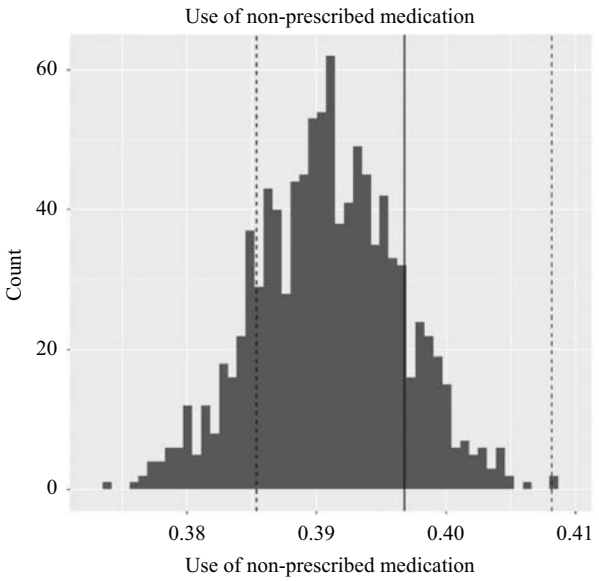

Fig. 5. Proportion of people using non-prescribed medicines decreases with adaptive design.

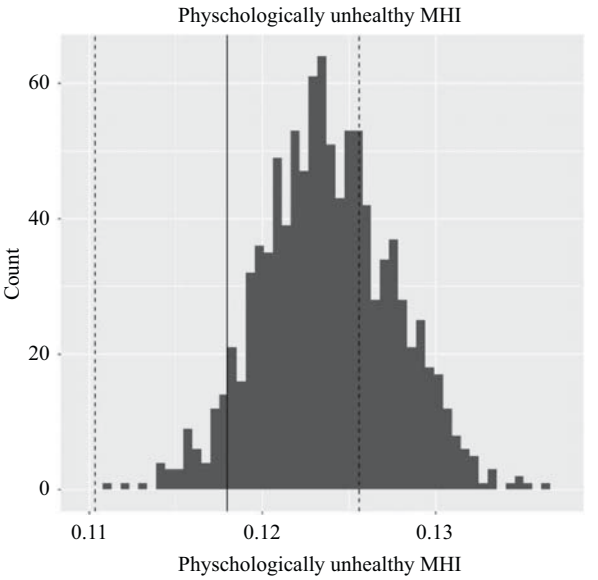

Fig. 6. Proportion of people who are psychologically unhealthy increases with adaptive design.

4. Dussion and Further Activities

This article describes and motivates choices that are made in the implementation of adaptive survey design at Statistics Netherlands. The focus is on sequential mixed-mode designs and the allocation of follow-up interviewer modes to nonrespondents of self-administered modes. The coefficient of variation of response propensities was adopted as the objective in optimisation of the designs. However, a range of logistical and cost constraints have been imposed and lead to a multifaceted optimisation problem. To facilitate easy management of the data collection, a case prioritisation approach was preferred over a mathematical optimisation. A case prioritisation approach is relatively easy to conduct and also is relatively robust to time change in survey design parameters such as costs and response propensities. However, improvement of the balance, that is a smaller coefficient of variation, is not guaranteed. A mathematical optimisation employing expected yield in follow-up interviewer modes does lead to improved balance, but is more sensitive to time change.

For the Health Survey case study, the implemented case prioritisation approach was compared to the mathematical optimisation approach. Results show that, as expected, on average the yield is smaller. However, balance is improved and the population strata that are allocated to face-to-face follow-up closely resemble each other. These results are promising.

There are a few limitations in this study: First, for ease of demonstration, the sampling design was restricted to simple random samples. Second, the role of mode-specific measurement bias was completely ignored. Third, the allocation of interviewer modes is posed as a simple yes-no decision, while it is clearly beneficial to also vary the amount of interviewer effort, for instance the number of contact attempts by the interviewers. Fourth,

since Statistics Netherlands' surveys are mostly repeated monthly surveys, adaptive survey designs are predominantly static, employing little paradata in making decisions.

The limitations lead to various future activities, as many surveys use unequal sampling probabilities and employ both telephone and face-to-face interview modes. Currently, the adaptive survey design framework is extended to stratified sampling designs and to multiple modes. An important decision is the choice of population strata. In this article, the focus was on explanation of nonresponse and strata were based on administrative variables that are used in post-survey adjustments. A general question is to what extent stratification should be survey-specific and to what extent a subset of general strata will always be imposed. This is especially important for regional variables, as they affect interviewer workloads over multiple surveys. Also, research into the stratification itself is conducted. Stratification of the target population can, for example, be performed using K-means clustering. This is a method that divides data into groups based on one or more characteristics, where outliers can be detected. The advantage of this method is that small groups with extremely high or low response rates can be identified as target groups. These target groups can be assigned a separate approach strategy. A disadvantage of the K-means method may be that the target groups are less homogeneous according to the characteristics used.

The desire to work with pre-defined monthly CAPI sample numbers per interview region makes it difficult to choose the right size per target group. Consideration can be given to whether pre-determined CAPI sample numbers for the entire country per month are sufficient for the interviewers' planning. In this case, the necessary sample numbers per target group for CAPI can be selected at random on a monthly basis from the CAWI nonresponse eligible for follow-up.

In this article and application, measurement error is neglected, while it is conjectured that mode-specific measurement biases are present in the Health Survey and also in other official surveys. Literature on the inclusion of mode-specific measurement biases in adaptive survey design is scarce. Two options have been proposed. One option is to estimate stratum mode-specific measurement biases relative to a bench measurement mode and add constraints on the absolute stratum sizes or relative sizes between strata in the optimisation, see Calinescu and Schouten (2015). Another option is to estimate stratum propensities of undesirable answer behaviour and include upper bounds to the prevalence rates of such behaviour in the optimisation, see Calinescu and Schouten (2016). Both options require extra data, either from a re- interview design or from powerful paradata or administrative data. In the application, these estimates were not available. Future work is oriented at efficient estimation of stratum biases or propensities and to include estimates in optimisation of the adaptive survey design.

Finally, the introduction of adaptive survey design requires rethinking and redesigning of the post- survey adjustment and of the estimation of precision of survey estimates. Since adaptive survey design strata are formed by variables that are present in weighting models, some adjustment may no longer be needed or could employ more parsimonious models.

5. Appendix: Characteristics for Segmentation of the Population

1. Wealth of household: 1% groups.

2. Home ownership: owner, rent without rent subsidy, rent with rent subsidy.
3. Income: 1% groups of standardised disposable household income.
4. Socio-economic category: employee of private company, government employee, director or large shareholder, self-employed, employed other, claiming unemployment benefit, claiming income support benefit, claiming other social provision, disabled, pensioner younger than 65 years, pensioner 65 years or older, unemployed other.
5. Household size: number of people in the household.
6. Household status: child living at home, single person, partner without children, partner with children, parent in single parent household, reference person in other household, other household member.
7. Type of household: Single person household, unmarried couple without children, married couple without children, unmarried couple with children, married couple with children, married couple with children, single parent household, other household.
8. Gender: male, female.
9. Marital status: unmarried, married, partnership, divorced, widowed.
10. Age: in years.
11. Age of eldest child: in years.
12. Age of youngest child: in years.
13. Duration of stay in the Netherlands: in years.
14. Part of the country: north, east, south, west.
15. Province: the 12 provinces of the Netherlands.
16. G32: the largest 32 municipalities, other.
17. G4: the largest 4 municipalities, other.
18. Ethnicity: NL residents, western migrants, non-western migrants, unknown.
19. Ethnicity of mother: same.
20. Ethnicity of father: same.
21. Generation: NL residents, first generation migrants, second generation migrants.
22. Highest attained educational level: primary education, secondary general education, secondary vocational education, higher professional education, university.
23. Highest level of education: same.
24. Urbanity of municipality: very strongly urban, strongly urban, moderately urban, few urban and non-urban.
25. Urbanity of neighbourhood: same.

6. References

Alabama R package. Available at: https://CRAN.Rproject=alabama (accessed June 2020).

Bethlehem, J.G. 1988. "Reduction of Nonresponse Bias Through Regression Estimation." *Journal of Official Statistics* 4(3): 251–260. https://www.scb.se/contentassets/-ca21efb41fee47d293bbee5bf7be7fb3/reduction-of-nonresponse-bias-through-regression-estimation.pdf (accessed May 2020).

Calinescu, M. and B. Schouten. 2015. "Adaptive survey designs to minimize mode effects. A case study on the Dutch Labour Force Survey." *Survey Methodology* 41(2): 403–425.

Available at: https://www150.statcan.gc.ca/n1/pub/12-001-x/2015002/article/14250-eng.htm (accessed june 2020).

Calinescu, M. and B. Schouten. 2016. "Adaptive survey designs for nonresponse and measurement error in multi-purpose surveys." *Survey Research Methods* 10(1): 35–47. DOI: https://doi.org/10.18148/srm/2016.v10i1.6157.

Chun, A.Y., S.G. Heeringa, and B. Schouten. 2018. "Responsive and adaptive design for survey optimization." *Journal of Official Statistics* 34(3): 581–597. DOI: https://doi.org/10.2478/jos-2018-0028.

De Heij, V., B. Schouten, and N. Shlomo. 2015. *RISQ m2.1 manual. Tools in SAS and R for the computation of R-indicators and partial R-indicators*. Available at www.risq-project.eu. (accessed june 2020).

Groves, M.R. and S. Heeringa. 2006. "Responsive design for household surveys: tools for actively controlling survey errors and costs." *Journal of the Royal Statistical Society* Series A: Statistics in Society 169 (Part 3): 439–457. DOI: https://doi.org/10.1111/j.1467-985X.2006.00423.x.

Kaputa, S.J., and K.J. Thompson. 2017. "Adaptive design strategies for nonresponse follow-up in economic surveys." *Journal of Official Statistics* 34(2): 445–462. DOI: https://doi.org/10.2478/jos-2018-0020.

Luiten, A. and B. Schouten. 2013. "Adaptive fieldwork design to increase representative household survey response. A pilot study in the Survey of Consumer Satisfaction." *Journal of Royal Statistical Society*, Series A, 176(1): 169–190. DOI: https://doi.org/10.1111/j.1467-985X.2012.01080.x.

Moore, J.C., G.B. Durrant, and P.W.F. Smith. 2018. "Data set representativeness during data collection in three UK social surveys: generalizability and the effects of auxiliary covariate choice." *Journal of the Royal Statistical Society*, Series A, 181(1): 229–248. DOI: https://doi.org/10.1111/rssa.12256.

Peytchev, A., S. Riley, J. Rosen, J. Murphy, and M. Lindblad. 2010. "Reduction of Nonresponse Bias through Case Prioritization." *Survey Research Methods* 4: 21–29. DOI: https://doi.org/10.18148/srm/2010.v4i1.3037.

rpart package. Available at: https://CRAN.Rproject. org/package=rpart. (accessed June 2020).

Schouten, B., M. Calinescu, and A. Luiten. 2013. "Optimizing quality of response through adaptive survey designs." *Survey Methodology* 39(1): 29–58. Available at: https://www150.statcan.gc.ca/n1/pub/12-001-x/2013001/article/11824-eng.htm (accessed June 2020).

Schouten, J.G., F. Cobben, and J. Bethlehem. 2009. "Indicators for the representativeness of survey response." *Survey Methodology* 35(1): 101–113. Available at: https://www150.statcan.gc.ca/n1/pub/12-001-x/2009001/article/10887-eng.pdf (accessed June 2020).

Schouten, B., F. Cobben, P. Lundquist, and J. Wagner. 2016. "Does balancing survey response reduce nonresponse bias?" *Journal of the Royal Statistical Society*, Series A, 179(3): 727–748. DOI: https://doi.org/10.1111/rssa.12152.

Schouten, B., A. Peytchev, and J. Wagner. 2017. *Adaptive Survey Design*. Series on Statistics Handbooks. Chapman and Hall/CRC.

Tourangeau, R., M. Brick, S. Lohr, and J. Li. 2017. "Adaptive and responsive survey designs: a review and assessment." *Journal of the Royal Statistical Society*, Series A, 180(1): 203–223. DOI: https://doi.org/10.1111/rssa.12186.

Wagner, J. 2008. "Adaptive Survey Design to Reduce Nonresponse Bias." PhD diss., University of Michigan, Ann Arbor, USA. Available at: https://deepblue.lib.umich.edu/bitstream/handle/2027.42/60831/jameswag_1.pdf?sequence=1&isAllowed=y (accessed June 2020).

Wagner, J. 2013. "Adaptive contact strategies in telephone and face-to-face surveys." *Survey Research Methods* 7(1): 45–55. DOI: https://doi.org/10.18148/srm/2013.v7i1.5037.

Wagner, J. and F. Hubbard. 2013. "Using propensity models during data collection for responsive designs: Issues with estimation." Paper presented at 68th AAPOR conference, May 16-19, Boston, USA. Available at: http://www.aapor.org/AAPOR_Main/media/AnnualMeetingProceedings/2013/Session_H-7-4-Wagner.pdf (accessed June 2020).

Received September 2018
Revised June 2019
Accepted October 2019

sciendo

Journal of Official Statistics, Vol. 36, No. 3, 2020, pp. 631–645, http://dx.doi.org/10.2478/JOS-2020-0032

The Effects of Nonresponse and Sampling Omissions on Estimates on Various Topics in Federal Surveys: Telephone and IVR Surveys of Address-Based Samples

Floyd J. Fowler[1], *Philip Brenner*[1], *Anthony M. Roman*[1], *and J. Lee Hargraves*[1]

With declining response rates and challenges of using RDD sampling for telephone surveys, collecting data from address-based samples has become more attractive. Two approaches are doing telephone interviews at telephone numbers matched to addresses and asking those at sampled addresses to call into an Interactive Voice Response (IVR) system to answer questions. This study used in-person interviewing to evaluate the effects of nonresponse and problems matching telephone numbers when telephone and IVR were used as the initial modes of data collection. The survey questions were selected from major US federal surveys covering a variety of topics. Both nonresponse and, for telephone, inability to find matches result in important nonresponse error for nearly half the measures across all topics, even after adjustments to fit the known demographic characteristics of the residents. Producing credible estimates requires using supplemental data collection strategies to reduce error from nonresponse.

Key words: Mixed modes; address-based samples.

1. Introduction

The theory behind making estimates from sample surveys is fairly straightforward. Find a list or other way to give most people in your study population a chance to be selected. Draw a probability sample of people in the population, then find a way to get a high percentage of them to answer some survey questions. If that is done, the statistics based on answers that the surveyed sample gives should do a good job of describing the entire population.

Since the 1940s, and perhaps before, the gold standard for how to do this in the United States was to draw an area probability sample of housing units and send interviewers in person to the selected households to do surveys. This is still the approach used for important federal surveys such as the Current Population Survey, the National Crime Victimization Survey, and the National Health Interview Survey. However, such surveys are comparatively expensive. For many years, an acceptable alternative for many purposes was to do telephone surveys based on random-digit dialing. In the 1980s and 1990s, over 90% of housing units in the United States had telephone service, and techniques were

[1] Center for Survey Research, University of Massachusetts Boston, 100 Morrissey Blvd, Boston, MA 02125, U.S.A. Emails: floyd.fowler@umb.edu, philip.brenner@umb.edu, anthony.roman@umb.edu and lee.hargraves@umb.edu

Acknowledgments: This work was funded by the National Science Foundation, Grant No.1424433.

developed to efficiently sample housing units by randomly sampling telephone numbers (Waksberg 1978). Telephone interviewers could obtain response rates that, while usually lower than in-person interviewers, were considered quite respectable, and the resulting data were quite comparable to those from in-person interviews (Groves and Kahn 1979; Groves et al. 1988).

Then, in the last 20 years, two fundamental changes made telephone surveys more problematic. On the sampling side, the growth of cell phones and decline in the use of landlines made sampling people or households via sampling telephone numbers much more complicated. Blumberg and Luke (2016) provide a recent estimate of the distribution of land lines and cell phones in the United States showing that a majority of American households now have only cell phones. In parallel, a variety of factors led people to be much less willing to answer their telephones when they receive calls from "unfamiliar" numbers. As a result, response rates to telephone surveys plummeted (Curtin et al. 2005; Tourangeau and Plewes 2013; Kohut et al. 2012), while concerns about the comprehensiveness of random-digit-dialed based samples grew. These trends, in turn, have led researchers to explore alternative ways to survey general populations (AAPOR Task Force 2017).

One approach is to go back to address-based samples, which give almost everyone living in a housing unit a chance to be selected, and then to experiment with ways other than in-person interviews to collect data. Approaches to data collection can include mailing respondents requests to return mail questionnaires, asking them to go on the internet to complete surveys, or asking them to call an 800 telephone number to provide answers to an automated interviewer (Interactive Voice Response, IVR). Another approach is to try to find a landline or mobile telephone number that matches selected addresses and to conduct interviews at selected households by telephone. All of these approaches can be used in various combinations in mixed-mode surveys (Groves et al. 2009; De Leeuw et al. 2008; Dillman et al. 2014).

The purpose of this study was to learn about the effects of nonresponse on estimates on various topics when data are collected using two of these approaches: by having interviewers call telephone numbers matched to the extent possible with households in an address-based sample and when data are collected by mailing a request to those in an address-based sample to call a telephone number to do an interview with an automated interviewer (IVR). Note that asking households in an addressed-based sample to call in to do interviews is a different methodology from having a computer-based interviewing system calling telephone numbers, asking for an interview. For the telephone survey, error can stem from both an inability to match a telephone number with an address and from an inability to complete interviews by telephone when targeted respondents refuse or simply never answer the telephone. For the IVR mode, error stems from the fact that some people choose not to call the IVR number and do the survey.

Studies of the effect of nonresponse on survey estimates have shown that it is inconsistent (Groves 2006; Groves and Peytcheva 2008; Keeter et al. 2000; Kohut et al. 2012; Keeter et al. 2006). Some estimates from surveys with low response rates are quite biased, while others from the same surveys are similar to estimates from more reliable sources. Thus, one of the important questions for any survey designer is whether or not low response rates are likely to affect estimates in the particular subject area that is the focus of

the survey. These issues are particularly important to those designing US federal surveys who may be considering the value and problems associated with alternative ways of collecting data from household samples.

The primary goal of this study was to assess the extent to which nonresponse and, in the case of telephone surveys, the inability to match telephone numbers to addresses affected estimates using the telephone or IVR to survey addressed-based samples. In addition, we wanted to assess the extent to which any nonresponse effects were or were not related to common topics covered in major government surveys.

2. Methods

2.1. The Sample

Two unclustered probability samples of 1,500 residential addresses each were drawn by Marketing Systems Group (MSG) from an address-based sampling frame covering five Boston area communities. These contiguous communities, including three neighborhoods of Boston (Dorchester, Jamaica Plain, and Mattapan) and two of Boston's immediate suburbs (Milton and Quincy), were chosen based on their demographic diversity. The samples were drawn proportionate to the size of each of the five neighborhoods and cities.

2.2. Data Collection Protocols and Results

One of the samples was devoted to studying nonresponse to efforts to do telephone interviews and the other was devoted to studying nonresponses to IVR invitations. All survey response rates are based on AAPOR definitions (AAPOR 2016).

The sample provider, MSG, attempted to match either a landline or cell telephone number to each selected address in the telephone half of the sample. They were successful for about 60% of the addresses. About 97% of the numbers that were matched were landlines.

Addresses that were matched to a telephone number were sent a letter, accompanied by a 2-USD cash incentive, explaining the background and purposes of the survey, assuring confidentiality, and informing them that an interviewer would be calling in the next few days. Soon thereafter, professional interviewers, working from a central phone facility, called each household. If there was more than one person 18 or older living in the household, the interviewer followed a randomized protocol that chose either the oldest or youngest to be the designated respondent. Each telephone number was called a maximum of 12 times; the median and modal number of calls was six. Calls were placed on various days of the week, primarily during evenings and on weekends. All telephone interviews were conducted in English. In this part of the subsample, 143 respondents (20 percent response rate, AAPOR RR 3) completed the survey.

For the IVR part of the study, a sample of 1,500 addresses received a letter, accompanied by a 2-USD cash incentive, inviting residents to call a toll-free number to complete an IVR survey with an automated interviewer. Invitation letters were printed in English and addressed to the household by name (e.g., "The Smith Household") with "or current resident" added to the addressee in case the matched name was out-of-date or otherwise incorrect. These materials described the topic of the survey as focused on "health and our community" and advised residents that it would take approximately 10–15

minutes to complete. In parallel with the telephone protocols, invitation letters included wording that randomly selected either the youngest or oldest adult 18 years of age or older in the household if there was more than one potentially eligible adult. In this subsample, 148 respondents (10% response rate, RR1) completed the survey.

There were three groups of nonrespondents that we then attempted to interview in person:

1. Those from the telephone sample for whom we had numbers but were unable to interview (whom we refer to as telephone nonrespondents),
2. Those initially selected for the telephone sample for whom we could not find a telephone match and hence could not even attempt a telephone interview, and
3. Those who were invited to complete an IVR survey who did not do so (whom we refer to as IVR nonrespondents).

For each of these groups, the protocols were quite similar.

Approximately half of the addresses for which there was a telephone number (N = 350) but that did not respond to telephone interview efforts were randomly selected for nonresponse follow-up in-person interviews. They were sent a second letter informing them that an interviewer would be visiting their home to complete a personal interview and promising a 20-USD post-paid incentive upon completion of the interview. Each address was visited a maximum of 12 times; the median and modal number of visits was six. In this nonresponding telephone subsample, 128 respondents (43 percent response rate, RR3) completed the survey.

All but a small sample of the addresses that were not matched to a telephone number received a letter, generally similar to the telephone interview letter, informing them that an interviewer would be visiting their homes to complete a personal interview. A 20-USD post-paid cash incentive was promised upon completion of the interview. Addresses were visited a maximum of 13 times; the median and modal number of visits was six. All personal interviews were completed in English. The same within household selection protocol was used as in the telephone interviews. In this subsample, 166 respondents (41 percent response rate, RR3) completed the survey.

The protocol for the 336 addresses selected from the IVR nonrespondents was virtually the same as for the telephone nonrespondents, and the level of interviewer effort was similar as well. For this subsample, it was estimated that a total of 262 were in fact eligible for an interview, and 124 interviews were completed, with a response rate of 47% (RR3).

Data collection began in the fall of 2015 and was completed in April 2016. All the data collection procedures were approved by the university's Institutional Review Board. Table 1 summarizes the samples and these results.

2.3. *The Survey Questions*

All of the questions in the survey were drawn from ongoing US federal surveys: The National Health Interview Survey (NHIS), the Behavior Risk Factor Surveillance Survey (BRFSS), the Health Information National Trends Survey (HINTS), The Current Population Survey (CPS), the National Crime Victimization Survey (NCVS), and the American Community Survey (ACS). The goal was to assess the extent to which a propensity for nonresponse error was related to the topic of the survey questions. In all, there were 35 estimates that we made from the surveys, which we categorized as follows:

Table 1. Data collection results by sample.

	Telephone sample with telephone number match	Telephone sample with no telephone number match	Interactive Voice Response (IVR) sample
Number of initial addresses sampled	921	579	1500
Estimated eligible*	727	NA	1500**
Number of responses by initial mode (Telephone interview or IVR)	143	NA	148
Initial response rates (AAPOR RR3)	20%	NA	10%**
Number assigned for in-person interviews	350	537	336
Estimated eligible*	294	409	262
Number of in-person interviews completed	128	166	124
Response rate for in-person interviews (AAPOR RR3)	43%	41%	47%

*Ineligible addresses included those that were non-residential or that were vacant plus those in which all adult residents were found to be away for an extended period of time or in which no adult could be interviewed in English.

**Address eligibility could not be evaluated, since there was no contact other than a recruitment letter with the addresses during the IVR phase of the survey. As a result, the calculation of the response rate was AAPOR RR1, since there could be no adjustment for eligibility.

- **Health status**: six chronic health conditions plus BMI, self-rated health, and total number of chronic conditions reported
- **Health services received**: doctor visits, dentists' visits, flu shots,
- **Health risks**: health insurance, getting enough sleep, smoking,
- **Giving**: to charity, donating blood, volunteering,
- **Use of technology**: use of phones and computers,
- **Work and income**: Employment status, number hours worked, work loss due to health, on welfare,
- **Household characteristics**: type house, own/rent, number of vehicles owned,
- **Crime**: victim of crime in last 12 months.

The exact wording of the questions, and their sources, are provided in the online Supplemental material.

2.4. Analysis

The sampling methodology used in this study was an application of one originally proposed by Hansen and Hurwitz (1946). An original data collection methodology was performed uniformly for all eligible sample cases. All respondents to that original

methodology were considered to be part of one stratum in which cases were selected with certainty. All nonrespondents (either due to nonresponse or, in the case of the telephone sample, inability to match a telephone number to an address) were then considered part of additional strata. Simple random samples with known probability of selection of these three groups of nonrespondents were then selected with face-to-face interviews being attempted. Weights were constructed for all sample cases that took into account the original probabilities of selection, as well as all subsampling probabilities of selection. Therefore, variable weights were obtained dependent on whether the sample case responded originally or through the follow-up procedures.

First, we looked at how the demographic characteristics of the various samples (respondents and interviewed nonrespondents) compared with Bureau of the Census estimates based on the American Community Survey (ACS) data for the population living in the study area. We also put the data from the test mode and the in-person interviews with nonrespondents together, weighted to adjust for probabilities of selection, to see how the estimates derived from the combined data from the samples targeting telephone and IVR respectively compared with the ACS estimates.

Then, we took two basic approaches to looking at the effects of nonresponse and sample frame limitations on the estimates.

The first analysis (shown in Table 3) addressed three questions about how the substantive survey results compared:

1. How do telephone respondents compare with telephone nonrespondents, those for whom we had telephone numbers but were unable to complete an interview by telephone but whom we were able to interview in person?
2. How do those interviewed who lived at addresses for which there was a telephone number match compare with those interviewed who lived at addresses for which no telephone match was found?
3. How do the IVR respondents compare with the IVR nonrespondents who were interviewed in person?

To make these comparisons, we did *t*-tests or Z-tests on each of the 35 estimates from the survey to see if the estimates were or were not different ($p < .05$). In this case, we did not adjust for demographic differences between the groups because we wanted to see how those who could be surveyed by telephone or by IVR compared with those who could not in the various areas covered in the survey. Demographic differences do not necessarily translate into particular substantive differences. These comparisons also allowed us to separately see the effects of not responding when we had a telephone number and of not being able to match a telephone number to an address on the estimates based on telephone interviews.

The second analysis (shown in Table 4) addressed the question of whether or not estimates based on either the telephone interviews or the IVR responses lay within the 95% confidence intervals around the best estimates we could make using all the data we had collected. The best estimate for the telephone sample combined all the results from the telephone interviews, the in-person interviews with nonrespondents and the in-person interviews in households for which there was not a telephone match. To make these estimates, we, of course, weighted to adjust for the different probabilities of selection for the various strata as described above. After those weights were constructed, a

Table 2. Demographic characteristics by sample type for respondents and for population as a whole.

Demographic characteristics	Telephone respondents	Telephone non-respondents interviewed in person	Households with no telephone match interviewed in person	Combined telephone samples	IVR respondents	IVR non-respondents interviewed in person	Combined IVR samples	ACS estimates for population
Percentage white, non-Hispanic	70%***	56%**	38%	56%***	67%***	50%	52%**	44%
Percentage college graduates	51%***	50%***	40%	45%***	68%***	43%*	46%***	35%
Percentage female	61%*	48%	52%	51%	62%*	56%	57%	53%
Percentage never married	26%***	27%***	47%	36%***	31%***	41%	40%*	46%
Percentage 65 or older	56%***	29%***	14%	26%***	22%**	16%	17%	15%

Note: Statistical significance was calculated by Z-test on differences between estimates from each sample and the estimates based on the American Community Survey (ACS) for the study area: $*p < .05$, $**p < .01$, $***p < .001$

post-stratified adjustment was also performed at the community level using age, race/ethnicity, education, gender and marital status in order to make weighted estimates agree with known demographic profiles for the area under study derived from the American Community Survey data.

We did the same analysis for the IVR respondents. However, in this case we did two comparisons: 1) How the IVR results compared with the estimates based on the combination of responses for the IVR results plus the interviews with the IVR nonrespondents; and 2) how the IVR responses compared with the estimates from the telephone sample, combining data from all three components of the data collection for that sample.

3. Results

Table 2 presents the results of the analysis of the demographic comparisons. The telephone respondents are very different from the population. They are much more likely to be non-Hispanic white, to have graduated from college and particularly to be over age 65; they are much less likely to have never married. The telephone nonrespondents who were interviewed in person tended to differ from the population in the same ways, but the differences were much smaller. In contrast, those interviewed in person at households for which there was not a telephone match look quite similar to the population as a whole; there were no statistically significant differences between those who lacked a telephone match and the ACS estimates. When we put the data from all three sources together, the combined estimates from the telephone sample differ significantly from the ACS estimates on four of the five demographic characteristics: race/ethnicity, education, marital status and age.

Turning to the IVR data, the IVR respondents differ from the ACS estimates in ways that are similar to the telephone respondents in being more likely to be non-Hispanic white, less likely than the population to have been "never married" and they are even more likely than the telephone respondents to be college graduates. In contrast, IVR respondents look more like the population than telephone respondents with respect to age, although they are also significantly more likely to be 65 or older than the ACS estimate.

The IVR nonrespondents who were interviewed in person are more like the population than the IVR respondents with respect to all the demographic characteristics shown in Table 2. When the data from respondents and nonrespondents are combined, however, the estimates are significantly different from the ACS estimates with respect to race/ethnicity and education, but are similar in other respects.

Table 3 summarizes the comparisons between respondents and nonrespondents, and those for whom there was and was not a telephone match across the various survey topics. The bottom line is that phone respondents differed significantly from telephone nonrespondents who were interviewed in person on nine of 35 measures; those with phone matches differed from those without telephone matches on eight measures. There was a little overlap on which measures were affected, but not much. Overall, 13 of the 35 variables had statistically significant differences between the telephone respondents and those interviewed in person from either the telephone nonrespondents or those in households for which there was not a telephone number match, or both.

Table 3. Number of statistically significant differences ($P < .05$) between respondents and nonrespondents by mode and topic, without demographic.

Topic	Telephone respondents versus telephone nonrespondents interviewed in person	Telephone respondents plus telephone nonrespondents interviewed in person versus households with no telephone match interviewed in person	IVR respondents vs IVR nonrespondents interviewed in person	Total number of items
Health conditions	2	1	1	9
Health services received	1	0	2	3
Health risks	0	1	0	4
Giving	1	1	1	3
Use of technology	3	2	2	6
Work and income	2	1	0	6
Household characteristics	0	2	0	3
Crime	0	0	0	1
Total	9	8	6	35

Note: Estimates from the different groups were compared by *t*-tests. Details of estimates and results by individual items are in online Supplemental material Table A1.

IVR respondents differed significantly from nonrespondents who were interviewed in person on only six of the 35 measures.

Table 4 puts all these effects together. It addresses the question of how the best estimates from the telephone interviews or the IVR respondents alone would compare with an estimate that included data from the in-person interviews with nonrespondents. The first column shows the number of estimates that lie outside the 95% confidence interval of the best estimate when the telephone interviews alone are compared with the estimates when data from the telephone interviews are combined with the data from the in-person interviews with the telephone nonrespondents and with those in households for whom there was not a matched telephone number. The second column compares the IVR estimates with the estimates one would make combining the IVR responses with the data from the in-person interviews with IVR nonrespondents. The third column compares the IVR estimates with the estimates from the combined estimates from those we tried to interview by telephone. All the estimates in the table are adjusted to match the characteristics of the population as a whole in the study area based on ACS data for age, education, race/ethnicity, gender and marital status.

Because the data from the original mode are embedded in the combined estimates, a direct test of statistical significance is not advisable. Instead, the criterion in the table for a

Table 4. Number of estimates that lie outside the 95% confidence interval of best estimate by mode, adjusted for demographics to match population characteristics.

Topic	Telephone interviews versus telephone interviews plus in person interviews with telephone nonrespondents and those in households with no telephone match	IVR responses versus IVR responses plus in-person interviews with IVR nonrespondents	IVR responses versus telephone interviews plus in person interviews with telephone nonrespondents and those in households with no telephone match	Total number of items
Health conditions	3	0	4	9
Health services received	1	2	3	3
Health risks	3	2	2	4
Giving	1	1	0	3
Use of technology	5	1	1	6
Work and income	2	0	4	6
Household characteristics	3	2	3	3
Crime	0	0	0	1
Total	18	8	17	35

Note: For all 35 items, two estimates were created: One combined data from the telephone interviews with the in-person interviews with telephone nonrespondents and those at addresses for which there was not a telephone match. The other combined data from IVR responses and the in-person interviews with IVR nonrespondents. When combining data, weights were applied for different probabilities of selection and number of adults in the household. The estimates from the telephone interviews and the IVR responses, as well as the two combined sets of estimates, were all adjusted to match the age, education, race/ethnicity, gender and marital status of the adult population living in area according to the ACS. Then the adjusted estimates from the telephone and IVR respondents were compared with the confidence intervals around the two combined estimates. The counts in the table are the number of items, by type, that fell outside two standard errors around the combined estimates. The details of these analyses are in online Supplemental material Table A2.

"difference" is whether or not the point estimate from the telephone interviews or the IVR responses, when adjusted for demographic differences, lies within the 95% confidence interval of the combined comparison estimate. Since point estimates generally are presented and most likely used in practice, we felt this was an acceptable manner to determine if there were errors in the estimates that should be of concern.

For the telephone interviews, 18 out of the 36 estimates lay outside the confidence interval around the estimate based on the combined data. The IVR estimates only differed from the combined estimates that included data from the in-person interviews with IVR

nonrespondents on eight of the 36 estimates. However, when we compared the IVR estimates to the combined estimates from the telephone sample, (Column 3 of Table 4) 17 of the 36 estimates were outside the confidence interval.

The higher rate at which IVR estimates differ from the telephone combined estimates (17 times) as compared to how they differ from combined IVR estimates with nonrespondents included (eight times) is primarily due to the higher telephone sample sizes and hence smaller 95% confidence intervals around those estimates. It is not because the combined estimates for the telephone samples and the combined IVR samples are fundamentally different. In fact, when we did *t*-tests to compare the two combined estimates, there were only four of 35 estimates that were different at $p < .05$, and two of those (cell phone ownership and landline ownership) were likely related. It should be noted that when conducting 35 tests, all at $p < .05$, on average one would expect about two tests to show up as significant by chance (see online Supplemental material, Table A2). Thus, the two independent samples of the same population produced approximately the same overall estimates even though the methodologies used to collect the data differed.

Finally, when we look at the topics on which differences were observed, it is clear that they occur across all the topics. It would be very difficult from the results in Table 4 to conclude that there is any topic that is immune to the effects of nonresponse. Moreover, when we examined the details of the differences, which are presented in Tables A1 and A2 in the online Supplemental material, the patterns seem to be quite unpredictable.

Compared to our best estimates that included telephone interviews combined with data from in-person interviews, the telephone respondents reported higher BMIs, more diabetes, more work missed due to illness and more uninsured but fewer had depression, had gotten flu shots, reported getting enough sleep recently, or reported smoking. They were more likely to report giving to charity, less likely to be on welfare, more likely to live in a single-family house, less likely to rent and reported having more cars. With respect to technology, they reported more landlines, fewer smart phones, less use of phones for e-mails and more computers at home.

The IVR respondents, compared to our best combined estimates, reported lower BMI, fewer chronic conditions, less lung disease and less depression. They reported more visits to dentists, but fewer visits to doctors. They reported fewer flu shots and less smoking. They were more likely to be students, a lower percentage reported working and those who did worked fewer hours; they also missed fewer work days due to illness. Finally they were more likely in single-family homes, less likely to be renters, had more cars and more cell phones.

4. Discussion

This study found that telephone interviews matched to an address-based sample and IVR interviews conducted by asking sampled individuals to call in to an IVR number are biased both in terms of demographically representing the population of interest and in the accuracy of estimates of characteristics of the population, such as health and health care, use of technology, and employment. Clearly, the estimates based on telephone interviews at households matched to an address-based sample are likely to be problematic. About half the measures we tested were not good. The error was driven about equally by those who did not respond and by the fact that those for whom telephone number matches could not

be found were different from those for whom telephone matches could be found. The differences persisted after matching demographic characteristics to the ACS. The differences cut across all the topics we covered.

The IVR results are not dissimilar. Like the telephone respondents, the IVR respondents overrepresented non-Hispanic whites and college graduates. However, profiles for age and marital status were more similar to the population. The substantive results comparing respondents and nonrespondents appeared to be less problematic than the telephone results, as there were only six significant differences. Of course, a big advantage for IVR results is that they included addresses with and without matched telephone numbers, whereas telephone estimates were missing those for whom there was not a telephone match with the selected address. Moreover, the apparently higher response rate for the telephone of 20%, compared with the 10% rate for IVR, was actually not an advantage. When the fact that only 60% of households had a chance to be interviewed is considered, the effective telephone response rate was only 12%.

We found that only eight estimates from the IVR responses lay outside the confidence interval for the adjusted estimate when data from interviewed nonrespondents were added in. However, when we compared the IVR responses to the same estimates as we used for the telephone estimate, we found that 17 of 35 estimates fell outside the confidence interval–almost the same as the 18 from the telephone interviews.

Thus, from both samples the clear conclusion is that even after adjusting for demographic differences in respondents, close to half the estimates from both the IVR and telephone interviews lay two standard errors or more from our best estimates that reduced the effects of nonresponse. Moreover, the estimates of the number of items "seriously" affected by nonresponse is conservative, as there were additional items that were very close to the edges of the confidence intervals that are not in our counts.

It should be pointed out that our "best" estimates included a good deal of nonresponse; the effective response rates when all data are combined were not much over 50%. Indeed, when we combined data from the primary modes with data from the in-person interviews with nonrespondents, there remained significant differences in demographic characteristics between the survey-based estimates and the ACS estimates. However, even though we did not have external data other than demographic to directly evaluate the survey estimates, it would be difficult to argue that the follow-up interview data collected from nonrespondents and households without telephone numbers, which clearly moved the demographics of the samples closer to the ACS estimates, did not move all the estimates closer to the true values when they raised the percentage responding from about 10% to over 50%.

One could ask if changing modes from the primary mode to in-person interviewing could have affected the comparisons. That seems unlikely. The most likely sources of mode effects are that people would be more likely to give socially desirable answers to an actual interviewer than to an automated interviewer, which would show up in comparing IVR responses to responses to the follow-up interviewers. Of the four significant differences between IVR respondents and in-person interviews with nonrespondents that we thought might include an element of social desirability, three of them had those reporting to an automated interviewer giving fewer socially desirable answers: they reported worse health, fewer dentist visits and less giving to charity. The only answer that went the other way was they reported more flu shots.

Another possible limitation of the study is that we focused on one community area that may or may not produce more broadly representative results. The area was chosen for being diverse with respect to ethnicity, age and education. Nonetheless, we certainly urge others to try to collect data to assess the effects of nonresponse for different data collection modes, different research topics, and different populations.

We think these results contribute to our understanding of nonresponse in several ways. They certainly further strengthen the argument for the importance of using multiple data collection strategies to reduce nonresponse associated with a primary data collection mode. As has been shown before, nonresponse does not affect every survey estimate, but it affects many estimates across an array of subject areas in ways that are very difficult to anticipate (Groves 2006; Groves and Peytcheva 2008; Kohut et al. 2012; Keeter et al. 2006). Moreover, obvious adjustments for demographic anomalies in who responds do not do much to make the estimates better. There may be further adjustments that could have been done that would improve the estimates. However, given the heterogeneity of the effects observed and the lack of a gold standard for estimates other than demographics, we think the potential of adjustments of these data or the data in most surveys to eliminate nonresponse error is limited.

A second contribution is to specifically assess the importance of addressing both survey nonresponse and limitations in the sample frame when doing telephone surveys of addressed-based samples. Although nonresponse seems to have been the most important source of error, the differences between those who did and did not have telephone numbers matched to their addresses contributed important error as well.

Third, although the potential biases in who will actually do a telephone survey and which addresses can be matched to telephone numbers have both been observed, providing information showing that nonresponse to IVR has similar levels and types of nonresponse error is important. On the other hand, these data may also provide some encouragement to try IVR as one of several modes to collect data from addressed-based samples, providing that a nonresponse follow-up is planned.

In short, at a time when response rates are dropping for traditional survey protocols (De Leeuw and De Heer 2002; Tourangeau and Plewes 2013; AAPOR Task Force 2017), there is great pressure to accept low response rates and to use imperfect sample frames in the hope that the estimates will be "good enough". These data are another contribution to the argument that for many survey purposes there is no substitute for comprehensive sample frames and mixed-mode efforts to have a high percentage of a target population represented in survey results. We hope these results will further stimulate research on how best to obtain responses from high percentages of selected survey samples.

5. References

AAPOR. 2016. "Standard Definitions Final Dispositions of Case Codes and Outcome Rates for Surveys (Revised 2016)." American Association for Public Opinion Research. Available at: https://www.aapor.org/AAPOR_Main/media/publications/Standard-Definitions20169theditionfinal.pdf (accessed September 2019).

AAPOR Task Force. 2017. "The Future Of U.S. General Population Telephone Survey Research." 2017. Available at: https://www.aapor.org/Education-Resources/Reports/

The-Future-Of-U-S-General-Population-Telephone-Sur.aspx (accessed September 2019).

American Community Survey (ACS) Available at: https://www.census.gov/programs-surveys/acs/ (accessed June 2020).

Behavioral Risk Factor Surveillance Survey (BRFSS). Available at: https://www.cdc.gov/brfss/index.html (accessed June 2020).

Blumberg, S.J. and J.V. Luke. 2016. "Wireless Substitution: Early Release of Estimates From the National Health Interview Survey, July-December 2016." Available at: https://www.cdc.gov/nchs/data/nhis/earlyrelease/wireless201705.pdf (accessed September 2019).

Current Population Survey (CPS). Available at: https://www.census.gov/programs-surveys/cps.html (accessed June 2020).

Curtin, R., S. Presser, and E. Singer. 2005. "Changes in Telephone Survey Nonresponse over the Past Quarter Century." *Public Opinion Quarterly* 69(1):87–98. DOI: https://doi.org/10.1093/poq/nfi002.

De Leeuw, E.D., D.A. Dillman, and J.J. Hox. 2008. "Mixed-Mode Surveys: When and Why." In *International Handbook of Survey Methodology*, edited by Edith D. de Leeuw, Joop J. Hox, and Don A. Dillman, 299–311. New York: Erlbaum. Also European Association of Methodology. Available at: https://www.eam-online.org.

De Leeuw, E.D. and W. de Heer. 2002. "Trends in Household Survey Nonresponse: A Longitudinal and International Comparison." In *Survey Nonresponse*, edited by Groves Robert M., Don A. Dillman, John L. Eltinge, and Roderick J.A. Little: 41–54. New York: Wiley.

Dillman, D.A., J.D. Smyth, and L.M. Christian. 2014. *Internet, Phone, Mail, and Mixed-Mode Surveys: The Tailored Design Method*. Hoboken, NJ: John Wiley & Sons., Inc.

Groves, R.M., F.J. Fowler, M.P. Couper, J.M. Lepkowski, E. Singer, and R. Tourangeau. 2009. *Survey Methodology*. Hoboken, NJ: John Wiley and Sons.

Groves, R.M. 2006. "Nonresponse Rates and Nonresponse Bias in Household Surveys: What Do We Know about the Linkage between Nonresponse Rates and Nonresponse Bias?" *Public Opinion Quarterly* 70(5):646–675. DOI: https://doi.org/10.1093/poq/nfl033.

Groves, R.M., P.P. Biemer, L.E. Lyberg, J.T. Massey, W.L. Nicholls, and J. Waksberg, eds. 1988. *Telephone Survey Methodology*. New York, NY: John Wiley & Sons.

Groves, R.M. and E. Peytcheva. 2008. "The Impact of Nonresponse Rates on Nonresponse Bias: A Meta-Analysis." *Public Opinion Quarterly* 72(2):167–189. DOI: https://doi.org/10.1093/poq/nfn011.

Groves, R.M. and R.L. Kahn. 1979. "Surveys by Telephone; a National Comparison with Personal Interviews." New York, N.Y.: Academic Press.

Hansen, M.H. and W.N. Hurwitz. 1946. "The Problem of Non-Response in Sample Surveys." *Journal of the American Statistical Association*, 41(236):517–529. DOI: https://doi.org/10.1080./01621459.1946.10501894.

Health Information National Trends Survey (HINTS). Available at: https://hints.cancer.-gov/ (accessed june 2020).

Keeter, S., C. Kennedy, M. Dimock, J. Best, and P. Craighill. 2006. "Gauging the Impact of Growing Nonresponse on Estimates from a National RDD Telephone Survey." *Public Opinion Quarterly* 70(5):759–779. DOI: https://doi.org/10.1093/poq/nfl035.

Keeter, S., C. Miller, A. Kohut, R.M. Groves, and S. Presser. 2000. "Consequences of Reducing Nonresponse in a National Telephone Survey." *Public Opinion Quarterly* 64(2):125–148. DOI: https://doi.org/10.1086/317759.

Kohut, A., S. Keeter, C. Doherty, M. Dimock, and L. Christian. 2012. "Assessing the Representativeness of Public Opinion Surveys." *Pew Research Center, Washington, DC*. Available at: http://www.people-press.org/2012/05/15/assessing-the-representativeness-of-public-opinion-surveys/ (accessed September 2019).

National Health Interview Survey (NHIS). Available at: https://www.cdc.gov/nchs/nhis/index.htm (accessed June 2020).

National Crime Victimization Survey (NCVS). Available at: https://www.census.gov/programs-surveys/ncvs.html (accessed June 2020).

Tourangeau, R. and T.J. Plewes. 2013. *Nonresponse in Social Science Surveys: A Research Agenda – National Research Council, Division of Behavioral and Social Sciences and Education, Committee on National Statistics, Panel on a Research Agenda for the Future of Social Science Data Collection*. Washington, DC: The National Academies Press.

Waksberg, J. 1978. "Sampling Methods for Random Digit Dialing." *Journal of the American Statistical Association* 73(361):40–46. DOI: https://doi.org/10.1080/01621459.1978.10479995.

Received August 2018
Revised May 2019
Accepted September 2019

sciendo

Journal of Official Statistics, Vol. 36, No. 3, 2020, pp. 647–674, http://dx.doi.org/10.2478/JOS-2020-033

Working with Response Probabilities

Jelke Bethlehem[1]

Sample surveys are often affected by nonresponse. These surveys have in common that their outcomes depend at least partly on a human decision whether or not to participate. If it would be completely clear how this decision mechanism works, estimates could be corrected. An often used approach is to introduce the concept of the response probability. Of course, these probabilities are a theoretical concept and therefore unknown. The idea is to estimate them by using the available data. If it is possible to obtain good estimates of the response probabilities, they can be used to improve estimators of population characteristics.

Estimating response probabilities relies heavily on the use of models. An often used model is the logit model. In the article, this model is compared with the simple linear model.

Estimation of response probabilities models requires the individual values of the auxiliary variables to be available for both the respondents and the nonrespondents of the survey. Unfortunately, this is often not the case. This article explores some approaches for estimating response probabilities that have less heavy data requirements. The estimated response probabilities were also used to measure possible deviations from representativity of the survey response. The indicator used is the coefficient of variation (CV) of the response probabilities.

Key words: Nonresponse; adjustment weighting; response propensity; representativity.

1. Introduction

There are various ways of selecting a sample for a survey, but over the years it has become clear that the scientifically sound way to do this is by means of selecting a *probability sample*. Objects (persons, households, businesses) must have a non-zero probability of selection, and all these selection probabilities must be known. This makes it possible to compute precise, unbiased estimates of population characteristics. Also, the precision of these estimates can be quantified, for example by means of a *confidence interval*, or a *margin of error*. These are the fundamental principles of survey sampling.

In practice, the situation is often not so ideal. All kinds of problems may affect the quality of the estimates. One of those problems is (unit) *nonresponse*. This means that no information is obtained about a number of objects in the sample. The questionnaire form remains empty for these objects. One of the effects of nonresponse is that the sample size is smaller than expected. This leads to less precise, but still valid, estimates of population characteristics. This is not a serious problem as it can be taken care of by increasing the initial sample size. A far more serious effect of nonresponse is that estimates of population characteristics may be *biased*. This occurs if, due to nonresponse, some groups in the population are over- or under-represented, and these groups behave differently with

[1] Leiden University, Institute of Political Science, Albert Verweystraat 21, 2394 TK Hazerswoude-Rijndijk, The Netherlands. Email: jelkeb@xs4all.nl.

respect to the characteristics being investigated. Consequently, wrong conclusions will be drawn from the survey data. Such a situation must be avoided as much as possible. Therefore, the amount of nonresponse must be kept small as much as possible. Nevertheless, in spite of all these efforts, a substantial amount of nonresponse usually remains. There are several books that treat the nonresponse problem in more detail. A general overview is given by Bethlehem et al. (2011). Stoop (2005) shows when nonresponse can cause bias and investigates causes of and reasons of nonresponse. Särndal and Lundström (2005) focus on estimation techniques that improve the accuracy of survey estimates. A more recent reference is Valliant et al. (2018). The goal of this book is to present a set of tools for handling nonresponse They also show how existing software can be used to solve survey problems.

Although probability sampling is the preferred way of sample selection, some researchers use different selection techniques. Particularly for online surveys, *self-selection sampling* is used. The questionnaire is made available on the internet. Respondents are those visitors of the website who spontaneously decide to participate in the survey. No random sampling is applied. Respondents are those who happen to know the survey is being conducted, happen to have internet access, decide to visit the survey website, and complete the questionnaire. As the selection mechanism of these online surveys is completely unknown and unclear, it is impossible to compute precise and valid estimates of population characteristics. For more on self-selection surveys, see, for example Bethlehem and Biffignandi (2012).

Both probability sampling (affected by nonresponse) and self-selection sampling have in common that their outcomes depend, at least partly, on a human decision whether or not to participate. If it would be completely clear how this decision mechanism works, the estimates could be corrected. Unfortunately, such information is not available. An often used approach to analyse the effects of and to correct for biased human participation decisions is to introduce the concept of the *response probability*. It is assumed that every object in the target population of the survey has a certain probability to respond in the survey if asked to do so. Of course, these probabilities are a theoretical concept and therefore unknown. The idea is now to estimate the response probabilities using the available data. If it is possible to obtain good estimates of the response probabilities, they can be used to improve estimators of population characteristics.

Estimated response probabilities can be used by survey researchers in several ways. The focus in this article is on:

- *Analysis of nonresponse*. By analysing the relationships between response probabilities and other survey variables, insight is obtained in how the nonresponse mechanism works.
- *Correction for nonresponse*. Once precise estimates of response probabilities are available, they can be used in weighting adjustment techniques that reduce the bias caused by nonresponse.
- *Representativity indicator*. The response probabilities can be used to construct a representativity indicator, which shows, in one number, how good or bad a survey is.

Estimating response probabilities relies heavily on the use of models. An often used model is the *logit model*. It attempts to predict the response probabilities by using a set of

auxiliary variables. This seems to work well in practical survey situations. Other models are the *probit model* and the *linear model*. These models are compared.

Estimation of these response probabilities models requires the individual values of the auxiliary variables to be available for both the respondents and the nonrespondents in the survey. Unfortunately, this is often not the case. This article explores some approaches for estimating response probabilities that have less heavy data requirements. The idea is to start by computing weights with some weighting adjustment technique. These weights can be seen as a kind of inverted response probabilities, and therefore they can be used to estimate response probabilities. Weighting techniques have more modest data requirements. They can compute weights without having the individual data of the nonrespondents. Two weighting techniques are considered: generalised regression estimation and raking ratio estimation.

By taking the logit model as a benchmark, it is explored whether approximately the same estimated response probabilities can be obtained using techniques requiring less information:

1. The linear model for response probabilities;
2. Transforming weights that have been obtained by generalised regression estimation into estimated response probabilities;
3. Transforming weights that have been obtained by raking ratio estimation into estimated response probabilities.

The various approaches are tested on a real survey data set of Statistics Netherlands. This is an anonymised data set that will be called here the General Population Survey (GPS). The sample for this survey was selected from the population register of the Netherlands. Therefore, auxiliary variables in the register are available for both respondents and nonrespondents. Moreover, the sample data file was linked to some registers, providing even more auxiliary variables. So logit models could be fitted, and they could be compared with approaches requiring less data.

The estimated response probabilities were used to measure possible deviations from representativity of the survey response. The indicator used is the *coefficient of variation* (CV) of the response probabilities. This CV can be seen as a normalised measure of dispersion of the response probabilities. The larger the value of the CV, the more the response probabilities will vary, and the more the survey response therefore will lack representativity. The CV also turns up as a component of the bias of estimators that are affected by nonresponse.

The weighting adjustment approach makes it possible to estimate response probabilities in situations in which the logit model cannot be used. An example is given. Response probabilities and the CV are computed for a self-selection panel. This is the 'EenVandaag Opiniepanel' of the Dutch national public television channel 'NOP1'.

2. The Concept of Response Probability

2.1. Nonresponse in a Simple Random Sample

Let the finite *survey population U* consist of a set of N identifiable objects that are labelled $1, 2, \ldots, N$. Associated with each object k is an unknown value Y_k of the *target variable*.

The vector of all values of the target variable is denoted by

$$Y = (Y_1, Y_2, \ldots, Y_N)'. \tag{1}$$

The symbol ' denotes transposition of a matrix or vector. The objective of the sample survey is assumed to be the estimation of the population mean

$$\bar{Y} = \frac{1}{N}\sum_{k=1}^{N} Y_k. \tag{2}$$

To estimate this population characteristic, a simple random sample of size n is selected without replacement. The sample can be represented by the N-vector

$$a = (a_1, a_2, \ldots, a_N)' \tag{3}$$

of indicators, where $a_k = 1$ if object k is selected in the sample, and otherwise $a_k = 0$.

In case of simple random sampling without replacement, the sample mean

$$\bar{y} = \frac{1}{n}\sum_{k=1}^{N} a_k Y_k \tag{4}$$

is an unbiased estimator of the population mean.

Now suppose there is unit nonresponse in the survey. It is assumed that each object k in the population has a certain, unknown probability ρ_k of response. If object k is selected in the sample, a random mechanism is activated that results with probability ρ_k in response and with probability $1 - \rho_k$ in nonresponse. A vector R of response indicators

$$R = (R_1, R_2, \ldots, R_N)' \tag{5}$$

is introduced, where $R_k = 1$ if the corresponding objects k responds, and where $R_k = 0$ otherwise. So, $P(R_k = 1) = \rho_k$, and $P(R_k = 0) = 1 - \rho_k$.

The survey response only consists of those elements k for which $a_k = 1$ (in the sample) and $R_k = 1$ (responds). Hence, the number of respondents is equal to

$$n_R = \sum_{k=1}^{N} a_k R_k. \tag{6}$$

Likewise, the number of nonrespondents is $n_{NR} = n - n_R$.

The values of the target variable only become available for the n_R responding objects. The mean of these values can be denoted by

$$\bar{y}_R = \frac{1}{n_R}\sum_{k=1}^{N} a_k R_k Y_k. \tag{7}$$

Bethlehem (2009) shows that the expected value of the response mean is approximately equal to

$$E(\bar{y}_R) \approx \frac{1}{N}\sum_{k=1}^{N} \frac{\rho_k}{\bar{\rho}} Y_k, \tag{8}$$

where

$$\bar{\rho} = \frac{1}{N}\sum_{k=1}^{N} \rho_k \tag{9}$$

is the mean of all response probabilities in the population. Expression (8) shows that, generally, the expected value of the response mean is unequal to the population mean to be estimated. Therefore, this estimator is biased. This bias is approximately equal to

$$B(\bar{y}_R) = E(\bar{y}_R) - \bar{Y} \approx \frac{R_{\rho Y} S_\rho S_Y}{\bar{\rho}}, \tag{10}$$

where $R_{\rho Y}$ is the correlation between the response probabilities ρ and a target variable Y of the survey, S_ρ is the standard deviation of the response probabilities ρ, and S_Y is the standard deviation of the target variable Y. From Equation (10) a number of conclusions can be drawn:

- The bias vanishes if there is no relationship between the target variable of the survey and the response behaviour. Then $R_{\rho Y} = 0$. The stronger the relationship between the target variable and response behaviour, the larger the bias will be.
- The bias vanishes if all response probabilities are equal. Then $S_\rho = 0$. Indeed, in this situation the nonresponse is not selective. It just reduces the sample size. The more the values of the response probabilities vary, the larger the bias will be.
- The magnitude of the bias increases as the mean of the response probabilities decreases. The response rate is an unbiased estimator of the mean response probability. Translated in practical terms, this means that low response rates will lead to large biases.

It is clear that analysis of the estimates of the response probabilities provides insight into the possible effects of nonresponse on the possible bias of estimates of population characteristics. Some authors (for example Groves 2006) discuss a possible relationship between the response rate of a survey and the bias of its estimates. They fear increased biases if response rates decline. Equation (10) shows that the magnitude of the bias is determined by more than just the response rate. Just as important are the variation of the response probabilities (S_ρ) and the correlation between the response probabilities and the target variable ($R_{\rho Y}$).

2.2. *Nonresponse in a Self-Selection Sample*

Self-selection means that researchers are not in control of the sample selection process. They just make the survey questionnaire available, and wait and see what happens. A typical example is a web survey where everyone can complete the questionnaire on the internet. Also people outside the target population of the survey can participate. It is sometimes even possible to fill in the questionnaire more than once.

Participation in a self-selection web survey requires that respondents are aware of the existence of the survey. Moreover, they must have access to the internet, they have to visit the website (for example by following up a banner, an e-mail message, or a commercial on radio or TV), and they have to decide to fill in the questionnaire. This means that each

object k in the population has unknown probability ρ_k of participating in the survey, for $k = 1, 2, \ldots, N$.

Assuming there are no under-coverage problems, everyone has a nonzero probability of participating in the survey. The survey response is denoted by the vector of indicators

$$R = (R_1, R_2, \ldots, R_N)', \tag{11}$$

where $R_k = 1$ if object k participates, and otherwise $R_k = 1$, for $k = 1, 2, \ldots, N$. The expected value $\rho_k = E(R_k)$ is the *response probability* of element k. The realised sample size is denoted by

$$n_S = \sum_{k=1}^{N} R_k \tag{12}$$

Lacking any knowledge about the values of the response probabilities, a naïve researcher would implicitly assume all these probabilities to be equal. In other words: simple random sampling is assumed. Consequently, the sample mean

$$\bar{y}_S = \frac{1}{n_S} \sum_{k=1}^{N} R_k Y_k \tag{13}$$

is used as an estimator for the population mean. Bethlehem (2009) shows that the expected value of this estimator is approximately equal to

$$E(\bar{y}_S) \approx \frac{1}{N\bar{\rho}} \sum_{k=1}^{N} \rho_k Y_k \tag{14}$$

where $\bar{\rho}$ is the mean of all response probabilities.

It is clear from Equation (14) that, generally, the expected value of this sample mean is not equal to the population mean. One situation in which the bias vanishes is that in which all response probabilities are equal. In terms of the theory of missing data, this comes down to Missing Completely At Random (MCAR). This is the situation in which the cause of missing data is completely independent of all variables measured in the survey. For more information on MCAR and other missing data mechanisms, see Little and Rubin (2002). Indeed, in the case of MCAR, self-selection does not lead to an unrepresentative sample because all elements have the same selection probability.

Bethlehem (2009) shows that the bias of the sample mean in Equation (13) is approximately equal to

$$B(\bar{y}_S) = E(\bar{y}_S) - \bar{Y} \approx \frac{R_{\rho Y} S_\rho S_Y}{\bar{\rho}}, \tag{15}$$

in which $R_{\rho Y}$ is the correlation between the values of target variable Y and the response probabilities ρ, S_ρ is the standard deviation of the response probabilities, S_Y is the standard deviation of the target variable, and $\bar{\rho}$ is the average response probability.

Equation (10) for the bias in a random sample affected by nonresponse is identical to Equation (15) for the bias in a self-selection survey. However, in practical situations their values will be substantially different. For example, the probability samples for surveys of Statistics Netherlands had response rates of around 60%. This means that the average

response probability was 0.6. There have been self-selection web surveys in the Netherlands with large samples. An example is *21minuten.nl*. Approximately 170,000 people completed the questionnaire in 2006. Assuming the target population to consist of all Dutch citizens from the age of 18, the average response probability was 170,000 / 12,800,000 = 0.0133. This is a much lower value than the 0.6 of probability sampling based surveys. So there is a risk of a much large bias in self-selection surveys.

From Equations (10) or (15) an upper bound for the bias can be computed. Given the mean response probability $\bar{\rho}$, there is a maximum value that the standard deviation S_ρ of the response probabilities cannot exceed:

$$S_\rho \leq \sqrt{\bar{\rho}(1-\bar{\rho})}. \tag{16}$$

This implies that in the worst case, S_ρ assumes its maximum value if the correlation coefficient $R_{\rho Y}$ is equal to either $+1$ or -1. Then the absolute value of the bias will be

$$|B_{max}| = S_Y \sqrt{\frac{1}{\bar{\rho}} - 1}. \tag{17}$$

In case of a survey based on probability sampling with a response rate of around 60%, the maximum absolute bias is equal to $0.816 \times S_Y$. In case of a self-selection survey a size 170,000 from a population of size 12,800,000, the maximum absolute bias is $8.619 \times S_Y$. This is more than ten times as large.

3. Estimating Response Probabilities

3.1. Models for Response Probabilities

Response probabilities are unknown. Therefore they must be estimated using the available data. To this end, the concept of the *response propensity* is introduced. Following Little (1986) and Bethlehem et al. (2011), the response propensity of object k is defined by

$$\rho_k(X_k) = P(R_k = 1|X_k) \tag{18}$$

where R_k is the response indicator, and $X_k = (X_{k1}, X_{k2}, \ldots, X_{kp})$' is a vector of values of p auxiliary variables. So the response propensity is the probability of response given the values of some auxiliary variables. The response propensities are also unknown, but they can be estimated provided the values of the auxiliary variables are available for both the respondents and nonrespondents. The estimated response propensity is denoted by $\hat{\rho}_k(X_k)$. If the set of auxiliary variables is sufficient to explain the response probabilities, the (estimated) response propensities will resemble the response probabilities.

To be able to estimate the response propensities, a model must be chosen. The most frequently used one is the *logistic regression model*. It assumes the relationship between response propensity and auxiliary variables can be written as

$$logit\big(\rho_k(X_k)\big) = log\left(\frac{\rho_k(X_k)}{1-\rho_k(X_k)}\right) = \sum_{j=1}^{p} X_{kj}\beta_j, \tag{19}$$

where $\beta = (\beta_1, \beta_2, \ldots, \beta_p)'$ is a vector of p regression coefficients. The *logit* transformation ensures that estimated response propensities are always in the interval [0, 1].

Another model sometimes used is the *probit model*. It assumes the relationship between the response propensity and auxiliary variables can be written as

$$probit\big(\rho_k(X_k)\big) = \Phi^{-1}\big(\rho_k(X_k)\big) = \sum_{j=1}^{p} X_{kj}\beta_j, \tag{20}$$

in which Φ^{-1} is the inverse of the standard normal distribution function. Both models are special cases of the *generalised linear model* (GLM)

$$g\big(\rho_k(X_k)\big) = \sum_{j=1}^{p} X_{kj}\beta_j, \tag{21}$$

where g is called the *link function* that has to be specified. Another special case of the link function is the *identity* link function. This means the relationship between the response propensity and the auxiliary variables can be written as

$$\rho_k(X_k) = \sum_{j=1}^{p} X_{kj}\beta_j. \tag{22}$$

This is a simple *linear model*. It has advantages and disadvantages. A first advantage of the linear model is that coefficients are much easier to interpret. They simply represent the effects of the auxiliary variables on the response propensity. These effects are 'pure' effects. The coefficient of an auxiliary variable is corrected for the interdependencies of the other auxiliary variables in the model. Interpretation of a logit or probit model is not so straightforward. The logit or probit transformation complicates the interpretation of the model parameters.

A second advantage of the linear link function is that the computations are simpler. Estimates of the coefficients can be obtained by ordinary least squares. Estimation of the logit and probit models requires maximum likelihood estimation.

An advantage of the probit and logit models is that estimated response propensities are always in the interval [0, 1]. The linear model does not prevent estimated probabilities to be negative or larger than 1. However, according to Keller et al. (1984) the probability of estimates outside the interval [0, 1] vanishes asymptotically if the model is correct and all response probabilities are strictly positive. If a linear model produces estimated response propensities outside [0, 1], this is often an indication that the model does not fit very well.

It should be noted that the linear model is not necessarily a worse approximation of reality than the probit or logit model. Particularly the logit transform was introduced for convenience only, and not because this model was 'more likely'.

Figure 1 contains the graphs of the logit and probit function. It can be observed that both functions are more or less linear for values of p between, say, 0.2 and 0.8. So, the linear link function can be seen as an approximation of the other two link functions.

3.2. *Application of the Logit and Linear Model*

As an example, the logit model and linear model are applied in the General Population Survey (GPS). The GPS was a face-to-face survey. The target population consisted of

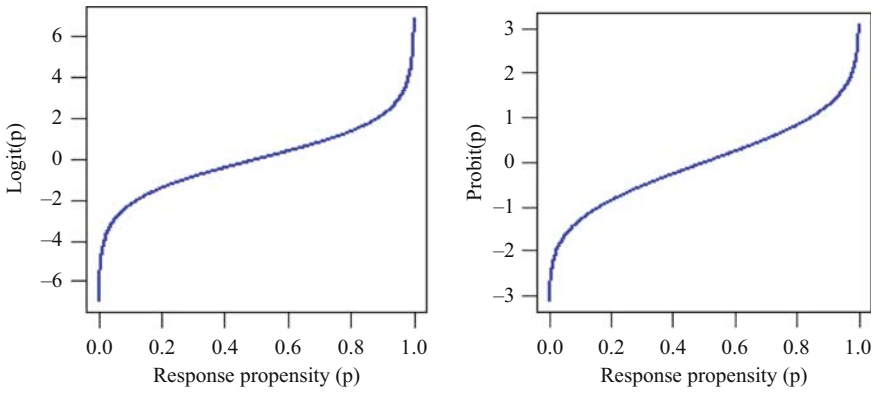

Fig. 1. The logit and probit link functions.

persons of age 12 and older. Persons were selected by means of a stratified two-stage sample. All persons had the same selection probability. The initial sample size was 32,019 people. The response consisted of 18,792 people. So, the response rate was 58.7%.

The GPS sample was linked to the Social Statistics Database (SSD) of Statistics Netherlands. The SSD contains a large set of variables for every person living in the Netherlands. These variables have been retrieved from registers and other administrative sources. By linking the GPS to the SSD, the values of all these variables became available for both respondents and nonrespondents.Table 1 lists the variables that have been used in this article.

Not all auxiliary variables were included in the models for the response propensities. A simple selection procedure was used to determine only the relevant ones. These are variables having a relationship with response behaviour. The strength of this relationship was measured with Cramér's V. It is defined by

$$V = \sqrt{\frac{\chi^2}{n \times min\,(r-1, c-1)}}. \tag{23}$$

χ^2 is the chi-square statistic for the contingency table obtained by crossing two categorical variables, n is the total number of observations in the table, r is the number of rows and c is

Table 1. Auxiliary variables that were available for the GPS.

Variable	Description	Categories
Gender	Gender	2
Married	Is married (yes / no)	2
Age13	Age in 13 age groups	13
Ethnic	Ethnic background	5
HHType	Type of household	5
Phone	Has listed phone number (yes / no)	2
HasJob	Has a job (yes / no)	2
HouseVal	Average house value in neighbourhood	5
Region	Region of the country	5
Urban	Degree of urbanisation	5

Table 2. Cramér's V for the auxiliary variables.

Variable	Cramér's *V*	In model
Region	0.163	Yes
Urban	0.153	Yes
Phone	0.150	Yes
HouseVal	0.112	Yes
Ethnic	0.112	Yes
HHType	0.106	Yes
Married	0.096	Yes
Age13	0.061	No
HasJob	0.037	No
Gender	0.011	No

the number of columns. V always assumes a value in the interval [0, 1]. $V = 1$ means a perfect relationship, and $V = 0$ means no relationship at all. Here, one of the variables is the auxiliary variable, and the other variable is the response variable (with categories *Yes* and *No*). Table 2 contains the values of Cramér's V for all available auxiliary variables.

It is clear from the table that response behaviour has the strongest relationship with the region in which people live (variable *Region*). The second variable is degree of urbanisation (variable *Urban*), which could partly measure the same aspect as region: people in rural areas are more likely to respond than people in urban areas. The relatively high value for the variable *Phone* (has a listed phone number) implies that people with a listed phone number are more likely to respond than those without it.

It was decided (rather arbitrarily) to include the seven auxiliary variables in the logit model for which Cramér's V has value larger than 0.090. Hence, the model contained the first seven variables in Table 2. It should be noted that this selection technique is only a simple one. There are more advanced techniques, like stepwise inclusion techniques that only add variables having a significant contribution, see, for example Bethlehem et al. (2011, chap. 9).

Note that all auxiliary variables are categorical. To be able to include them in the model, each variable is split into a set of dummy variables. There are as many dummy variables as the variable has categories. So there is a dummy variable for each category of each explanatory variable. Unique identification of this model requires some restrictions to be imposed. This can be done in various ways. Here, the coefficient of one of the dummy variables is set to 0. All other coefficients in the set represent deviations from this fixed value.

The logistic regression model was fitted, and subsequently used to estimate the response propensities. Figure 2 shows the distribution of the estimated response propensities. There is a substantial variation. The probabilities vary between 0.128 and 0.732.

Estimated response propensities can be used for the analysis of the nonresponse. Such a (numerical or graphical analysis) can give insight into possible relationships between response behaviour and auxiliary variables. Figure 3 shows an example. It is a boxplot of the response propensities by degree of urbanisation. There is a clear pattern: the lower the degree of urbanisation, the higher the response rate. Response is low in urban areas and high in rural areas.

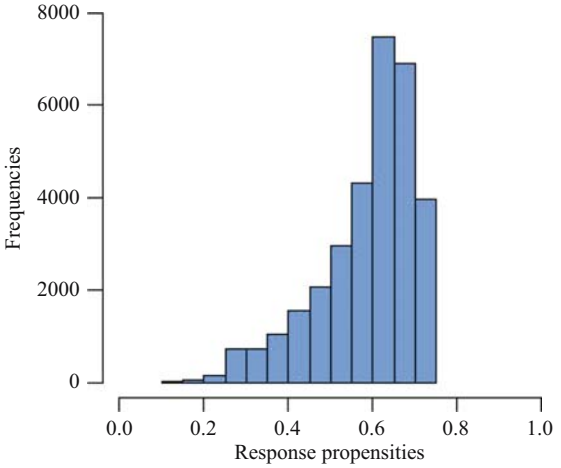

Fig. 2. Histogram of the estimated response propensities (logit model).

The response probabilities were also estimated with a linear model. The same auxiliary variables were included as for the logit model. Again, each explanatory variable was split into dummy variables, and extra restrictions were imposed to allow for unique identification: the last coefficient for each set of dummy variables was set to zero. For example, the variable *Phone* (has listed phone number) had two categories: *No* and *Yes*. The coefficient of *Yes* was set to 0. The estimate of the coefficient for *No* turned out to be equal to -0.108. So, not having a listed phone number reduced the response propensity by 0.108.

The estimated response propensities for the linear model varied between 0.050 and 0.738. So, all estimates were within the interval [0, 1]. Note that the smallest response probability for the linear model (0.050) is somewhat smaller than the one for the logit model (0.128). To see how much the estimated response propensities differ for the two models, they were plotted in a scatter plot, see Figure 4.

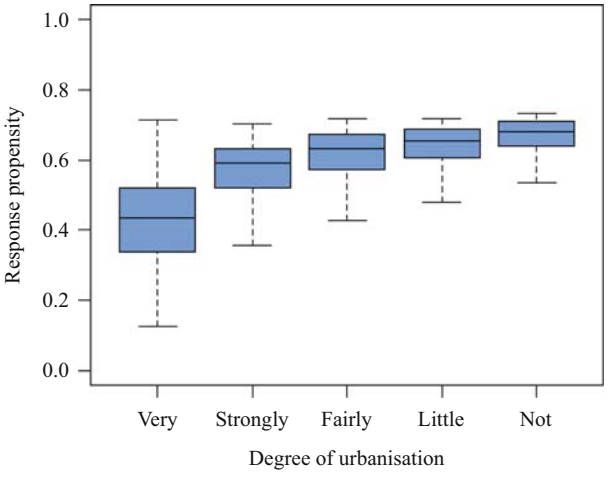

Fig. 3. Boxplot of estimated response propensities by degree of urbanisation for the General Population Survey (GPS).

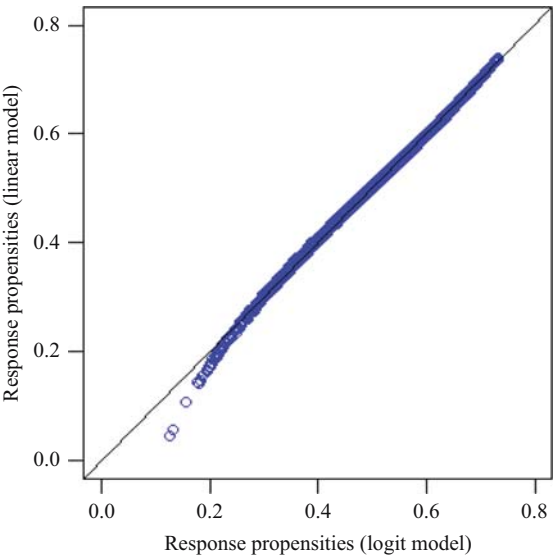

Fig. 4. Response propensities of the logit model and the linear model.

There is an almost perfect linear relationship between the response propensities of both models. This is confirmed by the value of the correlation coefficient, which is equal to 0.999573. Hence, one can conclude that, at least in this example, both models result in almost the same response propensities.

4. Adjustment Weighting With Probabilities

4.1. Weighting Adjustment

Selective nonresponse may cause estimators to be biased. To correct for such a bias, usually some *weighting adjustment technique* is applied. The basic idea is to assign weights to responding elements in such a way that over-represented groups get a weight smaller than one and under-represented groups get a weight larger than one.

There are several types of weighting techniques. The most frequently used ones are *post-stratification*, *generalised regression estimation*, and *raking ratio estimation*. Weighting is based on the use of *auxiliary information*. Auxiliary information is defined here as a set of variables that have been measured in the survey (auxiliary variables), and for which the distribution in the population, or in the complete sample, is available.

It is also possible to use response propensities for weighting adjustment. This can be done in several ways. Two approaches are described in this chapter. The first approach is *response propensity weighting*. It is based on the principle of Horvitz and Thompson (1952) that always an unbiased estimator can be constructed if the selection probabilities are known. In case of nonresponse, selection depends on both the sample selection mechanism and the response mechanism. The idea is now to adapt the Horvitz-Thompson estimator by including the (estimated) response probabilities.

A second approach is *response propensity stratification.* It is based on the fact that estimates will not be biased if all response probabilities are equal. In this case, selection problems will only lead to fewer observations, but the composition of the sample is not affected. The idea is to divide the sample in strata in such a way that all elements within a stratum have (approximately) the same response probability. Consequently, unbiased estimates can be computed within strata. Next, stratum estimates are combined into a unbiased population estimate.

First, the three traditional weight adjustment techniques (post-stratification, generalised regression estimation, and raking ratio estimation) are described. Then it is shown how response propensities can be used for weighting. The two mentioned approaches (response propensity weighting and response propensity stratification) are described. This section concludes with an example of the application of these approaches.

4.2. Post-Stratification

Post-stratification is a well-known and often used weighting technique, see, for example Cochran (1977) or Bethlehem (2002). To carry out post-stratification, categorical variables are needed. By crossing these variables, population and sample are divided into a number of non-overlapping subpopulations, called *strata*.

All objects in one stratum are assigned the same weight, and this weight is equal to the population proportion in that stratum divided by the response proportion in that stratum. Suppose that crossing the stratification variables produces L strata. The number of population objects in stratum h is denoted by N_h, for $h = 1, 2, \ldots, L$. Hence, the population size is equal to $N = N_1 + N_2 + \ldots + N_L$. The weight w_k for an object k in stratum h is now defined by

$$w_k = \frac{N_h/N}{m_h/m}, \tag{24}$$

where m_h is the number of respondents in stratum h (with $m_h < n_h$), and n is the total number of respondents (with $m < n$). If the values of the weights are taken into account, the result is the post-stratification estimator

$$\bar{y}_{ps} = \frac{1}{N}\sum_{h=1}^{L} N_h \bar{y}_h \tag{25}$$

where $\bar{y}_h$ is the response mean in stratum h. So, the post-stratification estimator is equal to a weighted sum of response means in the strata. The bias of the post-stratification estimator is equal to

$$B(\bar{y}_{ps}) = \frac{1}{N}\sum_{h=1}^{L} N_h B(\bar{y}_h) = \frac{1}{N}\sum_{h=1}^{L} N_h \frac{R_{\rho Y,h} S_{\rho,h} S_{Y,h}}{\bar{\rho}_h} \tag{26}$$

where $R_{\rho Y,h}$ is the correlation coefficient between the response probability and the target variable in stratum h, $S_{\rho,h}$ is the standard deviation of the response probabilities in stratum h, $S_{Y,h}$ is the standard deviation of the target variable in stratum h, and $\bar{\rho}_h$is the average response probability in stratum h.

It can be concluded that the bias of weighted estimates is small if there is a strong relationship between the target variable and the stratification variables. The variation in the values of the target variable should manifest itself between strata, but not within strata. In other words, strata should be homogeneous with respect to the target variables. In nonresponse correction terminology, this situation comes down to Missing At Random (MAR).

The bias of the estimator will also be small if the variation of the response probabilities is small within strata. This implies that there must be strong relationships between the auxiliary variables and the response probability.

In conclusion, application of post-stratification will successfully reduce the bias of the estimator if proper auxiliary variables can be found. Such variables should satisfy the following conditions:

- They must be measured in the survey;
- Their population (or complete sample) distribution must be available;
- They must be strongly correlated with all target variables;
- They must be strongly correlated with the response behaviour.

Unfortunately, such variables are often not available. If weakly correlated variables are used, the bias will only be partly removed.

4.3. *Generalised Regression Estimation*

Post-stratification is a simple and straightforward weighting technique. Unfortunately, it is not always possible to apply post-stratification. For example, if there are many auxiliary variables, cross-classifying them may result in empty strata. It is not possible to compute weights for such strata. These problems can be avoided by applying more advanced weighting adjustment techniques. Such techniques are described in Bethlehem (2002) and Särndal and Lundström (2005). One of these techniques is *generalised regression estimation*. It is sometimes also called *linear weighting*.

Generalised regression estimation assumes there is a set of auxiliary variables X_1, X_2, . . . X_p that can be used to predict the values of a target variable Y. The generalised regression estimator is defined by

$$\bar{y}_{GR} = \bar{y} + (\bar{X} - \bar{x})'b, \tag{27}$$

in which $\bar{y}$ is the sample mean of the target variable. $\bar{X}$ is the vector of population means of the auxiliary variables, and $\bar{x}$ is the vector of response means of these variables. Furthermore, b is the (estimated) vector of regression coefficients. The estimator reduces the bias if the underlying regression model fits the data well.

Post-stratification is a special case of generalised regression estimation. If the stratification is represented by a set of dummy variables, where each dummy variable denotes a specific stratum, Equation (27) reduces to Equation (25).

By rewriting Equation (27), it can be shown that generalised regression estimation is a form of weighting adjustment, see, for example. Bethlehem et al. (2011). The value of a weight for a specific respondent is determined by using the corresponding values of the auxiliary variables.

Generalised regression estimation can be applied in more situations than post-stratification. For example, post-stratification by age class and sex requires the population distribution of the crossing of age class by sex to be known. If just the marginal population distributions of age class and sex separately are known, post-stratification cannot be applied. At most, only one variable can be used. However, generalised regression estimation makes it possible to specify a regression model that contains both marginal distributions. In this way, more information is used, and this will generally lead to better estimates.

Generalised regression estimation has the disadvantage that some correction weights may turn out to be negative. Such weights are not wrong, but simply a consequence of the underlying model. Usually, negative weights indicate that the regression model does not fit the data too well. Some analysis software packages are able to take into account weights, but do not accept weights to be negative. This may be a reason not to apply generalised regression estimation.

It should be noted that generalised regression estimation will only substantially reduce the bias if Missing At Random (MAR) applies to the set of auxiliary variables used. For more about generalised regression estimation, see, for example Bethlehem and Keller (1987).

4.4. Raking Ratio Estimation

Correction weights produced by generalised regression estimation are the sum of a number of weight coefficients. It is also possible to compute correction weights in a different way, namely as the product of a number of weight factors. This weighting technique is usually called *raking ratio estimation, iterative proportional fitting*, *RIM weighting* (RIM stands for Random Iterative Method), or *multiplicative weighting*.

Raking ratio estimation can be applied in the same situations as generalised regression estimation, as long as only categorical auxiliary variables are used. Correction weights are the result of an iterative process, in which a series of post-stratifications is carried out repeatedly. This is shown schematically in Figure 5. After post-stratification 1 is carried out, the survey has become representative with respect to the variables of this post-stratification. Then post-stratification 2 is carried out. This adapts the weights. The survey becomes representative with respect to the variables of this post-stratification, but representativity with respect to the variables of post-stratification 1 is lost. However, the

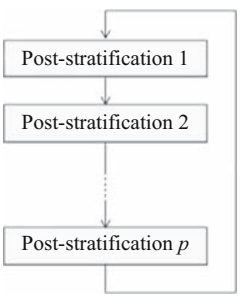

Fig. 5. Raking ratio estimation.

deviation from representativity is smaller than it was. After all (p) post-stratifications have been dealt with, the loop of p post-stratifications starts again. In every post-stratification in every loop the weights are adapted. The process stops if the values of the weights do not change any more. Then the weighting process has converged.

Multiplicative weighting has the advantage that computed weights are always positive. It has the disadvantage that there is no clear model underlying this approach. Moreover, there is no simple and straightforward way to compute standard errors of weighted estimates. Generalised regression estimation is based on a regression model, which allows for computing standard errors.

It should be mentioned that Deville and Särndal (1992) and Deville et al. (1993) have developed a general framework for weighting, of which raking ratio estimation and generalised regression estimation are special cases. They call this framework *calibration*. The paper by Haziza and Beaumont (2017) is also noteworthy. They present an overview of weighting adjustment procedures that are used by national statistical institutes.

4.5. *Response Propensity Weighting*

Horvitz and Thompson (1952) showed that it is always possible to construct an unbiased estimator if the following conditions are satisfied:

- The sample is selected by means of probability sampling;
- Each object in the target population has a positive probability of selection;
- All selection probabilities are known.

Again, let $a = (a_1, a_2, \ldots, a_N)'$ denote the vector of sampling indicators, where $a_k = 1$ if object k is in the sample, and $a_k = 0$ otherwise. The selection probability π_k is defined by $P(a_k = 1)$. It is also called the *first order inclusion probability*. The Horvitz-Thompson estimator is now defined by

$$\bar{y}_{HT} = \frac{1}{N}\sum_{k=1}^{N}\frac{a_k Y_k}{\pi_k} \tag{27}$$

In case of full response, this is an unbiased estimator. In case of nonresponse, however, only the data of the responding objects can be used, and this results in a biased estimator. One way of solving this problem is to include the nonresponse mechanism in the estimator. Let $R = (R_1, R_2, \ldots, R_N)'$ denote the response indicators, and $\rho = (\rho_1, \rho_2, \ldots, \rho_N)$ the corresponding response probabilities, then

$$\bar{y}_{HT} = \frac{1}{N}\sum_{k=1}^{N}\frac{a_k R_k Y_k}{\rho_k \pi_k} \tag{28}$$

would be an unbiased estimator. However, it is not possible to use this estimator, since the values of the response probabilities are not known. The way out is to replace each response probability ρ_k by its corresponding estimated response propensity $\hat{\rho}_k(X_k)$. See Subsection 3.2 on how to estimate response propensities. This results in the adjusted

Horvitz-Thompson estimator

$$\bar{y}_{HT,R} = \frac{1}{N}\sum_{k=1}^{N}\frac{a_k R_k Y_k}{\hat{\rho}_k(X_k)\pi_k} \tag{29}$$

The better the estimated response propensities resemble the 'true' response probabilities, the smaller the bias of the estimator will be.

4.6. Response Propensity Stratification

It was already made clear in Subsection 4.2 that post-stratification can reduce the bias of estimates. The bias of the post-stratification estimator was shown to be equal to

$$B(\bar{y}_{ps}) = \frac{1}{N}\sum_{h=1}^{L} N_h B(\bar{y}_h) = \frac{1}{N}\sum_{h=1}^{L} N_h \frac{R_{\rho Y,h} S_{\rho,h} S_{Y,h}}{\bar{\rho}_h} \tag{30}$$

This bias is small if the strata are homogeneous. This means that the target variable should vary between strata and not within strata. The same applies to the response probabilities: they should vary between strata and not within strata. So a post-stratification based on response probabilities helps to reduce the bias.

The idea is now to use one post-stratification variable and that is the response probability. Since the response probabilities are unknown, the estimated response propensities are used instead.

To construct strata based on estimated response propensities, a number of choices have to be made. One is how to construct the strata? They should at least be such that each stratum contains response propensities of approximately the same size. One way to do it is to divide the interval from 0 to 1, into a number of subintervals of equal length. This may result in some subintervals having many observations, and others only a few. Another way to do it is to make strata that all have the same amount of observations. More on this issue can be found in, for example, Bethlehem et al. (2011, chap. 11). Another choice to be made is for the number of strata to be constructed. According to Cochran (1968), five strata should be sufficient in most cases.

4.7. An Example

The various weighting techniques described in this section are applied in the GPS survey. The GPS was a face-to-face survey. The target population consisted of persons of age 12 and older. Persons were selected by means of a stratified two-stage sample. All persons had the same selection probability. The initial sample size was 32,019 people. The response consisted of 18,792 people. So, the response rate was 58.7%.

The auxiliary variables used were listed phone number (yes/no), married (yes/no), region, degree of urbanisation, ethnic background, house value, and type of household. Two target variables were considered: has a PC (yes/no) and owns a house (yes/no).

Five estimation approaches were applied: no adjustment, generalised regression estimation, raking ratio estimation, response propensity weighting, and response propensity stratification. The resulting estimates are summarised in Table 3.

Table 3. Results of a number of weighting adjustment techniques.

Weighting approach	Has PC	Owns house
No adjustment weighting	57.4 %	62.5 %
Generalised regression estimator	55.7 %	58.5 %
Raking ratio estimation	55.7 %	58.6 %
Response propensity weighting	55.7 %	58.6 %
Response propensity stratification	58.8 %	58.9 %

For the variable *HasPC* all adjustment weighting approaches have approximately the same effect: they produce smaller estimates, and all adjusted estimates are similar. The same can be observed for the variable *Ownhouse*: weighting has an effect, and all adjusted estimates are similar.

At first sight, the results in Table 3 seem to suggest that there are no differences between the various adjustment weighting approaches. One could conclude that the type of weighting adjustment does not matter as long as the right auxiliary variables are used. Of course, this is only one example. It takes more research to establish whether or not this conclusion can be generalised.

5. From Weights to Response Probabilities

5.1. Weighting Adjustment

Estimation of response propensities requires the values of the auxiliary variable to be known for the nonrespondents. This information is not available for many surveys. So then it is not possible to work with estimated response propensities. Still, there is a trick to do this. It makes use of the relation between adjustment weights and response propensities: inverse response propensities can be seen as adjustment weights. The idea is to first carry out some weighting technique and then to transform the weights into response propensities.

There are several types of weighting adjustment techniques. The most frequently used ones are post-stratification, generalised regression estimation and raking ratio estimation. Weighting is based on the use of *auxiliary information*. This is the set of variables that have been measured in the survey, and for which the distribution in the population, or in the complete sample, is available. Note that the individual values of the auxiliary variables are not required for the nonresponding objects. This is in contrast to the techniques discussed in Section 3. It is explored here whether it is possible to estimate the response probabilities using weights that are produced by a weighting model that only uses the marginal distributions of a set of auxiliary variables.

5.2. Estimating the Response Probabilities

It is now shown how weights, computed by means of generalised regression estimation or raking ratio estimation, could be transformed into response propensities.

Let there be p categorical auxiliary variables. The values of these variables for object k are denoted by the vector

$$X_k = (X_k^{(1)}, X_k^{(2)}, \ldots, X_k^{(p)})' \tag{31}$$

The number of categories of variable $X^{(j)}$ is denoted by C_j, for $j = 1, 2, \ldots, p$. The categories are assumed to be numbered $1, 2, \ldots, C_j$.

Whether generalised regression estimation or raking ratio estimation is applied, all responding objects with the same set of values for the auxiliary variables will be assigned the same weight. Suppose an object is in category k_1 of the first variable, category k_2 of the second variable, $\ldots$, and category k_p of the p-th variable. Let $w(k_1, k_2, \ldots, k_p)$ denote the corresponding weight. Furthermore, assume there are $r(k_1, k_2, \ldots, k_p)$ respondents in this group. The number of sample elements $n(k_1, k_2, \ldots, k_p)$ in the group can now be estimated by

$$\hat{n}(k_1, k_2, \ldots, k_p) = \frac{n}{n_R} \times w(k_1, k_2, \ldots, k_p) \times r(k_1, k_2, \ldots, k_p), \tag{32}$$

where n is the sample size and n_R is the total number of respondents. The response propensity for all objects in the group can now be estimated by

$$\hat{\rho}(k_1, k_2, \ldots, k_p) = \frac{r(k_1, k_2, \ldots, k_p)}{\hat{n}(k_1, k_2, \ldots, k_p)} = \frac{n_R}{n} \times \frac{1}{w(k_1, k_2, \ldots, k_p)}. \tag{33}$$

Indeed, the response propensities are inversely proportional to the weights.

5.3. Application to the GPS

The data of the GPS survey are now used to explore the behaviour of response propensities that have been computed from weights. First, the generalised regression estimator is applied. The auxiliary variables are the same as those in the logit model and the linear model of Subsection 3.2. There are seven variables: *Phone*, *Married*, *Region*, *Urban*, *Ethnic*, *HouseVal*, and *HHType*. Only their marginal distributions are used for computing the weights. So there are no interactions in the weighting model.

Note that not the population distributions of the auxiliary variables are used to compute the weights, but the complete sample distributions. The sample distributions are unbiased estimates of the population distributions. So they have some margin of error.

It is assumed that the individual values of the auxiliary variables are only available for the responding elements, and not for the nonresponding elements. So less information is used than in the case of the logit or linear model in Subsection 3.2. As a consequence, response propensities can only be computed for the responding elements.

Figure 6 shows the relationship between the logit model response propensities and the generalised regression estimation response propensities. There is a strong relationship. The correlation coefficient is equal to 0.9801535.

The linear relationship is somewhat less strong than that between the logit model response propensities and the linear model response propensities (with a correlation of 0.999573). Three clusters of points can be distinguished in the scatter plot of Figure 6. Further analysis showed that the two line-shaped clusters with lower response propensities mainly contain people in highly urbanised areas. Persons living in rural areas can all be found in the banana-shaped cluster of high response propensities.

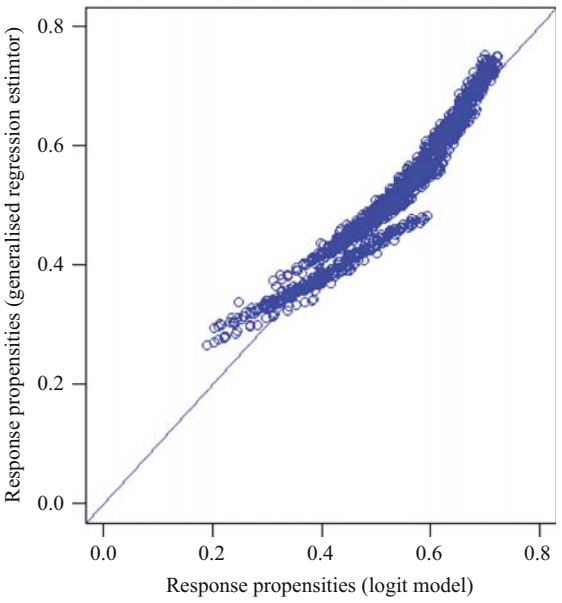

Fig. 6. Comparing response propensities produced by the logit model and generalised regression estimation.

The exercise was repeated using raking ratio estimation instead of the generalised regression estimation. Again, weights were transformed into response propensities. Figure 7 shows the relationship between the logit model response propensities and the raking ratio estimation response propensities. There is a strong relationship. The correlation coefficient is equal to 0.9937689.

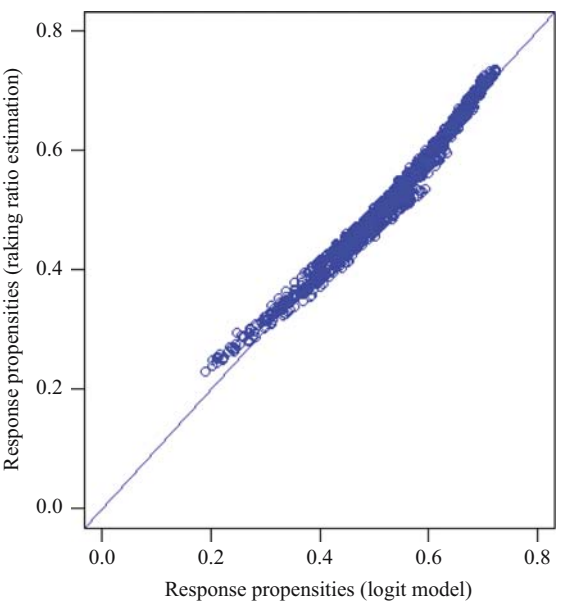

Fig. 7. Comparing response propensities produced by the logit model and raking ratio estimation.

In this example, raking ratio estimation seems to produce response propensities that are closer to those of the logit model than generalised regression estimation. Apparently, the individual values of the auxiliary variables are not needed in this case for estimating response propensities. However, this is just one example. More research is required to make clear whether or not this is a general phenomenon.

6. Using Response Propensities to Assess Representativity

6.1. The Coefficient of Variation

As was already described in Subsection 2.2, the bias of the response mean as an estimator for the population mean is equal to

$$B(\bar{y}_R) = \frac{R_{\rho Y} S_\rho S_Y}{\bar{\rho}}, \tag{34}$$

where $R_{\rho Y}$ is the correlation coefficient between target variable and the response behaviour, S_ρ is the standard deviation of the response probabilities, S_Y is the standard deviation of the target variable, and $\bar{\rho}$ is the average response probability. Equation (34) can be rewritten as

$$B(\bar{y}_R) = R_{\rho Y} \times S_Y \times CV_\rho, \tag{35}$$

where CV_ρ is the *coefficient of variation* (CV) of the response probabilities. It is the only component in the expression for the bias that purely depends on the response probabilities. A large coefficient of variation means that there is a potential risk of a large bias. How large the bias for a specific variable will be, depends on the strength of the relationship between the target variable and the response probabilities.

CV_ρ can be used as an indicator of representativity: the larger the value of CV_ρ, the larger the lack of representativity. A CV_ρ of 0 means that all response propensities are equal, which implies there is no bias.

Note that there are other indicators of representativity. The indicator presented in this section is based on the coefficient of variation of the (estimated) response probabilities. Schouten et al. (2009) propose the R-indicator, which is based on the standard deviation of the (estimated) response probabilities. Related to the concept of representativity is the concept of the imbalance of the response set, which was introduced by Särndal. More about this approach can be found in for example Särndal (2011), Lundquist and Särndal (2013), Särndal and Lundquist (2014a, 2014b, 2017) and Särndal et al. (2016).

6.2. Case 1: Individual Values for the Nonrespondents Are Available

If the individual values of the auxiliary variables are available for both respondents and nonrespondents, the logit model or the linear model, as described in Subsection 3.1, can be applied. For each sample element, the response propensity can be estimated. Therefore, CV_ρ can be computed for the sample, and this is an estimator of the population-based CV_ρ. Note that for small samples, this indicator may be somewhat biased.

The data of the GPS survey are used for an illustration. Response propensities were estimated using the seven auxiliary variables *Phone*, *Married*, *Region*, *Urban*, *Ethnic*,

Table 4. Computation of the CV for the GPS survey (case 1).

Model	Estimated response propensities				CV_ρ
	Minimum	Maximum	Mean	Standard deviation	
Logit	0.128	0.732	0.587	0.112	0.191
Linear	0.050	0.738	0.587	0.112	0.191

HouseVal, and *HHType*. Only main effects were used in the logit and linear model. The computations for both models are summarised in Table 4

Although the linear model produces a somewhat wider range of values for the response propensities, the values of the CV_ρ are approximately the same. At least in this example, the linear model can be used as an approximation of the logit model.

6.3. Case 2: Individual Values for the Nonrespondents are Not Available

If the individual values of the auxiliary variables are not available for the nonrespondents, the weighting approach may be considered for estimating response propensities. Section 5 describes how to do this. This approach requires the population distribution or the complete sample distribution to be known.

It should be noted that the response propensities can only be estimated for respondents and not for nonrespondents. These response propensities cannot be used without correction to estimate the standard deviation of the all response propensities in the sample. The reason is that elements with high response propensities will be over-represented in the response. Fortunately, there is a way out. Let

$$\bar{r}_R = \frac{\sum_{k=1}^{N} a_k R_k \rho_k}{\sum_{k=1}^{N} a_k R_k} \tag{36}$$

denote the response mean of the response probabilities. The expected value of this quantity is approximately equal to

$$E(\bar{r}_R) \approx \bar{\rho}_R \equiv \frac{1}{N\bar{\rho}} \sum_{k=1}^{N} \rho_k^2 \tag{37}$$

By rewriting Equation 37, it can be shown that the standard deviation of the response probabilities is equal to

$$S_\rho = \sqrt{\bar{\rho}(\bar{\rho}_R - \bar{\rho})} \tag{38}$$

In practice, the mean $\bar{\rho}$ of the response probabilities is estimated by the response rate n_R / n. The quantity $\bar{\rho}_R$ is estimated by the mean of the estimated response propensities of the respondents. This assumes simple random sampling. For unequal probability sampling designs the Horvitz-Thompson estimator should be used, which means that values are weighted with the inverse inclusion probabilities. Table 5 summarises the results of the computations for all four approaches considered in this article.

Table 5. Computation of the CV_ρ for the GPS survey.

Approach	Standard deviation	CV_ρ
Logit model	0.097	0.160
Linear model	0.097	0.160
Generalised regression estimation	0.107	0.176
Raking ratio estimation	0.102	0.168

Note that here computations are based on respondents only. This why the values for the logit and linear model differ from those in Table 4.

Although less information is used, raking ratio estimation seems to perform almost as well as the logit and the linear model. Generalised regression estimation performs slightly less than raking ratio estimation, but still produces an estimate that is close to the logit estimates.

Again, it must be remarked that this conclusion is based on application to just one data set. More research is required to find out whether this holds in general.

6.4. Application to a Self-Selection Web Survey

The theory developed for estimating response propensities from adjustment weights can be applied to self-selection surveys. There is no sample selection for such a survey. There are no selection probabilities, but only response probabilities. To say it differently: the whole population is the sample.

Typically, the values of auxiliary variables are only available for the participants, and not for the non-participants. It is assumed that it is possible to obtain the population distributions of the auxiliary variables for weighting purposes. After weights have been computed, they are transformed into response propensities, and they can be used to compute the CV_ρ. The CV_ρ takes the form

$$CV_\rho = \frac{\sqrt{\bar{\rho}(\bar{\rho}_R - \bar{\rho})}}{\bar{\rho}} = \sqrt{\frac{(\bar{\rho}_R - \bar{\rho})}{\bar{\rho}}}. \tag{39}$$

The mean response probability $\bar{\rho}$ is estimated by n_S / N, where n_S is the size of the realised response and N is the size of the target population. The quantity $\bar{\rho}_R$ is estimated by the mean of the estimated response propensities for the respondents.

The theory is applied in an example. There are three nationwide public TV channels in the Netherlands. One of these channels ('NOP1') has a current affairs program called 'EenVandaag'. This program maintains a web panel. It is used to measure public opinion with respect to topics that are discussed in the program. The 'EenVandaag Opinion Panel' started in 2004. In 2008, it contained approximately 45,000 members. The panel is a self-selection panel. Participants were recruited from the viewers of the program. For these reasons, the panel lacks representativity.

In the period before the start of the Olympic Games in Beijing in August of 2008 there was a lot of discussion in the Netherlands about a possible boycott of the games. Suggestions ranged from not showing up at the opening ceremony to athletes not

participating in the games at all. This boycott was proposed because of the lack of respect of the Chinese for the human rights of the Tibetan people. One of the waves of the opinion panel was conducted in April 2008 in order to determine the public opinion of the Dutch with respect to this issue. The members of the panel were invited to complete a questionnaire. This questionnaire also contained topics about other issues, like preference for political parties. The questionnaire was completed by 19,392 members of the panel aged 18 years and older.

The representativity of the response was affected by two phenomena. Firstly, the panel was constructed by means of self-selection. Secondly, not all members of the panel responded to the request to fill in the questionnaire (nonresponse).

If persons apply for membership of the panel, they have to complete a basic questionnaire with a number of demographic questions. These demographic variables can be used as auxiliary variables. The following variables were used for weighting adjustment:

- Gender in two categories: male and female;
- Age in five categories: 18–24, 25–39, 40–54, 55–64, and 65+;
- Marital Status in four categories: never married, married, divorced, widowhood;
- Province of residence in twelve categories: Groningen, Friesland, Drenthe, Overijssel, Flevoland, Gelderland, Noord-Holland, Zuid-Holland, Zeeland, Noord-Brabant and Limburg;
- Ethnic background in three categories: native, first-generation non-native, and second-generation non-native;
- Voting in the 2006 general elections in twelve categories: *CDA* (Christian-democrats), *PvdA* (social-democrats), *SP* (socialists), *VVD* (liberals), *PVV* (rightwing populists), *GroenLinks* (green party), *ChristenUnie* (right-wing Christians), *D66* (liberal-democrats), *PvdD* (party for the animals), *SGP* (right-wing Christians), other party, and did not vote.

The population distributions were available for all these variables. Note that the variables came from different sources, so that only marginal distributions could be used and not cross-classifications of variables.

The first step was to compute adjustment weights. Raking ratio estimation was used for this. The resulting weights turned out to have a large variation. The smallest weight was 0.089 and the largest was 34.570. This large variation clearly points to a substantial lack of representativity.

The next step was to estimate the response propensities using expression (Equation 29). The distribution of these response propensities is shown in Figure 8. It is clear that all response propensities are small. They vary approximately between 0.000 and 0.017. This is not surprising, as only 19,000 people out of a population of more than 12 million people responded.

The computations for the coefficient of variation are summarised in Table 6. The coefficient of variation is a little over one. This means that, compared to the GPS survey, the potential bias of the web survey is more than five times as large.

One should be careful when comparing the CV_ρ of different surveys. Differences are only meaningful if the estimated response probabilities are based on the same model. If

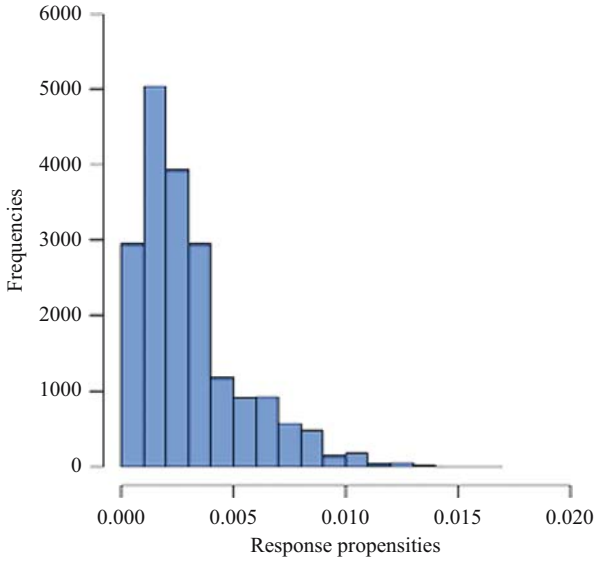

Fig. 8. Histogram of the response propensities in the self-selection survey.

this is not the case, differences may also be attributed to differences in models, and not to differences in the variation of the true response probabilities.

7. Conclusion

Nonresponse can have a serious impact on the quality of survey outcomes. Nonresponse affects the representativity of the survey and therefore the validity of its outcomes. Hence, it is important that survey researchers analyse the outcomes of their surveys. If there is a risk of biased outcomes, some kind of correction is called for.

One way of getting more insight into nonresponse is to introduce the concept of response probability. To that end, a model must be fit that is able to explain response probabilities from a set of auxiliary variables. The most frequently used model is the logistic regression model (or logit model). It is important that all relevant auxiliary variables are included in the model. It must have sufficient explanatory power.

Another model is the linear model. This is a simpler model. It can be seen as an approximation of the logistic regression model. Particularly when response probabilities

Table 6. Computation of the CV for the self-selection web survey.

Quantity	Value
Minimum response propensity (response)	0.000044
Maximum response propensity (response)	0.016878
Mean response propensity (response)	0.003051
Stand. dev. response propensity (response)	0.002334
CV_ρ (response)	0.764931
CV_ρ	1.011636

are within the range from 0.20 to 0.08, both models produce almost the same predictions. Application in the GPS showed that is does not matter which model is used.

To be able to estimate response probabilities with the logistic regression model or the linear model, the values of the explanatory variables must be available for both respondents and nonrespondents. Sometimes this is the case, for example if the sample is selected from a population register or from a sampling frame that is linked to registers. More often the values for the nonrespondents are not available. The article proposes a technique to circumvent this problem: first, a weighting adjustment technique is applied, and which does not require the individual values of the nonrespondents. Examples are generalised regression estimation and raking ratio estimation. These techniques assign weights to respondents. These weights can be seen as reciprocal response probabilities, and therefore response probabilities can be computed.

Application of this reciprocal weights technique to the GPS data showed that it worked in practice. The estimated response probabilities were similar to those obtained by using the logistic model or linear model, even though less information was used (no interactions).

The approach of estimating response probabilities by means of weighting model models has the attractive property that it can also be applied in the case of self-selection surveys. Application of the theory in the example of the web panel shows that the worst case bias can be very large.

Various models for response probabilities were explored in this article and applied to the example of the GPS. The conclusion could be drawn that it does not seem to matter which correction technique is used, as long as the model contains the relevant auxiliary variables. Of course, this conclusion is based on just one example. More research is needed to find out whether or not this conclusion can be generalised.

Response probabilities can be used for various purposes. Not only for corrections, but also for analysis. It was shown how a graphical technique like a boxplot can give more insight into the relation between response behaviour and auxiliary variables.

Another application is to base representativity indicators on response probabilities. A well-known example is the R-indicator. This article proposes an indicator based on the coefficient of variation of the response probabilities. It seems to work for the example of a self-selection web panel. More research is necessary.

8. References

Bethlehem, J.G. 2002. "Weighting Nonresponse Adjustments Based on Auxiliary Information." In *Survey Nonresponse*, edited by Groves, R.M., D.A. Dillman, J.L. Eltinge, and R.J.A. Little, 275–288. New York: John Wiley & Sons.

Bethlehem, J.G. 2009. *Applied Survey Methods, A Statistical Perspective*. Hoboken, NJ: John Wiley & Sons. DOI: https://doi.org/10.1002/9780470494998.

Bethlehem, J.G. and S. Biffignandi. 2012. *Handbook of Web Surveys*. Hoboken, NJ: John Wiley & Sons. DOI: https://doi.org/10.1002/978111812175.

Bethlehem, J.G., F. Cobben, and B. Schouten. 2011. *Handbook of Nonresponse in Household Surveys*. Hoboken, NJ: John Wiley & Sons. DOI: https://doi.org/10.1002/9780470891056.

Bethlehem, J.G. and W.J. Keller. 1987. "Linear Weighting of Sample Survey Data." *Journal of Official Statistics* 3: 141–153. Available at: https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/linear-weighting-of-sample-survey-data.pdf (accessed June 2020).

Cochran, W.G. 1968. "The Effectiveness of Adjustments by Subclassification in Removing in Observational Studies." *Biometrics* 24: 205–2013. DOI: https://doi.org/10.2307/2528036.

Cochran, W.G. 1977. *Sampling techniques*. Third Edition. New York: John Wiley & Sons.

Deville, J.C. and C.E. Särndal. 1992. "Calibration Estimators in Survey Sampling." *Journal of the American Statistical Association* 87: 376–382. DOI: https://doi.org/10.1080/01621459.1992.10475217.

Deville, J.C., C.E. Särndal, and O. Sautory. 1993. "Generalized Raking Procedures in Survey Sampling." *Journal of the American Statistical Association* 88: 1013–1020. DOI: https://doi.org/10.1080/01621459.1993.10476369.

Groves, R.M. 2006. "Nonresponse Rates and Nonresponse Bias in Household Survey." *Public Opinion Quarterly* 70: 646–675. DOI: https://doi.org/10.1093/poq/nfl033.

Hazizi, D. and J.F. Beaumont. 2017. "Constructions of Weights in Surveys." *Statistical Science* 32: 206–226. DOI: https://doi.org/10.1214/16-STS608.

Horvitz, D.G. and D.J. Thompson. 1952. "A Generalization of Sampling Without Replacement from a Finite Universe." *Journal of the American Statistical Association* 47: 663–685. DOI: https://doi.org/10.1080/01621459.1952.10483446.

Keller, W.J., A. Verbeek, and J.G. Bethlehem. 1984. *ANOTA: Analysis of Tables*. Voorburg: Statistics Netherlands, Department for Statistical Methods. (CBS-report 5766-84-M1-3).

Little, R.J.A. 1986. "Survey Nonresponse Adjustment for the Estimates of Means." *International Statistical Review* 54: 139–157. DOI: https://doi.org/10.2307/1403140.

Little, R.J.A. and D.B. Rubin. 2002. *Statistical Analysis with Missing Data*. Second edition. New York: John Wiley & Sons. DOI: https://doi.org/10.1002/9781119013563.

Lundquist, P. and C.E. Särndal. 2013. "Aspects of Responsive Design with Applications to the Swedish Living Conditions Survey." *Journal of Official Statistics* 29: 557–582. DOI: https://doi.org/10.1515/jos-2017-0033.

Särndal, C.-E. 2011. "The 2010 Morris Hansen Lecture: Dealing with Survey Nonresponse in Data Collection, in Estimation." *Journal of Official Statistics* 27: 1–21. Available at: https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7-be7fb3/the-2010-morris-hansen-lecture-dealing-with-survey-nonresponse-in-data-collection-in-estimation.pdf (accessed May 2020).

Särndal, C.E., K. Lumiste, and I. Traat. 2016. "Reducing the Response Imbalance: Is the Accuracy of the survey estimates improved?." *Survey Methodology* 42: 219–238. Available at: https://www150.statcan.gc.ca/n1/pub/12-001-x/2016002/article/14663-eng.htm (accessed June 2020).

Särndal, C.E. and P. Lundquist. 2014a. "Balancing the Response and Adjusting Estimates for Nonresponse Bias: Complementary Activities." *Journal de la Société Française de Statistique* 155: 28–50. Available at: https://www.semanticscholar.org/paper/Balancing-the-response-and-adjusting-estimates-for-S%C3%A4rndal-Lundquist/9686e52f8875ff2b0120303ed2e24b2008f463df (accessed June 2020).

Särndal, C.E. and P. Lundquist. 2014b. "Accuracy in Estimation with Nonresponse: A Function of Degree of Imbalance and Degree of Explanation." *Journal of Survey Statistics and Methodology* 2: 361–387. DOI: https://doi.org/10.1093/jssam/smu014.

Särndal, C.E. and P. Lundquist. 2017. "Inconsistent Regression and Nonresponse Bias: Exploring Their Relationship as a Function of Response Imbalance." *Journal of Official Statistics* 33: 700–734. DOI: https://doi.org/10.1515/jos-2017-0033.

Särndal, C.-E. and S. Lundström. 2005. *Estimation in Surveys with Nonresponse*. Chichester, UK: John Wiley & Sons.

Schouten, B., F. Cobben, and J.G. Bethlehem. 2009. "Measures for the Representativeness of Survey Response." *Survey Methodology* 36: 101–113. Available at: https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2009001/article/10887-eng.pdf?st=uZQq4P0J (accessed June 2020).

Stoop, I.A.L. 2005. *The Hunt for the Last Respondent*. The Hague: Social and Cultural Planning Office. Available at: https://www.researchgate.net/profile/Ineke_Stoop/publication/27686407_The_Hunt_for_the_Last_Respondent_Nonresponse_in_Sample_Surveys/links/569791b908ae1c427905094e/The-Hunt-for-the-Last-Respondent-Non-response-in-Sample-Surveys.pdf (accessed June 2020).

Valliant, R., J.A. Dever, and F. Kreuter. 2018. *Practical Tools for Designing and Weighting Survey Samples*. Heidelberg, Germany: Springer. DOI: https://doi.org/10.1007/978-3-319-93632-1

Received July 2018
Revised April 2019
Accepted September 2019

sciendo

Journal of Official Statistics, Vol. 36, No. 3, 2020, pp. 675–701, http://dx.doi.org/10.2478/JOS-2020-0034

A Validation of R-Indicators as a Measure of the Risk of Bias using Data from a Nonresponse Follow-Up Survey

Caroline Roberts[1], *Caroline Vandenplas*[2], *and Jessica M.E. Herzing*[1]

R-indicators are increasingly used as nonresponse bias indicators. However, their effectiveness depends on the auxiliary data used in their estimation. Because of this, it is not always clear for practitioners what the magnitude of the R-indicator implies for bias in other survey variables, or how adjustment on auxiliary variables will affect it. In this article, we investigate these potential limitations of R-indicators in a case study using data from the Swiss European Social Survey (ESS5), which included a nonresponse follow-up (NRFU) survey. First, we analyse correlations between estimated response propensities based on auxiliary data from the register-based sampling frame, and responses to survey questions also included in the NRFU. We then examine how these relate to bias detected by the NRFU, before and after adjustment, and to predictions of the risk of bias provided by the R-indicator. While the results lend support for the utility of R-indicators as summary statistics of bias risk, they suggest a need for caution in their interpretation. Even where auxiliary variables are correlated with target variables, more bias in the former (resulting in a larger R-indicator) does not automatically imply more bias in the latter, nor does adjustment on the former necessarily reduce bias in the latter.

Key words: Nonresponse; R-indicator; propensity score weighting; nonresponse survey.

1. Introduction

High response rates have traditionally been regarded as a guarantee of survey data quality. Over the past two decades, however, obtaining high response rates in social surveys has become increasingly challenging (De Leeuw and De Heer 2002; Brick and Williams 2013; Kreuter 2013; Williams and Brick 2017; Beullens et al. 2018), and questions have been raised regarding the extent to which they can protect survey estimates from nonresponse bias (Groves 2006; Groves and Peytcheva 2008, Brick and Tourangeau 2017). Indeed, bias depends not only on the rate of nonresponse, but also on the difference in characteristics between respondents and nonrespondents (Groves and Couper 1998), and according to the stochastic view of nonresponse, the covariance between variables influencing the probability of responding and a given survey variable

[1] Institute of Social Sciences, University of Lausanne, Bâtiment Géopolis, Quartier Mouline, CH-1015 Lausanne, Switzerland. Emails: caroline.roberts@unil.ch, and Jessica.herzing@unil.ch
[2] Consulting, Chemin du Cyclotron 6, 1348 Ottignies-Louvain-la-Neuve, Belgium. Email: caroline.vandenplas@b12-consulting.com
Acknowledgments: We are grateful to Michèle Ernst Stähli at FORS, the Swiss Centre of Expertise in the Social Sciences, for providing access to the sampling data analysed in this study. We would also like to thank the anonymous reviewers for their constructive feedback on earlier drafts of our manuscript, which enabled us to make a number of important improvements.

(Bethlehem 2002; Little and Rubin 2014; Brick and Tourangeau 2017). In other words, even in a survey with a high response rate, a variable that correlates highly with the probability of responding may have a larger nonresponse bias than a variable that does not, or a variable that only weakly correlates with the probability of responding in a survey with a lower response rate. Because nonresponse bias is, thus, variable-dependent, finding simple and intuitive methods for detecting its presence and assessing its impact poses an on-going challenge for survey methodologists and statisticians (Groves et al. 2008; Schouten 2018).

In response to this challenge, the last decade has seen the development of a number of new indicators for assessing the risk of nonresponse bias (Wagner 2012; Nishimura, et al. 2016). Of these, one that has rapidly gained popularity is the '*Representativity Indicator*', or *R-indicator* (Schouten et al. 2009), together with its related '*partial R-indicators*' (Schouten et al. 2011; Beullens and Loosveldt 2012). R-indicators offer an intuitive way of summarising the extent to which the respondents in a probability-based sample survey represent all the sample units that were selected from the sampling frame, and the risk of nonresponse bias in survey variables. Because of this, R-indicators have quickly attracted interest as a way to evaluate fieldwork outcomes and compare the effectiveness of different survey designs (e.g. Schouten et al. 2012; Luiten and Schouten 2013; Moore et al. 2018; Schouten and Shlomo 2017). Meanwhile, partial R-indicators are being used to plan adaptive survey designs or identify specific subgroups during fieldwork monitoring for targeted interventions, as in responsive designs (Groves and Heeringa 2006; Schouten et al. 2011a; Beullens and Loosveldt 2012; Schouten et al. 2016).

The utility of R-indicators depends in part, however, on the availability of suitable auxiliary data for their estimation (i.e. variables that correlate both with the probability of responding and key survey variables), which (in cross-sectional surveys at least) is often limited (e.g. Sakshaug and Antoni 2018). Given that R-indicators essentially summarise nonresponse bias in the auxiliary variables, the question is raised as to how effective they (and the auxiliary variables) are at identifying the risk of nonresponse bias on other survey variables – especially in the context of large-scale surveys covering a wide variety of topics. Relatedly, given that the same auxiliary data can also be used to adjust for nonresponse bias, a further question is raised as to what can be learned from the R-indicator about bias in other survey variables *after* adjustment on the auxiliary variables.

In this article, we investigate these potential limits of R-indicators in a case study using data from the Swiss European Social Survey (ESS), which in Round 5 (2010), included a nonresponse follow-up (NRFU) survey. We first evaluate the suitability of available auxiliary data (from the sampling frame based on population registers) for estimating R-indicators by examining how well they correlate with a selection of target variables. Then we assess how well the R-indicator predicts the presence of actual nonresponse bias on these variables, before and after adjustment, using data from the NRFU to estimate the difference between respondents and nonrespondents in the main survey. Before describing our research questions and analytic approach in more detail, we present an overview of R-indicators, their possible limitations, and the role of auxiliary variables, then review recent studies that have investigated their performance and interpretation.

2. Background

2.1. Using R-Indicators to Detect Nonresponse Bias – The Role of Auxiliary Variables

R-indicators describe the variance of the sample members' probability of responding to a given survey (for detailed accounts, including the statistical notation and formulae for estimating R-indicators, see Schouten and Cobben 2007; Schouten et al. 2009; Schouten et al. 2011a; Schouten and Shlomo 2017; De Heij et al. 2015). The higher the variance in the response probabilities, the more likely it is to have an unbalanced respondent sample (i.e., reduced 'representativity') and, theoretically, to have nonresponse bias on other variables that correlate with the response probability. R-indicators are normalised to range between zero and one, where one represents strong representativity, and zero the 'maximum deviation from representativity' (Schouten et al. 2009, 104). As such, they provide an intuitive summary statistic for describing survey quality.

As we do not know the actual probability of responding of all members of the survey sample, in practice, the R-indicator is based on the standard deviation of the *estimated* response probabilities (Schouten and Cobben 2007; Schouten et al. 2009; Schouten et al. 2011), calculated based on auxiliary variables available both for respondents and nonrespondents (e.g. frame variables, linked contextual or administrative data, survey paradata – Cornesse and Bosnjak 2018), typically using a logistic regression model. The higher the standard deviation of the estimated response probabilities, the less representative is the sample across categories of the auxiliary variables. Thus, the notion of 'representativity' relates specifically to the extent to which the response sample represents the complete sample on the auxiliary variables included as covariates in the model (Cornesse and Bosnjak 2018).

The utility of R-indicators lies partly in their ability to translate nonresponse impact on the auxiliary variables used in the estimation to just one value. Partial R-indicators, which can be estimated at the variable or the category level (for details see Schouten et al. 2011, 5–6), permit a more fine-grained investigation into which variables or subcategories of the auxiliary variables contribute most to a lack of representativeness (Schouten et al. 2011; Beullens and Loosveldt 2012). This makes them useful for fieldwork management for example, as a basis for decisions to direct additional fieldwork effort to under-represented groups with the aim of achieving a balanced sample (Schouten et al. 2016). The intended use of the R-indicator may imply different considerations about which auxiliary variables are most suitable for their estimation (assuming such data are available to begin with). If the aim is to compare designs or monitor the evolution of fieldwork, then the indicators should ideally be estimated using the *same* auxiliary variables for each comparison, to allow an evaluation of the relative quality of responding samples. If intended to be interpreted in an absolute sense (i.e., for a single survey design at a single point in time), ideally as broad a range of variables as possible should be used to ensure the definition of representativeness is not too restricted (Cornesse and Bosnjak 2018, 5). Irrespective of the intended use of R-indicators, the choice of auxiliary variables used in the estimation of response probabilities is key to their interpretation (Nishimura et al. 2016).

It is noteworthy that a number of other indirect nonresponse bias indicators have been proposed for similar purposes as the R-indicator (see Wagner 2012, and more recently,

Nishimura et al. 2016, for reviews). These include the closely-related coefficient of variation of the response propensities (CV), which is more optimally suited to assessing the risk of bias in population means and totals (Schouten et al. 2009; Schouten and Shlomo 2017; Schouten et al. 2016; Schouten 2018); the coefficient of variation of subgroup response rates (Groves 2006; Wagner 2012); the coefficient of variation of nonresponse adjustments (Särndal and Lundström 2010); and the area under the curve or pseudo-R^2 (Nagelkerke 1991). Like the R-indicator, the interpretation – and utility – of such indicators similarly depends on which auxiliary variables are used in their estimation (Nishimura et al. 2016; Cornesse and Bosnjak 2018), and so the questions raised and addressed in the case study presented here have a broader relevance beyond R-indicators.

Besides their capacity to summarise representativeness with respect to the auxiliary variables, the utility of R-indicators (and other nonresponse bias indicators) also lies in their ability to detect bias in other survey variables. R-indicators provide an estimate of the upper bound of the nonresponse bias of a *hypothetical* survey variable under 'worst case scenarios' (Schouten et al. 2009, 107) – referred to as the Maximal Absolute Bias (MAB), which is equivalent to the coefficient of variation (CV) of the response propensities (Schouten et al. 2009; Beullens and Loosveldt 2012). As such, the magnitude of the R-indicator and MAB should be informative about the extent of actual bias in other survey variables. However, they cannot identify which survey variables are affected or by how much (Nishimura et al. 2016), nor whether bias will remain after adjustment on the auxiliary variables (Groves et al. 2008; Brick and Jones 2008; Kreuter and Olson 2011; Sakshaug and Antoni 2018). To be optimally informative, the choice of auxiliary variables used in the estimation of R-indicators is, once again, key (Schouten et al. 2016; Nishimura et al. 2016; Cornesse and Bosnjak 2018; Schouten 2018). The chosen auxiliary variables should not only be strongly related to the 'real' response propensities, but also to the variables of interest. Understanding that relationship is essential for interpreting (and evaluating the utility of) the nonresponse bias indicator, as well as the potential for reducing bias through adjustment.

In practice, the availability of auxiliary data for both respondents and nonrespondents is typically limited, leaving researchers little choice over which variables to use to build indicators of nonresponse bias (or nonresponse adjustment weights) – especially in the context of cross-sectional surveys (e.g. Sakshaug and Antoni 2018). If auxiliary data do exist, they typically consist of socio-demographic variables (e.g. on sampling frames), which may correlate only weakly with response probabilities and the variables of most interest to data users (Peytcheva and Groves 2009; Schouten 2018; Cornesse and Bosnjak 2018). In the case of general purpose (cross-sectional) social surveys (e.g. the International Social Survey Programme, the European Social Survey, the General Social Survey, the European and World Values Studies), where users may be interested in nonresponse impact on a diverse range of subjective variables covering many different topics, this limitation may be especially frustrating, and may, in turn, limit the value of R-indicators for nonresponse bias assessments in such studies.

2.2. *Assessments of the Performance of R-Indicators*

Because the utility of R-indicators (and other related indicators) is so dependent on the availability and power of the auxiliary variables used in their estimation, assessments of

their performance should ideally be focused on the latter. To date, however, relatively few studies have investigated how R-indicators perform under different conditions or how the choice of auxiliary variables used (and other factors) influence the utility of the information the indicators provide. This is partly because nonresponse biases on survey variables are usually unknown, rendering the validation of indirect indicators of bias risk challenging. This section provides a short review of available studies and their conclusions.

Cornesse and Bosnjak (2018) conducted a meta-analysis investigating the effect of different survey design variables on the representativeness of survey samples, including the number (though not type) of auxiliary variables used in the estimation and its relation to the magnitude of the R-indicator and the MAB. They hypothesised that the more auxiliary variables included, the more likely it is to detect bias (i.e., the smaller the value of the R-indicator and the larger the MAB (Cornesse and Bosnjak 2018, 5). However, over 104 R-indicator studies, they did not find the anticipated relationship.

A theoretical contribution by Schouten (2018) considered the type of auxiliary variables used to detect bias (using the CV) and their degree of association with survey variables affects their capacity to detect bias on other variables. He presents a framework in which the socio-demographic auxiliary variables that are typically available (and used for bias detection and adjustment) are viewed as just one possible selection from the universe of potential variables on a population. Using simulations and an application to the problem of attrition bias in the Dutch online Longitudinal Internet Studies for the Social Sciences (LISS) panel, he attempts to show how the level of association between select survey and auxiliary variables (comparing standard socio-demographic variables with random draws of 20 alternative variables taken from prior waves of the panel) may influence the potential to detect bias. He concludes that auxiliary variables selected at random appear able to detect a (predictable) amount of the total bias (more bias in the auxiliary variables implying more expected bias in the survey variables). However, the standard socio-demographic covariates generally outperform any random selection.

Schouten's (2018) conclusions raise the question of whether larger bias detected by available (sociodemographic) auxiliary variables (which are not randomly selected) is also a sign of larger bias in other variables (i.e., above what would be predicted by a random selection of covariates). This question was addressed by Schouten et al. (2016) in a study investigating the usefulness of R-indicators in the context of adaptive survey design (the original motivation also for Schouten's 2018 article). As the aim of such designs is to reduce bias by targeting fieldwork strategies to particular subgroups to optimise the balance of the response sample on auxiliary variables, it is of interest to know whether the (logistically more costly) targeted approach is more effective at reducing bias in survey estimates than simply adjusting on the same variables Schouten et al. (2016, 728), and hence, how informative the magnitude of the R-indicators are about the extent of bias after adjustment. Across 14 data sets, the authors found that achieving a balanced sample through adaptive design guided by such indicators was generally beneficial, resulting, on average, in less nonresponse bias in target survey variables even after adjustment (though the need for adjustment was not eliminated completely; Schouten et al. 2016, 745). However, due to some inconsistencies they observed, they conclude that more research is needed to provide further guidance as to the conditions under which a higher R-indicator

or lower CV (i.e., a more balanced response sample with respect to the auxiliary variables) implies less bias in the survey variables (Schouten et al. 2016, 745).

Finally, Nishimura et al. (2016) used simulation studies to compare the R-indicator to a number of alternative nonresponse bias indicators, under different scenarios varying response rates, missing data mechanisms (i.e., whether data are missing at random (MAR), missing completely at random (MCAR) or not missing at random, NMAR), and at varying levels of correlation between the auxiliary data and the survey data. They found that R-indicators did not perform well at indicating the magnitude of the bias on survey variables, though their effectiveness in this regard depended on the missing data mechanism (Nishimura et al. 2016, 54). Especially at low values, R-indicators give some indication of whether the data are MAR rather than MCAR. However, it is not possible to distinguish between MCAR and NMAR mechanisms, especially when the value of the indicator is large. On the assumption that available auxiliary data could be used to adjust bias on survey variables (as in Schouten et al.'s 2016 research), the same authors extended their analysis to investigate the circumstances in which nonresponse weight-adjusted means showed less bias than the unadjusted means, and obtained mixed findings across different survey variables. Though this is to be expected theoretically, it warrants further investigation to inform our understanding of what the magnitude of the R-indicator implies for bias in other survey variables after adjustment on the auxiliary variables (Nishimura et al. 2016, 59).

2.3. *The Present Study*

We address some of the issues raised above in the present case study, in which we investigate the effectiveness of R-indicators (and the related CV) as a measure of the risk of nonresponse bias in the context of the European Social Survey (ESS). Specifically, the study addresses the following research questions:

- RQ1: To what extent are available auxiliary data suitable for estimating response propensities and the risk of nonresponse bias using R-indicators? Or, specifically, how well do response propensities estimated on the basis of available auxiliary variables correlate with target survey variables?
- RQ2: To what extent are R-indicators based on the available auxiliary variables good predictors of nonresponse biases on target variables?
- RQ3: To what extent is the magnitude of the R-indicator informative about bias in the target variables once bias in the auxiliary variables has been adjusted? In other words, does more nonresponse bias on auxiliary variables (i.e., a lower R-indicator) imply more bias on other variables, even after adjustment for the auxiliary variables?

To tackle these questions, we use auxiliary data from a sampling frame based on population registers to estimate sample members' response propensities and the R-indicator/CV. We then assess the correlation between the response propensities with a selection of target survey variables. Finally, we examine the extent of the 'actual' nonresponse biases in the target variables, estimated on the basis of a nonresponse follow-up survey, and compare these to the predicted risk of bias provided by the R-indicator, before and after adjustment on the auxiliary variables.

3. Data and Methods

3.1. Data

To address our research questions, we use data from Round 5 (2010) of the Swiss European Social Survey (ESS). The ESS is a biennial cross-national face-to-face survey of social values and attitudes. The questionnaire consists of a repeated core of items aimed to measure changing social attitudes and values, and two 'rotating' modules focused on specific topics, which change in each round. The ESS target population is defined as all resident adults (aged 15 and over) within private households, 'regardless of their nationality, citizenship, language or legal status' (ESS5 – 2010 Documentation Report 2012). The Swiss Federal Statistical Office (SFSO) supplied the Swiss ESS National Coordinator with a single-stage equal probability systematic sample of individuals from this population, with no clustering, proportionally stratified by the seven NUTS-2 regions of Switzwerland (CH01 – Région lé manique; CH02 – Espace Mittelland; CH03 – Nordwestschweiz; CH04 – Zürich; CH05 – Ostschweiz; CH06 – Zentralschweiz; CH07 – Ticino). The total number of issued sample units was 2,850, and the final number of valid interviews was 1,506 – a total response rate of 53.2% (equivalent to AAPOR Response Rate 1). For more details on the fieldwork protocol and response enhancement methods used for the main survey, see Roberts et al. (2014a).

As well as using questionnaire data from the main survey interview, we analyse data from the Swiss Federal Statistical Office (SFSO)'s sampling frame of residents in Switzerland, based on population registers maintained by municipalities. In addition to individual names and addresses, the frame contains a number of socio-demographic variables, including the individual's sex, date of birth, marital status, and nationality. On the basis of address information, additional contextual variables are derived, including the linguistic region of Switzerland (French, German, Romansch or Italian), and the degree of urbanicity. Telephone numbers were obtained for 61% of the sample via an automatic search by the fieldwork agency in the commercial database ('AZ Direct'), so the auxiliary variables additionally include an indicator of whether or not a telephone number was available. This variable is known to be an important correlate of response propensity, in part because telephone contacts are used in refusal conversion and as a means to reduce the noncontact rate.

The third source of data used in this study comes from a nonresponse follow-up (NRFU) survey (Ernst Stähli et al. 2018), which was a postal survey carried out two months after the end of the main survey fieldwork, and consisted of a single sheet (double-sided) paper questionnaire with around 20 questions. After removing ineligible sample units, the nonresponse questionnaire was sent to 1,047 non-respondents (186 refusals were not recontacted for reasons not known). Efforts to improve response rates if the questionnaire had not been completed and returned within four weeks included re-contacts by telephone (if a number was available) or by mail (if no number was available). The response rate for the NRFU was 55.7%, yielding a total of 583 cases for analysis. In total, therefore, 2,089 cases (73.3% of the total sample) responded to either the main survey or the NRFU, leaving 26.7% in a group we refer to here as 'persistent nonrespondents'. Further details of final outcome rates are available in Table 1. From now on, we refer to the group of respondents to the main survey together with the respondents of the NRFU as the 'reduced' sample, in contrast to the original 'complete' sample, which includes the persistent non-respondents.

Table 1. ESS5 2010 final outcome rates (Switzerland).

Break-down of final response and nonresponse:	N	%
Total number of issued sample units	2,850	100.0
Refusal by respondent	713	25.0
Refusal by proxy (or household or address refusal)	76	2.7
No contact	278	9.8
Language barrier	67	2.4
Respondent mentally or physically unable to participate	64	2.3
Respondent unavailable throughout fieldwork period	109	3.8
Address ineligible[1]	20	0.7
Respondent moved abroad	10	0.4
Respondent deceased	7	0.3
Number of valid interviews	1,506	52.8
Total non-respondents eligible for follow-up[2]:	1,047	100.0
Non-contacts	278	26.6
Refusals and refusals by proxy (excluding office refusals)	769	73.5
Completed NRFU questionnaires by non-respondents:		
On paper	530	50.6
By telephone	53	5.1

Notes. [1]Not residential, not occupied, not traceable or other ineligible. [2]Does not include respondents who were not sent the nonresponse follow-up questionnaire.

The selection of the 23 items included in the NRFU was based partly on decisions taken in collaboration with the Core Scientific Team of the ESS (see Stoop et al. 2010; Matsuo et al. 2010). Items were selected on the assumption that they might be particularly likely to correlate with variables influencing the decision to participate in the survey and thus be at risk of nonresponse bias. The complete NRFU questionnaires are available in online Supplemental material. Details of the variables analysed here are shown in Figure 1.

Data from nonresponse surveys can suffer from timing of the fieldwork, context, and mode of data collection effects, depending on the design of the shorter questionnaire and how it is administered (Voogt and Saris 2005). These artefacts may hinder comparisons with the answers given by respondents to the main survey questionnaire and hence, the overall assessment of nonresponse bias. To remedy this issue, the ESS NRFU questionnaire was additionally sent to a random subsample of 300 respondents to the main survey to enable an assessment of measurement differences resulting from the delayed timing of fieldwork, the change in mode and possible context effects from shortening the questionnaire. Based on an analysis of this sample, six out of the 23 items in the questionnaire were found to suffer from low reliability and were, therefore, excluded from the analysis of nonresponse bias reported here (see Vandenplas et al. 2015 for further details). This resulted in a total of target variables measured in the main survey and the NRFU, for the assessment of nonresponse bias. Moreover, to further minimise the potential impact of differences between the two sources in the distribution of responses across ordinal response categories, we recoded them into binary variables.

The success of the nonresponse survey approach also depends on the extent to which respondents to NRFUs are representative of all non-respondents to the main survey (Cobben 2009). Following the continuum of resistance theory (Lin and Schaeffer 1995; Stoop 2004),

1. **Living with a partner** (1 'Living with a partner' 0 'Not living with a partner')
2. **In paid work** (1 'In paid work' 0 'Other main activity')
3. **High school education only** (1 'Primary, secondary school or vocational/training school' 0 'Higher levels of education')
4. **Fixed line telephone** (1 'Fixed line telephone in accommodation' 0 'No fixed line telephone')
5. **Registered fixed line number** (1 'Fixed line number registered' 0 'No fixed line or fixed line number not registered')
6. **Mobile telephone** (1 'Respondent has a mobile phone' 0 'Respondent does not have a mobile phone')
7. **Registered mobile number** (1 'Registered' 0 'Not registered')
8. **Good health** (1 'Very good or good' 0 'Fair, bad or very bad')
9. **Extremely happy** (1 'Extremely happy (7, 8, 9, or 10 on 11-point scale)' 0 'Not extremely happy')
10. **Takes part in social activities** (1 'More or much more than most' 0 'About the same, less or much less than most')
11. **Meets people socially frequently** (‘1 Several times a week or everyday' 0 'Once a week or less often')
12. **Very or quite interested in politics** (1 'Very or quite interested' 0 'Hardly or not at all interested')
13. **Satisfied with democracy** (1 'Extremely satisfied (7, 8, 9, 10 on 11-point scale)' 0 'Not extremely satisfied')
14. **Immigrants make country better** (1 'Immigrants make Switzerland a better place to live (7, 8, 9, or 10 on 11-point scale)' 0 'Immigrants do not make Switzerland a better place to live')
15. **Has complete trust in justice** (1 'Almost complete trust (7, 8, 9, 10 on 11-point scale)' 0 'Less than complete trust')
16. **Number of children** (0, 1, 2+)
17. **Number of people in the household** (1,2,3+)

Fig. 1. Coding of variables in the nonresponse survey.

which places respondents at the first contact attempt at one end of the continuum and nonrespondents at the other end, respondents to the NRFU can be considered to be situated somewhere between the respondents to the main survey and the persistent nonrespondents in terms of the characteristics measured by the survey (Lin and Schaeffer 1995; Stoop 2004). In this article, our analysis rests on the assumption that the nonrespondents participating in the NRFU are representative of all the nonrespondents to the main survey, including the persistent nonrespondents (and where possible, we validate this assumption with the sampling frame data). Additionally, we assume the answers to the NRFU survey are a good measure of the answers the respondents would have given had they participated in the main survey. We come back to these assumptions in the Discussion.

3.2. Analytic Approach

3.2.1. Estimating Response Propensities

To estimate the response propensities, we estimated the parameters of logistic regression equations predicting each sample member’s probability of participating in the survey

using covariates from the sampling frame data (Roberts et al. 2014a). These were respondent sex (coded 1 if male); age categories (<30 years, 31–44, 45–64, leaving the group aged 65 and over as the reference); marital status (coded 1 if married or in a legal partnership, 0 if single, divorced or widowed); nationality (coded 1 for those without Swiss citizenship, 0 if Swiss); linguistic region (coded 1 if from the French or Italian-speaking regions), 0 if German or Romansch-speaking (the ten Romansch-speaking respondents were interviewed by German speaking interviewers); urbanicity (coded 1 if living in an urban area and 0 if an isolated town or rural community); and availability of a telephone number (coded 1 if available 0 if not).

For our main analyses, we estimate the response propensities twice: first, for the complete sample (i.e., predicting response to the main survey among all sample members) and second, for the reduced sample (i.e., predicting response to the main survey among respondents to either the main survey or the NRFU). To assess the implications of only focusing on the responding nonrespondents and not all nonrespondents, we also estimate the response propensities for respondents to the NRFU compared with all the persistent nonrespondents.

3.2.2. Assessing the Relation Between Auxiliary Variables and Target Variables

To evaluate the effectiveness of the auxiliary variables included in the propensity model as indicators of bias in the target variables (RQ2), we first examine the correlations between the predicted response propensities estimated from the logistic regression model for the reduced sample and responses given in the main survey to the questions that were also included in the NRFU (Gummer and Blumenstiel 2018; Sakshaug and Antoni 2018). The items included 13 questions from the core questionnaire covering a variety of topics, five of which came from the socio-demographic module. The remaining four were country-specific items about having a fixed line or mobile telephone and whether the fixed line/mobile numbers were registered. All the items were recoded into dichotomous variables, where 1 represented a positive or affirmative response to the question. Household size and number of children in the household were kept as continuous measures. However, we also recoded them into categorical indicators, and then created dummy variables for each category where the first category was the reference (single person household/ household with no children, see Figure 1).

The list of coefficients includes a mix of Pearson's r (for continuous variables), biserial and point-biserial correlations (depending on whether the dichotomy reflects a discrete or continuous relation between the response options). For this reason, we convert the coefficients to z-scores (standard normal distribution) to facilitate comparisons between them. We focus our interpretation on whether or not the correlation was statistically significant at the 95% level.

In general, if there is a strong correlation between the estimated response propensities and the survey variable, the auxiliary variables should be considered suitable as predictors of bias for this variable (Little and Vartivarian 2005). Where the correlation between the estimated response propensities and the survey variable is low, we can have less confidence in the ability of the auxiliary variables to predict bias. Low correlations could also occur if the considered variable does not suffer from nonresponse bias. As a result, it is important to conjointly look at the correlations with the response propensities alongside the nonresponse biases with the help of the NRFU.

3.2.3. Assessing Nonresponse Bias

To investigate the extent to which nonresponse resulted in bias on the 17 target variables (RQ2), we compare estimates based on respondents to the main survey, and respondents to the NRFU, before and after adjustment for nonresponse bias on the auxiliary variables. To assess the size of the bias we compute the difference (contrast) in the proportion of respondents to the main survey and respondents to the NRFU selecting the categories coded 1, and use Chi-square tests of association to test whether the difference in proportions are statistically significant before and after adjustment on the auxiliary variables. To adjust for nonresponse bias in the auxiliary variables, we computed weights on the basis of the propensity scores from the logistic regression models (Little 1986), once for the complete sample, and once for the reduced sample. The weight for the respondents was estimated as the inverse of the propensity score, while that for the nonrespondents was calculated as one minus the inverse of the propensity score.

3.2.4. Assessing the R-Indicator As a Predictor of the Risk of Bias

The R-indicators were also estimated on the basis of the predicted response propensities from the logistic regression models using the R tool developed by De Heij et al. (2015). In addition, we also compute the adjusted coefficient of variation (CV) in the response propensities (following the formula provided by De Heij et al. (2015, 18 (14)), which is relevant when considering population means and totals (De Heij et al. 2015) and is equivalent to the Maximum Absolute Bias (MAB). As mentioned, the MAB is defined as the largest possible nonresponse bias on an estimate of a population mean in a survey with a response rate of less than 100%. To assess whether the R-indicator is a good predictor of the risk of bias (RQ2), we additionally estimate the Maximal Absolute Contrast (MAC) following the formula provided by Schouten et al. (2010). The MAC is defined as the largest *possible* difference between the respondents and non-respondents on an estimate of a population mean in a survey with a response rate of less than 100% (Schouten et al. 2010). If the estimated R-indicator is a good predictor of bias, the MAC should give a realistic upper limit for any actual difference observed between main survey and NRFU. Note that while the MAC gives the maximum possible difference between respondents and nonrespondents, the CV/MAB represents the maximum bias, in other words, the maximum difference between the respondents to the main survey and the total complete and reduced samples.

Finally, to address the question of whether more nonresponse bias on the auxiliary variables (i.e., a lower R or higher CV) implies more nonresponse bias on the survey variables after adjustment for the auxiliary variables (RQ3), we estimate the difference in the contrast between respondents and nonrespondents in the reduced sample, after adjustment on the auxiliary variables, and compare the effects of the two different propensity score weights.

4. Results

The presentation of the results is organised around our research questions. We start by presenting the results of the logistic regression models for estimating the response

propensities for the complete and the reduced sample, and the correlations between response propensities and the target variables measured in the main survey (RQ1). Then, we present the unadjusted biases on the target variables, and the R-indicators and related risk-of-bias indicators (RQ2). Finally, we present the nonresponse adjusted biases on the target variables for both samples, together with their contrasts (RQ3).

4.1. *Predicted Response Propensities*

Coefficients for the parameters of the logistic regression models estimated for the complete and reduced samples are given in Table 2. With the exception of sex, all variables included in the model for the complete sample were significantly associated with the propensity to respond to the main survey. Living in an urban area, residing in the French or Italian regions of Switzerland compared with the German region, and being a foreigner were all negatively associated with responding to the survey, while being aged 15–30 or 45–65 (compared to being over 65), being married, and having a registered telephone number were positively associated with responding. When we replace the complete sample with the reduced sample, living in an urban area, being married, being a foreigner and being aged 45–65 (compared to older) are no longer significantly associated with responding to the survey. This could be an indication that the NRFU fails to increase the level of participation for certain subgroups. Overall, however, it implies that there is somewhat less nonresponse bias on the auxiliary variables in the reduced sample compared with the complete sample.

Table 2. Parameter coefficients for logistic regression equations estimating response propensities for the reduced and complete samples.

	Complete sample			**Reduced sample**		
Parameter	$\hat{\beta}$	p	SE	$\hat{\beta}$	p	SE
Male	0.08		0.08	0.18^{+}		0.10
Urban	−0.31***		0.09	−0.24		0.11
Linguistic region (ref. German)						
French	−0.34***		0.09	−0.48***		0.11
Italian	−0.55**		0.19	−0.70**		0.23
Nationality (ref. Swiss)						
Bordering countries	−0.41**		0.14	−0.07		0.19
Other countries	−0.68***		0.13	−0.12		0.17
Age category (ref. 65+)						
15–30 years	0.54***		0.13	0.37*		0.17
31–44 years	0.09		0.12	−0.09		0.15
45–65 years	0.30**		0.11	0.12		0.14
Married	0.25**		0.09	0.13		0.11
Telephone number available	0.38***		0.08	0.34**		0.11
Constant	−0.07		0.14	0.82***		0.18
N	2,850			2,089		
Nagelkerke R^2	0.07			0.04		
Hosmer-Lemeshow's Test	0.73			0.31		

Notes: $\hat{\beta}$ = unstandardized beta coefficient; SE = standard error; ref. = reference category; $^{+}p < .1$, $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$; Data source: ESS 2010.

Note that both models have a poor fit (p-value for the Hosmer-Lemeshow's Test of 0.31 for the reduced sample and 0.73 for the complete sample and 60% versus 62% correspondence between expected and observed outcomes), indicating that taken together, the available auxiliary data explain little of the variance in the probability of responding to the survey. Nagelkerke's R^2 for both models is also low, which could be taken as a positive indication of the overall magnitude of bias on these variables.

4.2. *Correlations Between the Response Propensities and the Target Variables*

Of the 17 target variables analysed, 11 were significantly correlated with the predicted response propensities (right-hand side of Table 3, columns 4 and 5). Nine of these variables were positively correlated with the response propensities, while the other two were negatively correlated (number of children reported to be living in the household and believing that immigrants do not make Switzerland a better place to live). The six variables not significantly correlated with the response propensities were: having a registered mobile telephone number, being in paid work, being in good health, frequently meeting people socially, taking part more often in social activities, and having complete trust in the justice system. Of these, the first five were selected for the NRFU because they were presumed to relate to a sample member's contactability/time availability to participate in the survey and hence be at risk of bias if noncontact rates indeed vary as a function of these characteristics. If this is the case and these target variables are consequently affected by bias, then the socio-demographic variables used to estimate the response propensities would be ineffective for predicting the risk of bias on these measures. However, as mentioned, low correlations with the response propensities may also result from a lack of bias in the target variables, so for this reason, we need to assess the correlations alongside the nonresponse biases in the target variables, which we do in more detail in the following.

4.3. *Actual Nonresponse Bias*

Shown in the left-hand side of Table 3 are the unadjusted estimates for the target variables based on the main survey respondents (column 1) and the NRFU respondents (column 2), together with the contrast (column 3 – ordered according to size). In total, nine out of 17 estimates were affected by nonresponse bias (statistically significant differences between the respondents and nonrespondents). Seven of these were among the variables that were correlated significantly with the response propensities (shaded in grey). These included five factual variables: number of people and number of children in the household, education, having a fixed line telephone and a registered fixed line telephone number; and two subjective variables: extremely happy, and satisfied with democracy. Note that the correlation with the 'telephone' variables and to a certain extent the number of people in the household (from the main survey and nonresponse follow-up) was to be expected given that variables about having a registered phone number and marital status (from the frame data) were included in the response propensity model. For all these variables, the differences between respondents and nonrespondents were in the expected direction, depending on the sign of the correlation coefficient: nonrespondents scoring lower if the correlation was positive and higher if it was negative (which, in fact, was only the case of

Table 3. Differences in unadjusted estimates based on the main survey and NRFU, with correlation coefficients between predicted response propensities and target variables.

Target variables	*n*	(1) Main survey (n = 1506) %(SE)	(2) NRFU (n = 583) %(SE)	(3) Contrast %(SE) *p*	(4) Correlation coefficients *r*	(5) *z* *p*
1. Takes part in social activities more	2,063	18.0 (1.0)	18.0 (1.6)	0.0 (1.8)	rb = −0.06	−0.43
2. In paid work	2,057	57.7 (1.3)	57.5 (2.1)	−0.0 (2.5)	rpb = −0.02	−0.66
3. Very or quite interested in politics	2,083	58.9 (1.3)	58. 3 (2.1)	−0.6 (2.5)	rb = 0.16	5.62***
4. Mobile telephone	2,062	87.7 (0.9)	88.5 (1.3)	0.8 (1.6)	rpb = 0.04	2.24*
5. Living with a partner	2,079	62.3 (1.3)	63.2 (2.0)	0.9 (2.4)	rpb = 0.05	1.90*
6. Good health	2,084	81.4 (1.0)	79.2 (1.7)	2.2 (1.9)	rb = −0.01	−0.21
7. Immigrants make country better	2,044	29.2 (1.2)	26.1 (1.8)	−3.2 (2.1)	rb = −0.12	−3.95***
8. Registered mobile number1	1,798	7.3 (0.7)	9.8 (1.3)	2.5 (1.5)	rpb = 0.08	0.01
9. Number of people in household (mean)	2,052	2.8 (0.0)	2.6 (0.06)	−0.2 (0.1)*	r = 0.10	3.20**
Single household	378	17.8 (1.0)	20.0 (1.7)	−2.1 (1.9)	rpb = −0.16	−6.31***
Two household members	750	35.5 (1.2)	39.6 (2.1)	−4.1 (2.4)$^{+}$	rpb = 0.05	1.85^{+}
Three or more household members	924	46.7 (2.1)	40.5 (1.3)	6.2 (2.4)*	rpb = 0.08	3.03***
10. High school education only	2,076	59.7 (1.2)	54.8 (2.1)	−4.9 (2.4)*	rpb = 0.10	3.83***
11. Number of children in household (mean)	2,071	0.5 (0.0)	0.58 (0.04)	0.1 (0.0)**	r = −0.06	2.40**
Household without child	1,513	74.1 (1.1)	70.3 (1.9)	3.8 (2.1)$^{+}$	rpb = 0.12	4.55**
Household with one child	256	11.8 (0.8)	14.0 (1.5)	−2.2 (1.6)	rpb = −0.07	−2.52**
Household with two or more children	302	14.1 (0.9)	15.8 (1.5)	−1.6 (1.7)	rpb = −0.09	−3.37***
12. Has complete trust in justice	2,046	53.8 (1.3)	46.8 (2.1)	−7.0 (2.2)**	rb = 0.03	0.47
13. Fixed line telephone	2,082	90.6 (0.8)	83.7 (1.5)	−6.9 (1.7)***	rpb = 0.18	6.88***
14. Meets people socially frequently	2,082	52.3 (1.3)	43.1 (2.1)	−9.2 (1.4)***	rb = 0.01	0.19
15. Extremely happy	2,081	89.2 (0.8)	80.0 (1.7)	−9.2 (1.9)***	rb = 0.10	2.44*
16. Satisfied with democracy	2,045	69.7 (1.2)	57.5 (2.1)	−12.1 (2.4)***	rb = 0.08	2.30*
17. Registered fixed line number1	1,798	94.9 (0.6)	79.5 (1.9)	−15.4 (2.0)***	rpb = 0.15	8.75***

Notes. [1]Questions only asked of respondents who reported having a fixed-line or mobile telephone. *r* denotes Pearson's *r*, r_{pb} the point-biserial correlation and r_b the biserial correlation. ***p < .001, **p < .01, *p < .05,$^{+}$p < .1. Shaded cells show variables with both significant contrasts and correlations.

number of children in the household). For this set of target variables, therefore, a post-survey nonresponse adjustment or a targeted fieldwork based on the available auxiliary variables should substantially reduce the nonresponse bias (and this is indeed the case – see below).

The two remaining variables that were affected by nonresponse bias but not significantly correlated with the response propensities were frequently meeting people socially and having complete trust in justice. This finding suggests that these two variables are strongly related to the nonresponse mechanism (nonrespondents having less trust in justice and meeting people less frequently), but that the estimated response propensities used to build the R-indicator fail to predict bias because the variables are not related to the auxiliary variables. For these variables, therefore, using the available auxiliary data for predicting the risk of bias, or for the purposes of post-survey adjustment or targeted fieldwork strategies would fail to correct the nonresponse bias (and this is also confirmed below).

For the eight variables where no nonresponse bias was observed (i.e., where there were no significant differences between the respondents and nonrespondents), four (living with a partner, having a mobile phone, being very or quite interested in politics, and believing immigrants make the country better) were among those that were significantly correlated with the estimated response propensities. The correlations suggest that bias could potentially arise as a result of nonresponse, but with the reduced sample of nonrespondents observed here, no bias is detected. For the remaining four variables where no bias was detected, the correlation with the estimated response propensities was not significantly different from zero. These included taking part in social activities more often than other people, being in paid work, having a registered mobile phone and being in good health.

4.4. R-Indicators As Predictors of the Risk of Bias

The response rate for the complete sample (i.e., the actual survey outcome, without taking the ineligibles into account) was 52.8% (see column 1 of Table 4), and the adjusted R-indicator was 0.79. The response rate increases to 72.1% and the value of the R-indicator increases to 0.86 (column 2 of Table 4), when they are calculated on the basis of the reduced sample (i.e., when the 'persistent nonrespondents' are removed from the sample).

Table 4. Response rates, adjusted R-indicators, coefficients of variation (CV) and maximum absolute contrast (MAC) for the complete and reduced samples and for the non-respondents.

	(1) Complete sample (n = 2,850)	**(2) Reduced sample (n = 2,089)**	**(3) NRFU compared to all non-respondents (n = 1,344)**
Response sample size	1,506	1,506	583
Response rate[1]	52.8%	72.1%	43.4%
Adjusted R-indicator	0.79	0.86	0.83
Confidence interval	(0.75–0.82)	(0.82–0.90)	(0.78–0.88)
Adjusted CV	0.20	0.10	0.20
MAC	0.44	0.35	0.35

Notes. [1]Response rate calculated here as total number of interviews divided by the sample size (i.e., it does not take account of ineligibles). Data source: ESS 2010.

The coefficient of variation (MAB) for the complete sample is 0.21 (21%), while for the reduced sample it is 0.10 (10%), so using the reduced sample instead of the complete sample underestimates the risk of bias by almost 11% (according to R-indicator estimated on the basis of the available auxiliary variables). This suggests the respondents to the NRFU are indeed situated somewhere 'between' the respondents to the main survey and the extreme non-respondents, and shows how the NRFU may fail to detect bias on certain variables.

To fully assess predictions of the risk of bias in target variables provided by the R-indicator and CV, we also consider their implications for the bias that remains after adjusting on the auxiliary variables (RQ3). We compare the contrasts (i.e., the differences between estimates based on the respondents to the main survey and respondents to the NRFU) before and after applying the propensity score weighting adjustment a) based on the logistic regression predicting participation for the complete sample (columns 1–3 of Table 5) and b) based on the logistic regression predicting participation for the reduced sample (columns 4–6 of Table 5). With both the complete and reduced sample adjustment, bias is 'removed' from three of the variables correlated with the response propensities: mean number of people (though for two person households specifically, it persists) and number of children in the household and having only completed high school education. However, bias remains in six of the target variables, of which four were significantly correlated with the response propensities – having a fixed line telephone and registered telephone number, being extremely happy and being satisfied with democracy. In all but the latter, the size of the contrast is reduced by the adjustment, but not removed. For the two variables not correlated with the response propensities – trust in justice and meets people socially – bias remains after adjustment (the size of the contrast increases slightly for the former and reduces slightly for the latter).

The pattern of results when the two sets of weights are applied is very similar (the absolute size of the contrasts when adjusting to the reduced sample propensity scores is actually slightly larger for six of the variables). There is one exception, however. In one other variable – believing immigrants makes the country better – adjustment to the complete sample increases the contrast such that it becomes statistically significant. This variable was one of the four variables correlated with the response propensities for which no bias was observed initially. With the reduced sample adjustment, the contrast on this variable also increases, but the difference between the respondents and nonrespondents is not significant. Thus, adjustment on the auxiliary variables (irrespective of whether the complete or reduced sample is considered) has – as anticipated – a mixed, and not altogether positive, impact on the estimates. In this respect, the lower R-indicator associated with the complete sample (where there is more nonresponse bias in the auxiliary variables) does imply more bias on the other variables, even after adjustment – but also due to the adjustment. However, overall, the gain in bias from the complete sample adjustment is greater than with the reduced sample adjustment.

A final observation can be made about the size of the bias on each of the variables (see column 3 of Table 4). We hypothesised that a 'good' R-indicator estimation based on the 'reduced' sample should predict bias 'correctly' and that this can be verified by examining whether the value for the MAC – the maximum difference between respondents and nonrespondents – exceeds the observed differences between respondents and

Table 5. Differences in estimates based on the main survey and NRFU respondents, adjusting for bias on the sociodemographic variables using a) response propensity scores predicted for the complete sample; and b) response propensity scores predicted for the restricted sample.

		Complete vample weights				Restricted sample weights				
Target variables		**(1) Main survey (n=1,506) %(SE)**	**(2) NRFU (n=583) %(SE)**	**(3) Contrast %(SE) *p***	**(4) Gain in bias[2] %**	**(5) Main survey (n=1,506) %(SE)**	**(6) NRFU (n=583) %(SE)**	**(7) Contrast %(SE) *p***	**(8) Gain in bias[2]**	**(9) Diff. (7)−(3) %**
1. Takes part in social activities more	2,063	17.9 (1.0)	18.6 (1.7)	0.7 (2.0)	0.7	18.1 (1.0)	18.7 (1.7)	0.6 (2.0)	0.6	−0.1
2. In paid work	2,057	57.5 (1.3)	57.3 (2.1)	−0.2 (2.5)	−0.2	57.9 (1.3)	57.2 (2.1)	−0.8 (2.4)	−0.8	−0.6
3. Very or quite interested in politics	2,083	57.1 (1.3)	59.1 (2.1)	2.0 (2.5)	2.6	58.0 (1.3)	59.6 (2.1)	1.5 (2.5)	2.1	−0.1
4. Mobile telephone	2,062	87.2 (0.9)	88.8 (1.4)	1.6 (1.6)	0.8	87.6 (0.9)	88.7 (1.4)	1.1 (1.6)	0.3	−0.5
5. Living with a partner	2,079	60.9 (1.3)	64.0 (2.1)	3.2 (2.4)	2.3	62.0 (1.3)	63.4 (2.1)	1.4 (2.4)	0.5	−1.8
6. Good health	2,084	81.2 (1.0)	79.5 (1.7)	−1.7 (2.0)	−3.9	81.5 (1.0)	79.1 (1.8)	−2.4 (2.0)	−4.6	−0.1
7. Immigrants make country better	2,044	30.6 (1.3)	24.1 (1.8)	−6.4 (2.2)**	−3.2	29.8 (1.2)	24.3 (1.8)	−5.5 (2.2)	−2.3	−1.0
8. Registered mobile number	1,798	7.2 (0.7)	10.3 (1.4)	3.1 (1.6)	0.6	7.2 (0.7)	10.3 (1.5)	3.1 (1.6)	0.6	−0.0
9. Number of people in household (mean)	2,052	2.7 (0.0)	2.6 (0.1)	−0.1 (0.1)	0.1	2.7 (0.0)	2.6 (0.1)	−0.1 (0.1)	0.1	−0.0
Single household	378	19.8 (1.1)	18.5 (1.7)	−1.3 (1.9)	0.8	18.5 (1.0)	18.6 (1.7)	0.1 (1.9)	2.2	−1.4
Two household members	750	34.6 (1.3)	39.8 (2.2)	5.2 (2.5)*	9.3	35.2 (1.3)	39.9 (2.2)	4.7 (2.5) +	8.8	0.5
Three or more household members	924	45.7 (1.3)	41.7 (2.2)	−3.9 (2.6)	−10.1	46.3 (1.3)	41.5 (2.2)	−4.8 (2.6) +	−11.0	0.9
10. High school education only	2,076	59.3 (1.3)	56.6 (2.1)	−2.8 (2.5)	2.1	59.1 (1.3)	56.7 (2.1)	−2.4 (2.5	2.5	−0.4
11. Number of children in household (mean)	2,071	0.5 (0.0)	0.5 (0.0)	0.1 (0.1)	0.0	0.5 (0.0)	0.5 (0.0)	0.0 (0.0)	−0.1	−0.0
Household without child	1,513	73.5 (1.2)	69.4 (2.0)	−4.0 (2.3) +	−7.8	73.5 (1.2)	69.9 (2.0)	−3.7 (2.3)	−7.5	−0.3

Table 5. Continued

Target variables		Complete vample weights (1) Main survey (n = 1,506) %(SE)	(2) NRFU (n = 583) %(SE)	(3) Contrast %(SE) *p*	(4) Gain in bias[2] %	Restricted sample weights (5) Main survey (n = 1,506) %(SE)	(6) NRFU (n = 583) %(SE)	(7) Contrast %(SE) *p*	(8) Gain in bias[2]	(9) Diff. (7)–(3) %
Household with one child	256	12.2 (0.9)	13.9 (1.5)	1.8 (1.7)	4.0	12.0 (0.9)	14.1 (1.5)	2.1 (1.8)	4.3	−0.3
Household with two or more children	302	14.4 (0.9)	16.6 (1.6)	2.2 (1.9)	3.8	14.5 (0.9)	16.1 (1.6)	1.6 (1.7)	3.2	0.6
12. Has complete trust in justice	2,046	54.0 (1.3)	44.9 (2.1)	−9.0 (2.5)***	−2.0	53.7 (1.3)	45.4 (2.1)	−8.3 (2.5)***	−1.3	−0.7
13. Fixed line telephone	2,082	89.0 (0.9)	85.5 (1.4)	−3.5 (1.7)*	3.4	89.9 (0.8)	85.4 (1.5)	−4.5 (1.6)**	2.4	−1.0
14. Meets people socially frequently	2,082	52.1 (1.3)	43.2 (2.1)	−8.9 (2.5)***	0.3	52.3 (1.3)	43.2 (2.1)	−9.1 (2.5)***	0.1	−0.2
15. Extremely happy	2,081	88.5 (0.9)	80.4 (1.7)	−8.1 (1.9)***	1.1	89.0 (0.8)	80.6 (1.7)	−8.5 (1.9)***	0.7	−0.4
16. Satisfied with democracy	2,045	69.4 (1.2)	57.4 (2.1)	−12.0 (2.5)***	0.1	69.4 (1.2)	57.8 (2.2)	−11.6 (2.5)***	0.5	−0.4
17. Registered fixed line number[1]	1,798	94.0 (0.7)	82.1 (1.8)	−11.9 (1.9)***	3.5	94.6 (0.6)	82.3 (1.8)	−12.2 (1.9)***	3.2	−0.3

Notes. Data source: ESS 2010. [1]Questions only asked of respondents who reported having a fixed-line or mobile telephone. [2]Gain in bias = adjusted contrast – unadjusted contrast shown in table 3, column 3 (positive values indicate reduction in the absolute size of the contrast). + $p < .1$, * $p < .05$, ** $p < .01$, *** $p < .001$. Shaded cells show variables with both significant contrasts and correlations when unadjusted.

nonrespondents. This is indeed the case – none of the observed differences were greater than 0.35 (the value for MAC for the reduced sample).

5. Discussion and Conclusion

In the search for indicators of the risk of nonresponse bias to supplement response rates, indicators of the representativeness of the responding sample ('R-indicators' – and the related coefficient of variation (CV) of response propensities) offer considerable appeal. Yet the utility of such indicators depends on a) the availability of suitable auxiliary data for their estimation, b) how well they predict nonresponse bias on other variables in the survey, and c) whether their magnitude (i.e., what they tell us about the extent of bias in the auxiliary variables) is also informative about bias on other variables after adjustment for bias on the auxiliary variables. We investigated these issues in a case study using data from a nonresponse survey to assess the extent of actual bias in estimates of socio-demographic and attitudinal measures from the Swiss ESS, by treating respondents to an NRFU survey as though they were the complete sample of non-respondents (or at least, perfectly representative of them). Though not unproblematic (discussed further below), this set-up allowed us to address questions raised in previous research (e.g. Schouten 2018; Schouten et al. 2016; Nishimura et al. 2016) about whether the presence of more nonresponse bias in auxiliary variables necessarily translates into more bias in survey variables (the issue raised by one anonymous reviewer of whether 'where there's smoke there's fire'), and how adjusting for nonresponse on auxiliary variables affects this relationship (e.g. Schouten 2018; Schouten et al. 2016; Nishimura et al. 2016).

By examining the correlations between estimated response propensities used to build the R-indicator and variables included in the NRFU, we assessed the suitability of the available auxiliary variables (socio-demographic variables from a sample frame based on population registers) for detecting the observed bias, before and after adjustment. We then assessed the value of the R-indicator, the CV (maximum absolute bias) and the maximum absolute contrast as summary statistics of the risk of nonresponse bias, by comparing their predictions with the biases detected by the NRFU.

Our results with respect to the auxiliary variables (RQ1) were, on the one hand, reassuring. Of the nine variables that were affected by bias, seven were significantly correlated with the estimated response propensities used to calculate the R-indicator, and the observed bias was consistent with the sign of the correlation coefficients. Five of these variables were socio-demographic or other factual variables, of which three (fixed telephone number, registered telephone number and number of people in the household) were directly correlated with two auxiliary variables that were included in the response propensity model (telephone number available and marital status). On the other hand, the absence of significant correlations for the remaining variables included in the NRFU (including some affected by bias) suggests some limits to the socio-demographic variables used for detecting bias on subjective measures. The two variables affected by bias that were not correlated with the estimated response propensities were 'has complete trust in justice' (nonrespondents had lower levels of trust), and 'meets socially frequently' (nonrespondents were less likely to meet). Thus, the R-indicator and CV based on the auxiliary variables available in this study were only partially informative about the extent of bias in the survey variables (RQ2).

Whilst perhaps not surprising given that their primary purpose is to translate bias in the auxiliary variables to smaller dimensions, this finding highlights potential limitations of R-indicators for practitioners and analysts. In addition, it has implications for the possibility to adjust nonresponse bias using the same auxiliary variables in post-stratification weights. Indeed, we found that when adjusting for nonresponse using a propensity score weighting method based on the response propensities predicted for the complete sample, the contrast in estimates for respondents and non-respondents for 'trust in justice' and 'meets people socially' remained statistically significant (i.e., bias was, unsurprisingly, not reduced by the adjustment). We also found that the contrast for the variable 'immigrants make the country better' became significant only *after* adjustment on the propensity scores for the complete sample (which was not the case when using the reduced sample weights).

Thus, in this case study, the presence of *more* nonresponse bias in the auxiliary variables (resulting in a lower R-indicator), did imply slightly more bias in the target variables both before and after adjustment (RQ3) – more smoke indicating more fire. Nevertheless, there were relatively few differences in the effectiveness of the complete and reduced sample adjustment methods, so the slightly larger R-indicator obtained for the reduced sample did not imply *much less* bias in the target variables than was the case for the complete sample. Even after adjustment on the auxiliary variables, bias remained on four of the variables correlated with the response propensities for which the unadjusted contrasts were significant (as well as the two which were not). These included self-reported measures of having a fixed-line telephone and telephone number (which apparently do not concur with the auxiliary data on number availability, which came from a commercial database), and feeling extremely happy and being satisfied with democracy. The results, therefore, make it difficult to draw strong conclusions about whether a survey design with less nonresponse bias on auxiliary variables also has, on average, less bias on other variables. We recommend that future research investigates this question further.

The finding that variation in the magnitude of the R-indicator is only partially informative of the risk of bias on other variables (irrespective of the effects of weighting) concurs with the findings of other studies (e.g. Nishimura et al. 2016; Schouten et al. 2016). This limitation may be particularly relevant where subjective variables are concerned, however, and may not be entirely due to a lack of, or only weak correlations with the auxiliary variables. In particular, subjective variables may additionally be affected by substantial measurement biases (Roberts and Vandenplas 2017), which could account for some of the results observed in the comparison between the main survey and the NRFU. Nevertheless, the results suggest the need for some caution when interpreting the magnitude of the R-indicator – as well as that of related bias indicators. For example, while the MAC represents an upper limit of the contrast detected by the NRFU, we found that the values of the CV and MAC were somewhat exaggerated; a maximum nonresponse bias of 10% (reduced sample) or 21% (complete sample) as given by the CV, or a maximum contrast of 35% (reduced) or of 44% (complete) as predicted by the MAC, would likely be unacceptably high. In this sense, the predictions of the R-indicator do not map directly onto the observed nonresponse bias, meaning the R-indicator provides only a broad-brushed measure of the likely impact of nonresponse error. While it may be unrealistic to expect one value to represent the impact of nonresponse on multiple

variables, it is important for practitioners using R-indicators to be aware of the need for a more narrow interpretation of their meaning.

This having been said, and as previously discussed, partial R-indicators (and partial CVs) can, and do, provide far richer insight into how different variables (and categories of variables) are affected by nonresponse and in turn, contribute to a reduction in sample representativeness. In recognition of this, we extended the analyses presented here (Tables A1 and A2 in the online Supplemental material.) by calculating partial indicators for the reduced sample, for the auxiliary variables, and again for the target variables (using the latter to predict the response propensities). The results illustrate the advantages of having variable-level information (on the same metric) about the extent of nonresponse bias on specific variables, and lend further support to the findings reported here relating to the impact of adjusting on the auxiliary variables (namely, that adjustment has a mixed and not altogether positive effect on bias in the survey variables, and hence the magnitude of R-indicators).

While our findings are informative about some of the potential drawbacks of using R-indicators, it is important to recognise the limitations of the case study presented. As already alluded to, a principal concern is that our conclusions are sensitive to the methodological limits of the NRFU survey used to estimate bias (namely, nonresponse and differential measurement errors in the reduced sample). We treated the latter as though it was the complete sample, but our results suggest that the assumption that the NRFU respondents are representative of all the non-respondents to the ESS may not hold. While, the R-indicator for the reduced sample was substantially higher thanfor the complete sample, and the value for the CV and the MAC was substantially lower, there is evidence to suggest that the respondents to the NRFU were more similar to the respondents to the main survey than they were to the persistent nonrespondents, and that consequently, bias would be underestimated by the nonresponse survey. These findings are in line with previous research using these data. Roberts et al. (2014a) analysis found that the NRFU survey in the ESS Round 5 was successful in bringing into the overall responding sample more people from urban areas, from the French-speaking region of Switzerland, and without an available telephone number, as well as in balancing the different age categories. However, they found that it failed to improve the representation of non-Swiss citizens and the unmarried population. In this respect, it is perhaps not surprising that the nonresponse survey underestimates nonresponse bias on certain variables. It implies, however, that our analysis of the utility of the available auxiliary variables for detecting bias was restricted to a somewhat peculiar target population (of main survey and NRFU respondents), which may arguably somewhat limit the validity of our conclusions.

Another factor alluded to that is likely to hinder comparisons between the main survey and the nonresponse survey concerns the possibility that survey participation propensity and measurement error are interrelated. A large literature exists on whether reluctant respondents (refusals or hard-to-contact) are more likely to give inaccurate answers than motivated respondents (e.g. Roberts et al. 2014b; Peytchev et al. 2010; Olson 2006; Tancreto and Bentley 2005; Yan et al. 2004). In Olson's (2013) review of the published literature, she found considerable support for the conclusion that data from respondents recruited after many follow-ups or refusal conversion procedures were of lower quality.

However, it is not clear what to expect in terms of data quality for the NRFU surveys conducted for the purpose of nonresponse bias detection. On the one hand, there is a risk that respondents to the NRFU may be less motivated to answer and more inclined to reduce the effort to give accurate answers, which would result in bias due to measurement, and not necessarily due to selection effects. At the same time, the NRFU questionnaire is a lot shorter than the full interview, which should decrease response burden, and consequently, improve measurement quality. Similarly, it is not always clear whether and how mode differences in measurement will affect estimates. Given the lack of clarity on this matter, we consider the threat of persistent selection biases in the nonresponse bias estimates to be a greater cause for concern, but measurement bias should not be ignored. However, in the absence of alternative sources of information about the actual nonresponse bias on target variables in the ESS, we believe that the NRFU survey analysed here is still able to offer valuable insights. Nevertheless, it should be noted that the specificities of the design of this case study may restrict the possibility to extrapolate to other surveys, and so future studies should consider replicating our approach on other types of survey, with access to more and different auxiliary data, to see whether different conclusions are drawn.

Other methodological decisions we made might also have affected our conclusions, such as the use of logistic regression to estimate the response propensities, which may not produce the best propensity scores (e.g. Olmos and Govindsamy 2015). Alternative approaches such as Classification and Regression Tree models or Generalised Boosted Regression may outperform logistic regression in this regard (Olmos and Govindsamy 2015; McCaffrey et al. 2004), and deserve consideration, particularly where large numbers of auxiliary variables are available. Similarly, the method of propensity score weighting we used is not the only way to adjust for nonresponse bias in the auxiliary data. We opted for both methods for pragmatic reasons and consistency, but also because of the limited number of auxiliary variables available from the sampling frame.

These caveats aside, our analyses underline the need to carefully consider how to select auxiliary variables and survey variables for nonresponse bias assessments, and the implications of this for interpreting R-indicators. Both influence conclusions drawn about sample representativeness and the risk of bias, and sometimes unpredictably. Nevertheless, we believe the findings of this study also highlight the considerable value to be gained from using the R-indicator to summarise bias in the auxiliary variables, given the additional information it contains, compared to the response rate. Combined with partial indicators (Schouten and Shlomo 2017), the R-indicator as a summary statistic has proven useful for monitoring and managing fieldwork progress, as well as for comparing the representativeness of different survey designs equipped with the same auxiliary variables (Luiten and Schouten 2013; Schouten et al. 2012). Together, such indicators can be used to ensure that population subgroups are adequately represented in surveys to allow meaningful comparisons between subgroups or to highlight where adjustments should be made to ensure greater balance in the response sample. Given the implications of greater balance for bias on other survey variables may be less obvious, however, practitioners should be encouraged to fully assess the implications of alternative model specifications when estimating response propensities for conclusions about representativeness and bias risk.

6. References

AAPOR, American Association for Public Opinion Research. 2016. Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys. 9th edition.

Bethlehem, J.G. 2002. "Weighting Non-response Adjustment Based on Auxiliary Information." In *Survey Non-response*, edited by R.M. Groves, D.A. Dillman, J.L. Eltinge, and R.J.A. Little, 41–54. New York: Wiley.

Beullens, K. and G. Loosveldt. 2012. "Should High response Rates Really be the Primary Objective?" *Survey Practice* 5(3): 1–5. DOI: https://doi.org/10.29115/SP-2012-0019.

Beullens, K., G. Loosveldt, C. Vandenplas, and I. Stoop. 2018. "Response Rates in the European Social Survey: Increasing, Decreasing, or a Matter of Fieldwork Efforts?" *Survey Methods: Insights from the Field.* DOI: https://doi.org/10.13094/SMIF-2018-00003.

Brick, J.M. and M.E. Jones. 2008. "Propensity to respond and nonresponse bias." *METRON – International Journal of Statistics*, LXVI(1), 51–73.

Brick, J.M. and D. Williams. 2013. "Explaining rising Non-response Rates in Cross-sectional Surveys." *The ANNALS of the American Academy of Political and Social Science* 645: 36–59. DOI: https://doi.org/10.1177/0002716212456834.

Brick, J.M. and R. Tourangeau. 2017. "Responsive Survey Designs for Reducing Non-response Bias." *Journal of Official Statistics* 33(3): 735–752. DOI: http://dx.doi.org/10.1515/JOS-2017-0034.

Cobben, F. 2009. *Non-response in Sample Surveys. Methods for Analysis and Adjustment*. PhD thesis, The Hague: Statistics Netherlands. Available at: https://hdl.handle.net/11245/1.312964 (accessed June 2020).

Cornesse, C. and M. Bosnjak. 2018. "Is there an association between survey characteristics and representativeness? A meta-analysis." *Survey Research Methods* 12(1): 1–13. DOI: https://doi.org/10.18148/srm/2018.v12i1.7205.

De Heij, V., B. Schouten, and N. Shlomo. 2015. *RISQ manual 2.1. Tools in SAS and R for the computation of R-indicators, partial R-indicators and partial coefficients of variation*. RISQ Project. Available at: www.risq-project.eu.

De Leeuw, E. and W. de Heer. 2002. "Trends in Household Survey Non-response: A Longitudinal and International Comparison." In *Survey Non-response*, edited by R.M. Groves, D.A. Dillman, J.L. Eltinge, and R.J.A. Little, 41–54. New York: Wiley.

Ernst Stähli, M., D. Joye, M. Sapin, A. Pollien, M. Ochsner, and A. van den Hende. 2018. "Non Response Surveys (NRS): ESS 2006, EVS 2008, ESS 2010, MOSAiCH 2011, ESS 2012, ESS 2014." [Dataset]. Distributed by FORS, Lausanne. https://doi.org/10.23662/FORS-DS-697-1.

ESS Round 5: European Social Survey 2016: ESS-5 2010 Documentation Report. Edition 4.1. Bergen, European Social Survey Data Archive, NSD – Norwegian Centre for Research Data for ESS ERIC. Available at: https://www.europeansocialsurvey.org/docs/round5/survey/ESS5_data_documentation_report_e04_2.pdf (accessed June 2020).

ESS Round 5: European Social Survey Round 5 Data. 2010. *Data file edition 3.4*. NSD – Norwegian Centre for Research Data, Norway – Data Archive and distributor of ESS data for ESS ERIC. Available at: https://www.europeansocialsurvey.org/download.html?file=ESS5CH&c=CH&y=2010.

Groves, R.M. 2006 . "Non-response Rates and Non-response Bias in Household surveys." *Public Opinion Quarterly* 70, Special Issue: 646–675. DOI: https://doi.org/10.1093/poq/nfl033.

Groves, R.M. and S. Heeringa. 2006. "Responsive design for household surveys: tools for actively controlling survey errors and costs." *Journal of the Royal Statistical Society*: Series A (Statistics in Society), 169(3): 439–457. DOI: https://doi.org/10.1111/j.1467-985X.2006.00423.x

Groves, R.M. and E. Peytcheva. 2008. "The Impact of Non-response Rates on Non-response Bias – A Meta-Analysis." *Public Opinion Quarterly* 72: 167–189. DOI: https://doi.org/10.1093/poq/nfn011.

Groves, R.M. and M.P. Couper. 1998. *Non-Response in Household Interview Survey*. New York: John Wiley & Sons.

Groves, R.M., J.M. Brick, M.P. Couper, W. Kalsbeek, B. Harris-Kojetin, F. Kreuter, B.-E. Pennell, T. Raghunathan, B. Schouten, T. Smith, R. Tourangeau, A. Bowers, M. Jans, C. Kennedy, R. Levenstein, K. Olson, E. Peytcheva, S. Ziniel, and J. Wagner. 2008. "Issues Facing the Field: Alternative Practical Measures of Representativeness of Survey Respondent Pools." *Survey Practice* 1(3): 1–6. DOI: https://doi.org/10.29115/SP-2008-0013.

Gummer, T. and J.E. Blummenstiel. 2018. "Experimental Evidence on Reducing Nonresponse Bias through Case Prioritization: The Allocation of Interviewers." *Field Methods* 30(2): 124–139. DOI: https://doi.org/10.1177/1525822X18757967.

Kreuter, F. 2013. "Facing the Nonresponse Challenge." *The Annals of the American Academy of Political and Social Science* 645(1): 23–35. DOI: https://doi.org/10.1177/0002716212456815.

Kreuter, F. and K. Olson. 2011. "Multiple Auxiliary Variables in Nonresponse Adjustment." *Sociological Methods & Research* 40(2): 311–332. DOI: https://doi.org/10.1177/0049124111400042.

Lin, I.-F. and N.C. Schaeffer. 1995. "Using Survey Participants to Estimate the Impact of Nonparticipation." *Public Opinion Quarterly* 59(2): 236–258, DOI: https://doi.org/10.1086/269471.

Little, R.J.A. and S. Vartivarian. 2005. "Does Weighting for Non-Response increase the Variance of Survey Means?" *Survey Methodology* 31(2): 161–68. DOI: https://doi.org/10.2307/1403140.

Little, R.J.A. 1986. "Survey Nonresponse Adjustments for Estimates of Means." *International Statistical Review / Revue Internationale de Statistique* 54(2), 139–157. JSTOR. DOI: https://doi.org/10.2307/1403140.

Little, R.J.A. and D.B. Rubin. 2014. *Statistical Analysis with Missing Data*. London, UK: John Wiley & Sons.

Luiten, A. and B. Schouten, 2013. "Tailored Fieldwork Design to increase Representative Household Survey Response: An Experiment in the Survey of Consumer Satisfaction." *Journal of the Royal Statistical Society:* Series A (Statistics in Society) 176(1): 169–189. DOI: https://doi.org/10.1111/j.1467-985X.2012.01080.x.

Matsuo, H., J. Billiet, G. Loosveldt, F. Berglund, and Ø. Kleven. 2010. "Measurement and Adjustment of Non-response Bias based on Non-response Survey: the Case of Belgium

and Norway in the European Social Survey Round 3." *Survey Research Methods* 4(3): 165–178. DOI: http://dx.doi.org/10.18148/srm/2010.v4i3.3774.

McCaffrey, D.F., G. Ridgeway, and A.R. Morral. 2004. "Propensity Score Estimation with Boosted Regression for Evaluating Causal Effects in Observational Studies." *Psychological Methods* 9: 403–425. DOI: https://doi.org/10.1037/1082-989X.9.4.403.

Moore, J.C., G.B. Durrant, and P.W. Smith. 2018. "Data Set Representativeness during Data Collection in Three UK Social Surveys: Generalizability and the Effects of Auxiliary Covariate Choice." *Journal of the Royal Statistical Society:* Series A (Statistics in Society) 181(1): 229–248. DOI: https://doi.org/10.1111/rssa.12256.

Nagelkerke, N.J. 1991. "A Note on a General Definition of the Coefficient of Determination." *Biometrika* 78(3): 691–692. DOI: https://doi.org/10.1093/biomet/78.3.691.

Nishimura, R., J. Wagner, and M. Elliott. 2016. "Alternative Indicators for the Risk of Non-response Bias: A Simulation Study." *International Statistical Review* 84(1): 43–62. DOI: https://doi.org/10.1111/insr.12100.

Olmos, A. and P. Govindsamy. 2015. "A Practical Guide for using Propensity Score Weighting in R." *Practical Assessment, Research & Evaluation*, 20(13). Available at: http://pareonline.net/getvn.asp?v=20&n = 13. (accessed June 2020).

Olson, K. 2006. "Survey participation, Non-Response Bias, Measurement Error Bias, and Total Bias." *Public Opinion Quarterly* 70(5): 737–758. DOI: https://doi.org/10.1093/-poq/nfl038.

Olson, K. 2013. "Do Non-response Follow-ups Improve or Reduce Data Quality?: A Review of the Existing Literature." *Journal of the Royal Statistical Society:* Series A: Statistics in Society 176(1): 129–145. DOI: http://doi.org/10.1111/j.1467-985X.2012.01042.x.

Peytchev, A., E. Peytcheva, and R.M. Groves. 2010. "Measurement Error, Unit Nonresponse, and Self-Reports of Abortion Experiences." *Public Opinion Quarterly* 74: 319–327. DOI: https://doi.org/10.1093/poq/nfq002.

Peytcheva, E. and R.M. Groves. 2009. "Using Variation in Response Rates of Demographic Subgroups as Evidence of Non-response Bias in Survey Estimates." *Journal of Official Statistics* 25: 193–201. Available at: https://www.researchgate.net/publication/282119961_Using_Variation_in_Response_Rates_of_Demographic_Subgroups_as_Evidence_of_Nonresponse_Bias_in_Survey_Estimates (accessed June 2020).

Roberts, C. and C. Vandenplas. 2017. "Estimating Components of Mean Squared Error to Evaluate the Benefits of Mixing Data Collection Modes." *Journal of Official Statistics* 33(2): 303–334. DOI: https://doi.org/10.1515/jos-2017-0016.

Roberts, C., C. Vandenplas, and M. Ernst Stähli. 2014a. "Using Register Data to assess the Impact of Response Enhancement Methods on the Risk of Non-response Bias." *Survey Research Methods* 8(2): 67–80. DOI: http://dx.doi.org/10.18148/srm/2014.v8i2.5459.

Roberts, C., N. Allum, and P. Sturgis. 2014b. "Non-response and Measurement error in an online panel: Does additional Effort to Recruit Reluctant Respondents Result in Poorer Quality Data?" In *Online Panel Research: A Data Quality Perspective*. edited by M. Callegaro, R. Baker, J. Bethlehem, A.S. Göritz, J.A. Krosnick, and P.J. Lavrakas. Hoboken: Wiley, Survey Methodology Series.

Sakshaug, J.F. and M. Antoni. 2018. "Evaluating the Utility of Indirectly Linked Federal Administrative Records for Nonresponse Bias Adjustment." *Journal of Survey Statistics and Methodology* 7(2): 227–249. DOI: https://doi.org/10.1093/jssam/smy009.

Särndal, C.-E. and S. Lundström. 2010. "Design for Estimation: Identifying Auxiliary Vectors to Reduce Nonresponse Bias." *Survey Methodology* 36(2): 131–144.

Schouten, B. 2018. "Statistical inference based on randomly generated auxiliary variables." *Journal of the Royal Statistical Society:* Series B (Statistical Methodology) 80(1): 33–56. DOI: https://doi.org/10.1111/rssb.12242.

Schouten, B. and N. Shlomo. 2017. "Selecting Adaptive Survey Design Strata with Partial R-indicators." *International Statistical Review* 85(1): 143–163. DOI: https://doi.org/10.1111/insr.12159.

Schouten, B. and F. Cobben. 2007. "R-Indexes for the Comparison of Different Fieldwork Strategies and Data Collection Modes." *Statistics Netherlands. Discussion Paper*, 07002, CBC, Voorburg. Available at: http://hummedia.manchester.ac.uk/institutes/cmist/risq/schouten-cobben-2007-a.pdf (accessed June 2020).

Schouten, B., F. Cobben, and J. Bethlehem. 2009. "Indicators for the Representativeness of Survey Response." *Survey Methodology* 35(1): 101–113. Available at: https://pdfs.semanticscholar.org/aa59/4bf03a7cc219ccc6da01d1e3cb14a125d67a.pdf (accessed June 2020).

Schouten, B., F. Cobben, P. Lundquist, and J. Wagner. 2016. "Does more Balanced Response imply less Non-response Bias?" *Journal of the Royal Statistical* Society, Series A (Statistics in Society) 179(3): 727–748. DOI: https://doi.org/10.1111/rssa.12152.

Schouten, B., J. Bethlehem, K. Beullens, Ø. Kleven, G. Loosveldt, A. Luiten, K. Rutar, N. Shlomo, and C. Skinner. 2012. "Evaluating, Comparing, Monitoring, and Improving Representativeness of Survey Response Through R-Indicators and Partial R-Indicators." *International Statistical Review* 80(3): 382–399. DOI: https://doi.org/10.1111/j.1751-5823.2012.00189.x.

Schouten, B., N. Shlomo, and C. Skinner. 2010. "Indicators for Representative Response." Paper presented at Quality in Official Statistics Conference 2010, Helinski, Finland. Available at: https://q2010.stat.fi/ (accessed June 2020).

Schouten, B., N. Shlomo, and C. Skinner. 2011. "Indicators for Monitoring and Improving Representativeness of Response." *Journal of Official Statistics* 27(2): 231–253. Available at: https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/indicators-for-monitoring-and-improving-representativeness-of-response.pdf (accessed June 2020).

Stoop, I.A. 2004. "Surveying Nonrespondents." *Field Methods* 16(1): 23–54. DOI: https://doi.org/10.1177/1525822X03259479.

Stoop, I., J. Billiet, A. Koch, and R. Fitzerald. 2010. *Improving Survey Response: Lessons Learned from the European Social Survey*. London, UK: John Wiley and Sons Ltd.

Tancreto, J.G. and M. Bentley. 2005 . "Determining the Effectiveness of Multiple Nonresponse Follow-up Contact Attempts on Response and Data Quality." In Proceedings of the Section on Survey Research Methods: American Statistical Association, 2005. 3626–3632. Minneapolis, MN: American Statistical Association. Available at: http://www.asasrms.org/Proceedings/y2005f.html (accessed June 2020).

Vandenplas, C., D. Joye, M. Ernst Stähli, and A. Pollien. 2015. "Identifying Pertinent Variables for Nonresponse Follow-Up Surveys. Lessons Learned from 4 Cases in Switzerland." *Survey Research Methods* 9(3): 141–158. DOI: https://doi.org/10.18148/srm/2015.v9i3.6138.

Voogt, R.J.J. and W.E. Saris, 2005. "Mixed Mode Designs: Finding the Balance Between Non-response Bias and Mode Effects." *Journal of Official Statistics*, 21(3): 367–387. Available at: https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/-mixed-mode-designs-finding-the-balance-between-nonresponse-bias-and-mode-effects.pdf (accessed June 2020).

Wagner, J. 2012. "A Comparison of Alternative Indicators for the Risk of Nonresponse Bias." *Public Opinion Quarterly* 76(3): 555–575. DOI: https://doi.org/10.1093/poq/nfs032.

Williams, D. and J.M. Brick. 2017. "Trends in U.S. Face-To-Face Household Survey Nonresponse and Level of Effort." *Journal of Survey Statistics and Methodology* 6(2): 186–211. DOI: https://doi.org/10.1093/jssam/smx019.

Yan, T., R. Tourangeau, and Z. Arens. 2004. "When less is more: Are reluctant respondents poor reporters?" In Proceedings of the Section on Survey Research Methods: American Statistical Association, 2004. 4632–4651. Toronto: American Statistical Association. Available at: http://www.asasrms.org/Proceedings/y2004/files/Jsm2004-000169.pdf (accessed July 2020).

Received July 2018
Revised March 2020
Accepted May 2020

sciendo

Journal of Official Statistics, Vol. 36, No. 3, 2020, pp. 703–728, http://dx.doi.org/10.2478/JOS-2020-0035

Proxy Pattern-Mixture Analysis for a Binary Variable Subject to Nonresponse

Rebecca R. Andridge[1] *and Roderick J.A. Little*[2]

Given increasing survey nonresponse, good measures of the potential impact of nonresponse on survey estimates are particularly important. Existing measures, such as the R-indicator, make the strong assumption that missingness is missing at random, meaning that it depends only on variables that are observed for respondents and nonrespondents. We consider assessment of the impact of nonresponse for a binary survey variable Y subject to nonresponse when missingness may be not at random, meaning that missingness may depend on Y itself. Our work is motivated by missing categorical income data in the 2015 Ohio Medicaid Assessment Survey (OMAS), where whether or not income is missing may be related to the income value itself, with low-income earners more reluctant to respond. We assume there is a set of covariates observed for nonrespondents and respondents, which for the item nonresponse (as in OMAS) is often a rich set of variables, but which may be potentially limited in cases of unit nonresponse. To reduce dimensionality and for simplicity we reduce these available covariates to a continuous proxy variable X, available for both respondents and nonrespondents, that has the highest correlation with Y, estimated from a probit regression analysis of respondent data. We extend the previously proposed proxy-pattern mixture (PPM) analysis for continuous outcomes to the binary outcome using a latent variable approach for modeling the joint distribution of Y and X. Our method does not assume data are missing at random but includes it as a special case, thus creating a convenient framework for sensitivity analyses. Maximum likelihood, Bayesian, and multiple imputation versions of PPM analysis are described, and robustness of these methods to model assumptions is discussed. Properties are demonstrated through simulation and with the 2015 OMAS data.

Key words: Missing data; nonignorable nonresponse; nonresponse bias; survey data; bayesian methods.

1. Introduction

Response rates for large-scale surveys have been steadily declining in recent years (Curtain et al. 2005; Brick and Williams 2013), increasing the need for methods to analyze the impact of nonresponse on survey estimates. Andridge and Little (2011) argue that the assessment of impact depends primarily on three features of the available data: the nonresponse rate, differences between respondents and nonrespondents on characteristics that are observed for the entire sample, and the relationship between these fully observed covariates and the survey outcome of interest. Current methods for handling nonresponse in surveys have tended to focus on a subset of these features, but all three are important.

[1] The Ohio State University College of Public Health Division of Biostatistics, 242 Cunz Hall, 1841 Neil Ave., Columbus, OH 43210, U.S.A. Email: andridge.1@osu.edu
[2] University of Michigan Department of Biostatistics, M4071 SPH II, 1415 Washington Heights, Ann Arbor, MI 48109, U.S.A. Email: rlittle@umich.edu

In addition, missing-data methods have usually assumed that the missing data are missing at random (MAR, see Rubin 1976), which means that missingness does not depend on missing values in the data, after conditioning on the variables observed for respondents and nonrespondents. This assumption is often questionable, particularly in the case of unit nonresponse, where the set of variables observed for nonrespondents and respondents is limited. Data that are not MAR are called missing not at random (MNAR) or nonignorable. We propose measures to assess the impact of nonresponse for binary survey variables, when data are potentially MNAR.

A limited set of methods have been developed for MNAR nonresponse for categorical outcomes in survey data. Stasny (1991) used a MNAR hierarchical Bayes selection model to study victimization in the National Crime Survey. This work was extended by Nandram and Choi (2002a, 2002b). Similar methods are developed for multinomial outcomes in Nandram et al. (2002) and Nandram et al. (2005). Previous methods for categorical nonresponse have tended to require that auxiliary data are also categorical, but our methods allow auxiliary variables to be continuous, that is, do not require that continuous variables be categorized before inclusion in the model.

Outside of survey applications there have been more recent developments in MNAR nonresponse for binary outcomes. Magder (2003) formulated departures from MAR (towards MNAR) in terms of a response probability ratio, defined as the ratio of the probability of the outcome being observed ("response probability") comparing subjects with and without the outcome. In a similar vein, Hedeker et al. (2007) and Jackson et al. (2014) formulate departures from MAR in terms of odds ratios comparing the probability of success for subjects with and without missing outcomes, and Higgins et al. (2008) used this approach (and referred to it as "informative missingness odds ratios") for meta-analyses. More recently, Liublinska and Rubin (2014) used tipping point displays to perform sensitivity analyses for binary outcomes in two-arm clinical trials.

Little and Rubin (2019) describes five different strategies for handling MNAR problems:

1. Follow up a sample of nonrespondents and incorporate this information into the main analysis.
2. Adopt a Bayesian approach, assigning the parameters prior distributions. Bayesian inference does not generally require that the data provide information for all the parameters, although inferences tend to be sensitive to the choice of prior distribution.
3. Impose additional restrictions on model parameters, such as on the regression coefficients in regressions for missingness or the survey variables.
4. Conduct analysis to assess sensitivity of inferences for quantities of interest to different choices of the values of parameters poorly estimated from the data.
5. Selectively discard data to avoid modeling the missingness mechanism.

Our methods apply approaches (2) and (4) to handle MNAR missingness in binary survey outcomes. The work is motivated by missing income data in the 2015 OMAS (RTI International 2015), though it has broad applicability to missing binary outcomes in survey data. For example, a recent analysis used survey data from the 2016 Behavioral Risk Factor Surveillance System (BRFSS) to estimate the proportion of adults who served as caregivers in the United States, but approximately 10% of survey respondents did not answer the

question about caregiving (Barnhart et al. 2019). The resulting analysis assumed missingness was completely at random; our methods would allow investigation into how estimates would change under varying missingness assumptions, including if caregivers were less likely to respond to the question (MNAR missingness). Other possible applications include estimation of the prevalence of socially undesirable or risky behaviors, as these often are assumed to be MNAR (i.e., people who engage in the behavior are more likely to have missing data on these behaviors). For example, estimating the prevalence of intimate partner violence among pregnant women using survey data from the Pregnancy Risk Assessment Monitoring System (PRAMS), or estimating opioid drug use using survey data from the National Survey on Drug Use and Health (NSDUH).

The 2015 OMAS has been conducted seven times since 1998, and is one of the largest state-sponsored health surveys in the United States. The sampling design was a stratified (by county/sub-county), dual-frame (cell phone and landline) random digit dialing sample of Ohio's non-institutionalized population, with oversampling of certain minority populations. Clusters were defined as a household/ family, and within each cluster one adult was randomly selected to participate. Details on the design and implementation of the 2015 OMAS are available elsewhere (RTI International 2015).

A total of $n = 42{,}876$ adults provided responses to some or all of the survey, but 22.2% of subjects ($n = 9{,}511$) were missing the categorical income variable. Income is particularly important for OMAS, as the primary focus of the survey is to assess the health and health care utilization of Ohio's Medicaid, Medicaid-eligible, and non-Medicaid populations, and Medicaid eligibility is partially determined by income. Having a measure of income was even more important for the 2015 OMAS, since this was the first administration of OMAS after Ohio expanded Medicaid coverage in January 2014. Post-expansion, Medicaid was available for all adults aged 19 through 64, regardless of parental status, with family incomes at or below 138% of the Federal Poverty Level (FPL). Thus, for the 22.2% of survey participants with missing income data, eligibility for Medicaid based on income under the new criteria could not be established.

We extend to categorical outcomes our previously-described proxy pattern-mixture (PPM) analysis for a continuous outcome (Andridge and Little 2011), in order to estimate the potential impact of MNAR nonresponse on the proportion of people at various income levels estimated by 2015 OMAS data. In Section 2 we describe the binary PPM model and in Section 3 we discuss three estimation approaches: maximum likelihood (ML), Bayes, and multiple imputation (MI). In Section 4 we describe the sensitivity of each method to model misspecification and propose modifications to produce more robust estimates. The method is illustrated through simulation in Section 5 and then applied to the 2015 OMAS data in Section 6. Section 7 presents some concluding remarks.

2. Proxy Pattern-Mixture Model for a Binary Outcome

Two main approaches for formulating MNAR models can be distinguished. For unit i in the survey, let Y_i be a survey variable subject to missing data, M_i a missing data indicator with value 1 if Y_i is missing and 0 if Y_i is observed, and Z_i known variables. Assuming independent units, selection models factor the joint distribution of M_i and Y_i as

$$f(M_i, Y_i|Z_i, \theta, \psi) = f(Y_i|Z_i, \theta)f(M_i, |Z_i, Y_i, \psi), \quad (1)$$

where densities are distinguished by their arguments. The first factor characterizes the distribution of Y_i in the population, the second factor models the missingness mechanism, and θ and ψ are parameters. Alternatively, pattern-mixture models factor the joint distribution as

$$f(M_i, Y_i|Z_i, \xi, \omega) = f(Y_i|Z_i, M_i, \xi)f(M_i, |Z_i, \omega), \tag{2}$$

where the first distribution characterizes the distribution of Y_i given Z_i in the strata defined by different patterns of missingness, M_i; the second distribution models the probabilities of the different patterns (Rubin 1977; Little 1993), and ξ and ω are parameters. The selection model formulation was used to characterize MAR in Rubin (1976), and early MNAR models were based on this factorization; see, for example Heckman (1976). However, the selection Equation (1) requires specifying the missingness model via the density $f(M_i|Z_i, Y_i, \psi)$. In contrast, pattern-mixture Equation (2), which we use as the basis of our proposed method, avoids the need for an explicit parametric model for the missingness mechanism. For more discussion comparing these modeling approaches, see, for example Little and Rubin (2019) or Muthen et al. (2011).

Andridge and Little (2011) first introduced the proxy pattern-mixture (PPM) model for assessing the potential for nonresponse bias for continuous outcomes. Their model was based on an assumption of multivariate normality, which is not suitable for categorical survey outcomes. We now extend the normal proxy pattern-mixture model to a binary outcome Y using a latent variable framework. Consider, initially, a simple random sample of size n from an infinite population. Let Y_i, denote the value of a binary survey outcome and $Z_i = (Z_{i1}, Z_{i2}, \ldots, Z_{ip})$ the values of p covariates for unit i in the sample. Only r of the n sampled units provide a response for Y, so observed data consist of (Y_i, Z_i) for $i = 1, \ldots, r$ and Z_i for $i = r + 1, \ldots, n$. This data pattern could be seen with item nonresponse, as in our 2015 OMAS application, where Y is the indicator for being below 138% FPL (missing for 22% of respondents) and Z is a rich set of variables collected in the remainder of the survey. This data structure could also occur with unit nonresponse, where Y is a single survey item and the covariates Z are design variables known for the entire sample, paradata (e.g., interviewer observations), or data available through linkage with administrative registers. Our main interest is assessing the impact of nonresponse on inference for the proportion of units in the population with $Y = 1$.

In order to use the framework of Andridge and Little (2011), we assume that Y is related to a normally distributed latent variable U through the rule that $Y = 1$ when the latent variable $U > 0$. Following the approach of previous PPM models (Andridge and Little 2011; Andridge and Thompson 2015), we reduce the covariates Z to a single proxy variable X that is a linear combination of the Z using a regression model for $Y|Z$. Letting M denote the missingness indicator (i.e., $M = 1$ when Y is missing), we take the regression of Y on Z for the respondents to be a probit regression model given by

$$\Pr(Y = 1|Z, M = 0) = \Phi(\alpha_0 + \alpha Z) \tag{3}$$

where $\Phi(\cdot)$ denotes the standard normal CDF. We then create the proxy X as the linear predictor from the probit regression, $X = \hat{a}_0 + \hat{a}Z$. In doing this, covariates that are not associated with Y (among respondents) will have $\hat{a} \approx 0$ and thus will not contribute to the proxy variable X. Though the $\hat{a}$ are estimated using respondent data only, the proxy X can be

created for nonrespondents as well, since Z is fully observed. The regression coefficients $\{\hat{a}_0, \hat{a}\}$ are subject to sampling error, so in practice X is estimated rather than known.

We now apply the normal PPM model of Andridge and Little (2011) to X, U and M. Specifically, we assume that the joint distribution of $[U, X, M]$ follows the bivariate pattern-mixture model discussed in Little (1994):

$$\begin{aligned} (U, X|M = m) &\sim N_2\left((\mu_u^{(m)}, \mu_x^{(m)}), \Sigma^{(m)}\right) \\ M &\sim Bernoulli(1 - \pi) \\ \Sigma^{(m)} &= \begin{bmatrix} \sigma_{uu}^{(m)} & \rho^{(m)}\sqrt{\sigma_{uu}^{(m)}\sigma_{xx}^{(m)}} \\ \rho^{(m)}\sqrt{\sigma_{uu}^{(m)}\sigma_{xx}^{(m)}} & \sigma_{uu}^{(m)} \end{bmatrix}, \end{aligned} \tag{4}$$

where N_2 denotes the bivariate normal distribution. In this pattern-mixture model, the parameters of the joint distribution of U and X are allowed to differ across patterns, as indicated by the superscript (m).

Of primary interest is the marginal mean of Y, which can be obtained from the pattern-mixturemodel by averaging over the two patterns ($m = 0, 1$). Specifically, we want to estimate

$$\begin{aligned} \mu_y = \Pr(Y = 1) &= \Pr(U > 0) \\ &= \Pr(U > 0|M = 0) \times \Pr(M = 0) + \Pr(U > 0|M = 1) \times \Pr(M = 1) \\ &= \pi\Phi\left(\mu_u^{(0)}/\sqrt{\sigma_{uu}^{(0)}}\right) + (1 - \pi)\Phi\left(\mu_u^{(1)}/\sqrt{\sigma_{uu}^{(1)}}\right). \end{aligned} \tag{5}$$

For parameters of the respondent distribution ($m = 0$), estimates are easily obtained, though since U is a latent variable we cannot estimate both its mean and variance. Without loss of generality we fix the variance at $\sigma_{uu}^{(0)} = 1$ as is conventional for latent variables and estimate the mean $\mu_u^{(0)}$ (for details on estimation methods, see Section 3). For nonrespondents, we can estimate the mean and variance of X, $\{\mu_x^{(1)}, \sigma_{xx}^{(1)}\}$, but there is no information in the data with which to estimate the parameters $\mu_u^{(1)}$, $\sigma_{uu}^{(1)}$, and $\rho^{(1)}$. Thus the model is underidentified.

Importantly, note that $\rho^{(m)}$ is the correlation between X and the latent variable U, not the correlation between X and Y. This correlation between latent U and X is referred to as the biserial correlation between binary Y and continuous X, and is always larger than the Pearson correlation between Y and X. This correlation is defined for both respondents ($\rho^{(0)}$) and nonrespondents ($\rho^{(1)}$), but it is only estimable (without further assumptions) for respondents. We refer to this correlation as the "strength of the proxy", as higher $\rho^{(0)}$ indicates a stronger model for Y, that is, that the covariates Z are more predictive of Y in the probit model for respondents.

In order to identify the model and obtain estimates for the parameters $\mu_u^{(1)}$, $\sigma_{uu}^{(1)}$, and $\rho^{(1)}$ we use parameter restrictions induced by assumptions on the missing data mechanism (Little 1994; Andridge and Little 2011). Specifically, we assume that the probability that $M = 1$ is an unspecified function f of a linear combination of X and U:

$$\Pr(M = 1|U, X) = f((1 - \phi)X^* + \phi U), \tag{6}$$

where $X^* = X/\sqrt{\sigma_{xx}^{(0)}}$ is the proxy rescaled to have unit variance.

Here ϕ is a sensitivity parameter that ranges from 0 to 1 and determines the missingness mechanism. A value of $\phi = 0$ corresponds to missingness being entirely dependent on the proxy X, which is fully observed, implying missingness is at random (MAR). At the other extreme, $\phi = 1$ corresponds to missingness being entirely dependent on U, the completely unobserved latent variable, and thus a missing not at random (MNAR) mechanism. Intermediate values of ϕ allow for "less extreme" MNAR mechanisms, for example, $\phi = 0.5$ which would equally weight the contributions of X and U. For MNAR mechanisms, missingness is a function of the latent variable U, thus allowing for a "smooth" missingness function. In other words, conditional on X the probability of missingness may lie on a continuum instead of only taking two values, as would be the case if missingness depended on Y itself. We note that in previous PPM models the sensitivity parameter was defined differently (as $\lambda \in [0, \infty]$ in Andridge and Little 2011; Andridge and Thompson 2015; Andridge et al. 2017); we elect to use the more recent reparameterization of Little et al. (2019). We note that the sensitivity analyses in the earlier papers using $\lambda = \{0, 1, \infty\}$ correspond to $\phi = \{0, 0.5, 1\}$.

With the missingness assumption in Equation (6), the unidentified parameters can be expressed as functions of the identified parameters and the sensitivity parameter ϕ. The mean and variance of U for nonrespondents (given $M = 1$) are given by

$$\begin{aligned} \mu_u^{(1)} &= \mu_u^{(0)} + \left(\frac{\phi + (1-\phi)\rho^{(0)}}{\phi\rho^{(0)} + (1-\phi)}\right)\frac{\mu_x^{(1)} - \mu_x^{(0)}}{\sqrt{\sigma_{xx}^{(0)}}} \\ \sigma_{uu}^{(1)} &= 1 + \left(\frac{\phi + (1-\phi)\rho^{(0)}}{\phi\rho^{(0)} + (1-\phi)}\right)^2\frac{\sigma_{xx}^{(1)} - \sigma_{xx}^{(0)}}{\sigma_{xx}^{(0)}}. \end{aligned} \tag{7}$$

These formulae follow from the corresponding formulae for the normal proxy pattern-mixture model, with U in place of Y; see Andridge and Little (2011) and the appendix of Sullivan and Andridge (2015) for details on their derivation.

Looking closely at these formulas, we see that the mean and variance for nonrespondents ($\mu_u^{(1)}$, $\sigma_{uu}^{(1)}$) are shifted from the mean and variance for respondents ($\mu_u^{(0)}$,$\sigma_{uu}^{(0)} = 1$), with the amount of shift based on the sensitivity parameter ϕ, the correlation between U and X for respondents ($\rho^{(0)}$), and the differences between the respondent and nonrespondent distributions of X. Larger shifts correspond to larger nonresponse bias for the respondent mean of Y, that is, the overall mean is further from the respondent mean. The size of the bias is also determined by the response rate π, since as shown in Equation (5) the overall mean is a weighted combination of the respondent and nonrespondent means.

A few observations can be made about the formulae in Equation (7). Regardless of ϕ or $\rho^{(0)}$ or π, if the mean and variance of X are the same for respondents and nonrespondents (i.e., $\mu_x^{(0)} = \mu_x^{(1)}$ and $\sigma_{xx}^{(0)} = \sigma_{xx}^{(1)}$), then the mean and variance of U are identical for respondents and nonrespondents. Thus the marginal mean of Y in Equation (5) equals the respondent mean of Y. This property seems reasonable, as if there is no evidence of nonresponse bias in X, which is a proxy for Y, then there is no evidence of nonresponse bias in Y. If there is a difference between respondents and nonrespondents in means or variances of X, then the difference in the respondent and nonrespondent means and

variances of U are functions of the differences for X, the strength of the proxy as measured by $\rho^{(0)}$, and the sensitivity parameter ϕ. The higher the correlation $\rho^{(0)}$, the closer the shift in means and variances for U is to the shift for X, regardless of ϕ. This makes sense, as in the extreme, if X is a perfect proxy for U, that is, $\rho^{(0)} = 1$, then the shift in the mean and variance of U should be exactly equal to the (standardized) shift for X (i.e., the nonresponse bias for U is equal to the nonresponse bias for X). To assess the effect of the sensitivity parameter ϕ, note that as ϕ goes from 0 to 1, we move from an MAR assumption ($\phi = 0$) towards an increasingly strong MNAR assumption. For a given difference in means and variances for X and correlation $\rho^{(0)}$, moving further away from MAR (increasing ϕ towards 1) leads to larger differences between respondent and nonrespondent means of U. This, in turn, leads to larger differences between the respondent mean and the overall mean, that is, a larger nonresponse bias for the overall mean. To help visualize these effects, Subsection 5.1 shows an illustration of how these parameters impact nonresponse bias under this model.

For a given partially missing survey outcome Y and covariates Z, all components of the PPM model will be identified under the missingness assumption in Equation (6) with a specified value of ϕ. Thus, to use our model to assess the sensitivity of the mean of Y to varying assumptions about the missingness mechanism, we specify values of the sensitivity parameter ϕ and estimate the marginal mean of Y for each ϕ value. In order to capture a range of nonignorability, we suggest using the values $\phi = \{0, 0.5, 1\}$, which correspond to MAR ($\phi = 0$), the most "extreme" MNAR ($\phi = 1$) where missingness depends entirely on unobserved U, and an intermediate value ($\phi = 0.5$) that lies between these two extremes. How much or little the estimates vary, along with the size of confidence intervals for the mean, can provide insight into the potential magnitude of nonresponse bias.

3. Estimation Methods

3.1. Maximum Likelihood

Maximum likelihood (ML) estimates for the pattern-mixture model are obtained by substituting estimates for the parameters into Equations (5) and (7). Estimators of $\mu_x^{(m)}$ and $\sigma_{xx}^{(m)}$ for $m = 0,\ 1$ are the usual sample means and covariance matrices for X for respondents and nonrespondents. The estimate for π is the response rate, that his, the proportion of cases with Y observed. ML estimates of the biserial correlation $\rho^{(0)}$ and $\mu_u^{(0)}$ (referred to as the "cutpoint") do not have closed form (Tate 1955a, 1995b) and require an iterative algorithm such as Newton-Raphson. Substituting these estimates into Equations (5) and (7) yields the ML estimate of the mean of Y. The large-sample variance estimate of μ_y is obtained through Taylor series expansion and inversion of the information matrix.

Though quite easy to compute, ML estimation for the PPM model has several drawbacks. First, the resulting ML estimate of $\mu_u^{(0)}$ is not the inverse probit of the respondent mean of Y, and thus the ML estimate of the respondent mean of Y is not $\overline{y}_R$. We propose an alternative approach that avoids this consequence in Subsection 4.1. Additionally, the proxy X is constructed using estimated coefficients from the probit regression of Y on Z, but the ML estimate treats these coefficients as known. A more principled approach is to incorporate this uncertainty with a Bayesian approach, described

below. Finally, incorporation of survey design features, such as weights, in the ML estimates is not straightforward, leading us to consider a multiple imputation approach that enables easily application of design-based inference.

3.2. Bayesian Inference

In order to obtain Bayesian estimates for the parameters of the PPM model, we place non-informative priors on the probit regression parameters α (which includes the intercept) and use a Gibbs sampler to draw the latent U for respondents (Albert and Chib 1993). At the jth iteration of the Gibbs sampler, U follows a truncated normal distribution conditional on Y and $\alpha_{(j)}$ (and the proxy $X_{(j)} = \alpha_{(j)}Z$),

$$(U_{(j)}|Y, \alpha_{(j)}, M = 0) = (U_{(j)}|Y, X_{(j)}, M = 0) \sim N(X_{(j)}, 1) = N(\alpha_{(j)}Z, 1)$$

$$\text{truncated at the left by 0 if } Y = 1 \quad (8)$$

$$\text{truncated at the right by 0 if } Y = 0,$$

where the subscript (j) denotes the jth draws of the parameters. Note that the truncation arises because $Y = 1$ if $U > 0$ and $Y = 0$ otherwise. Then, given the augmented continuous $U_{(j)}$ we draw $\alpha_{(j+1)}$ from its posterior distribution, which also follows a normal distribution,

$$(\alpha_{(j+1)}|Y, U_{(j)}, M = 0) \sim N((Z^TZ)^{-1}Z^TU_{(j)}, (Z^TZ)^{-1}), \quad (9)$$

and recreate the proxy $X_{(j+1)} = \alpha_{(j+1)}\,Z$.

The data augmentation of Y to U allows us to exploit the straightforward Bayesian estimation methods for the normal PPM model. For a chosen value of ϕ, we apply the PPM model algorithm as described in Andridge and Little (2011) to the pair (X, U) to obtain draws of the parameters of the joint distribution of X and U, $\{\mu_x^{(0)}, \sigma_{xx}^{(0)}, \mu_x^{(1)}, \sigma_{xx}^{(1)}, \mu_u^{(0)}, \rho^{(0)}\}$. Since U is unobserved even for the respondents, after each draw of the parameters from the PPM model, X is recreated for the entire sample and U is redrawn for the respondents given the current set of parameter values as described in the data augmentation approach above. Note that this does not require a draw of the latent data for nonrespondents. Draws from the posterior distribution of μ_y are obtained by transforming the draws from the Gibbs sampler as in Equation (7) to obtain $\{\mu_u^{(1)}, \sigma_{uu}^{(1)}\}$, and then substituting these draws into Equation (5).

3.3. Multiple Imputation

An alternative method of inference is multiple imputation (MI) (Rubin 1978). For a selected ϕ we create K complete data sets by filling in missing (binary) Y values with draws from the posterior distribution based on the pattern-mixture model. At the jth draw of the model parameters from their posterior distribution as described in Subsection 3.2 we draw the latent U for nonrespondents based on the conditional distribution,

$$\left[u_{i(j)}|x_{i(j)}, m_i = 1, \phi_{(j)}\right] \sim N\left(\mu_{u(j)}^{(1)} + \frac{\sigma_{ux(j)}^{(1)}}{\sigma_{xx(j)}^{(1)}}\left(x_{i(j)} - \mu_{x(j)}^{(1)}\right), \sigma_{uu(j)}^{(1)} - \frac{\sigma_{ux(j)}^{(1)\,2}}{\sigma_{xx(j)}^{(1)}}\right). \quad (10)$$

The missing y_i are then imputed as $y_{i(j)} = I(u_{i(j)} > 0)$, where $I()$ is an indicator function taking the value 1 if the expression is true. In order to reduce auto-correlation between the imputations due to the Gibbs sampler, we thin the chain for the purpose of creating the imputations. For the kth completed data set the estimate of μ_y is the imputed sample mean $\bar{Y}_k$ with estimated variance W_k. A consistent estimate of μ_y is then given by $\hat{\mu}_y = \frac{1}{K}\sum_{k=1}^{K} \bar{Y}_k$ with $\text{Var}(\hat{\mu}_y) = \bar{W}_K + \frac{K+1}{K} B_K$, where $\bar{W}_K = \frac{1}{K}\sum_{k=1}^{K} W_k$ is the within-imputation variance and $B_K = \frac{1}{K-1}\sum_{k=1}^{K}\left(\bar{Y}_k - \hat{\mu}_y\right)^2$ is the between-imputation variance.

As with the normal PPM analysis, an advantage of the MI approach is that complex design features like clustering, stratification and unequal sampling probabilities can be readily incorporated in the complete-data component of the MI combining rules. Once the imputation process has created complete data sets, design-based methods can be used to estimate μ_y and its variance; for example the design-weighted Horvitz-Thompson estimator can be used to calculate $\bar{Y}_k$. It has been shown that the multiple imputation variance estimator can be biased when applied to data from a complex sample survey unless the sample weights are included as a predictor in the imputation (Kim et al. 2006), and thus we include the sample weights as part of Z when performing imputation with the binary PPM model. For stratified designs, strata can be included as a set of indicator variables in Z.

4. Reducing Sensitivity to Violations of Normality

The normal PPM model is relatively robust to departures from normality, since it relies on first and second moments in estimating the mean of Y. However, for binary outcomes the normality of X plays a more crucial role, as can be seen in the following example. Under the latent variable framework, the conditional distribution of U given X for respondents is normal by definition. Thus, if X has a normal distribution, the marginal distribution of U is also normal, and $\mu_y^{(0)} = \Pr(U > 0|M = 0) = \Phi(\mu_u^{(0)})$. Substituting the ML estimate for $\mu_u^{(0)}$ into the expression for $\mu_y^{(0)}$yields an unbiased estimate of the mean of Y. Suppose, however, that X does not have a normal distribution among the respondents. The conditional distribution $[U\,|X, M = 0]$ remains normal (by definition), but the marginal distribution $[U\,|M = 0]$ is no longer normal. As a consequence, the ML estimate of $\mu_u^{(0)}$ is unbiased, but the ML estimate of $\mu_y^{(0)}$ obtained as $\hat{\mu}_y^{(0)} = \Phi(\hat{\mu}_u^{(0)})$ is biased. The transformation from U to Y gives $\Pr(Y = 1|M = 0) = \Pr(U > 0|M = 0) = \int_0^\infty f_U(u)du$ where $f_U^{(0)}(u)$ is the convolution of the normal distribution of $[U\,|X, M = 0]$ and the non-normal distribution of $[X|M = 0]$. Thus, the ML estimate of the respondent mean of Y is biased, despite the fact that Y is fully observed for the respondents. The Bayesian approach produces biased estimates of the respondent mean for the same reason; the transformation of draws of $\mu_u^{(0)}$ to draws of $\mu_y^{(0)}$ produces biased results.

Both the ML and Bayesian estimation approaches will produce increasingly biased estimates of $\mu_y^{(0)}$ the further X deviates from normality. MI is less sensitive to departures from normality since imputations are based on the conditional distribution $[U|X, M]$ which is normal by definition of the latent variable and is not affected by non-normal X. In the next section, we describe modifications to the ML and Bayesian methods that are robust against deviations from normality of the proxy X.

4.1. Robust Estimation Methods

To reduce sensitivity to non-normal X we use the *two-step method* proposed by Olsson et al. (1982) to estimate the biserial correlation coefficient, $\rho^{(0)}$. In the first step, the cutpoint$\mu_u^{(0)}$ is estimated by $\hat{\mu}_u^{(0)} = \Phi^{-1}(\bar{y}_R)$, so that the estimate of the respondent mean of Y is $\bar{y}_R$. Then a conditional ML estimate of $\rho^{(0)}$ is computed, given $\hat{\mu}_u^{(0)}$. ML estimates of the other parameters of the PPM model are computed as before. This method is computationally simpler than ML, has the attractive property of returning the natural estimate $\hat{\mu}_y^{(0)} = \bar{y}_R$, and is less sensitive to non-normality of the proxy. We estimate the variance using the bootstrap. We refer to this method as Modified ML (MML).

We modify the Bayesian method to reduce sensitivity to non-normal X by applying the Gibbs sampler to draw the latent U for nonrespondents conditional on X and the current parameter values at each iteration, which is not required for the standard Bayesian approach (but is part of the MI approach, see Equation (10)). Draws of the nonrespondent mean of Y, $\mu_y^{(1)}$, are then taken to be $\mu_y^{(1)} = \frac{1}{n-r}\sum_{i=r+1}^{n} I(U_i > 0)$, instead of transformations of the parameters $\{\mu_x^{(0)}, \sigma_{xx}^{(0)}, \mu_x^{(1)}, \sigma_{xx}^{(1)}, \mu_u^{(0)}, \sigma_{uu}^{(0)}, \rho^{(0)}\}$ as given in Equations (7) and (5). A similar modification for the respondent mean is not appropriate, as draws of U for the respondents in the Gibbs sampler are conditional on the observed Y (as in Equation (8)) and thus the resulting draw of $\mu_u^{(0)}$ will always be $\bar{y}_R$. To avoid this, we propose two approaches. An obvious extension is to draw U *without* conditioning on Y, and condition only on the current draws of the proxy X and the parameters. With the subsequent draw of $\mu_y^{(0)}$ taken to be $\mu_y^{(0)} = \frac{1}{n-r}\sum_{i=r+1}^{n} I(U_i > 0)$, the draws of $\mu_y^{(0)}$ will not always be $\bar{y}_R$. To distinguish this method from the unmodified posterior distribution intervals (PD), we call this method Modification 1: PD-redraw. The drawback of this method is that variances may actually be *overestimated* since we are essentially imputing the observed binary outcome Y for the respondents. An alternative approach is to not use the latent U, but instead directly use the draws of the proxy X to calculate predicted probabilities from the probit model. The average of these predicted probabilities for the respondents can then be taken as a draw of the respondent mean, that is, $\mu_y^{(0)} = \frac{1}{r}\sum_{i=1}^{r} \Phi^{-1}(X_i)$. This is actually a draw of the conditional mean of Y (conditional on X) and so its posterior distribution will underestimate the variance of $\mu_y^{(0)}$. To combat this, we take a simple random sample with replacement of the X_i before calculating the mean of the predicted probabilities (Modification 2: PD-predprob).

5. Simulation Studies

We now describe a set of simulation studies designed to (1) illustrate the effects of proxy strength $\rho^{(0)}$, differences in the mean of X between respondents and nonrespondents, and sample size on PPM model estimates of the mean of a binary outcome Y, (2) assess confidence coverage of ML, Bayes and MI inferences when model assumptions are met, and (3) assess confidence coverage of the robust estimation methods when the normality assumption is incorrect. All simulations and data analysis were performed using the software package R (R Core Team 2017). Code for implementing the PPM model and example analyses are available at http://github.com/randridge/PPMA.

5.1. Numerical Illustration of Binary PPM Analysis

We first illustrate the taxonomy of evidence concerning bias based on the strength of the proxy ($\rho^{(0)}$) and the standardized difference between the mean of the proxy for the entire sample and the mean for respondents, $\delta^* = \left(\mu_x - \mu_x^{(0)}\right)/\sqrt{\sigma_{xx}^{(0)}}$. We created a total of eighteen artificial data sets in a 3x3x2 factorial design with a fixed nonresponse rate of 50%. A single data set was generated for each combination of $\rho^{(0)} = \{0.8, 0.5, 0.2\}$, $\delta^* = \{0.1, 0.3, 0.5\}$, and $n = \{100, 400\}$ (with corresponding $r = \{50, 200\}$). A single covariate Z was generated for both respondents and non-respondents, with $z_i \sim N(0, 1)$, $i = 1, \ldots, r$ for respondents and $z_i \sim N\ (\delta^*/(1-r/n),\ 1),\ i = r+1, \ldots, n$ for nonrespondents. For respondents only, a latent variable u_i was generated as $[u_i|z_i] \sim N(a_0 + a_1 z_i,\ 1)$, with an observed binary Y then created as $y_i = 1$ if $u_i > 0$. We set $a_1 = \rho^{(0)}/\sqrt{1-\rho^{(0)^2}}$ so that $\text{Corr}(U,X|M=0) = \rho^{(0)}$ where the proxy $X = a_0 + a_1 Z$. We chose $a_0 = \Phi^{-1}(0.3)\sqrt{1+a_1^2}$ so that the expected value of Y for respondents was 0.3. In this and all subsequent simulations, the latent variable U was used for data generation and then discarded; only Y and Z were used for the proxy pattern-mixture analysis.

For each of the eighteen data sets, estimates of the mean of Y and its variance were obtained using the PPM model for $\phi = (0, 0.5, 1)$. For each value of ϕ, three 95% intervals were calculated:

1. ML: the ML estimate ± 2 standard errors (large-sample approximation),
2. PD: the posterior median and 2.5th to 97.5th posterior interval based on 2000 cycles of the Gibbs sampler as outlined in Subsection 3.2, with a burn-in of 20 iterations,
3. MI: mean ± 2 standard errors from 50 multiply imputed data sets, with a burn-in of 20 iterations and imputing on every hundredth iteration of the Gibbs sampler.

The robust estimation methods described in Subsection 4.1 designed to handle non-normal proxies were also calculated. Since the simulated covariate data were normally distributed, the modified estimators yield similar results and are not shown. The complete case estimate, that is, the respondent mean, was also computed (± 2 standard errors) for each data set.

5.1.1. Results

Figure 1 shows the resulting 95% intervals using each of the three estimation methods for the nine data sets with n=400, plotted alongside the respondent mean (complete case). The relative performances of each method for the data sets with $n = 100$ are similar to the results with $n = 400$ (with larger interval lengths); results are not shown. We note that in this simulation the true mean of Y is not known (indeed, we did not simulate Y for nonrespondents); we simply illustrate the effect of various values of proxy strength ($\rho^{(0)}$) and proxy mean deviation (δ^*) on the sensitivity analysis and compare the different estimation methods.

For populations with strong proxies ($\rho^{(0)} = 0.8$), ML, PD, and MI give nearly identical results. For these populations, there is not a noticeable increase in the length of the intervals as we move from $\phi = 0$ to $\phi = 1$, suggesting that even in the case of a large deviation ($\delta^* = 0.5$) there is good information to correct the potential bias.

For weaker proxies, we see differences among the three methods. When $\phi = 0$ (MAR) the three methods yield similar inference, but for MNAR mechanisms the intervals for PD

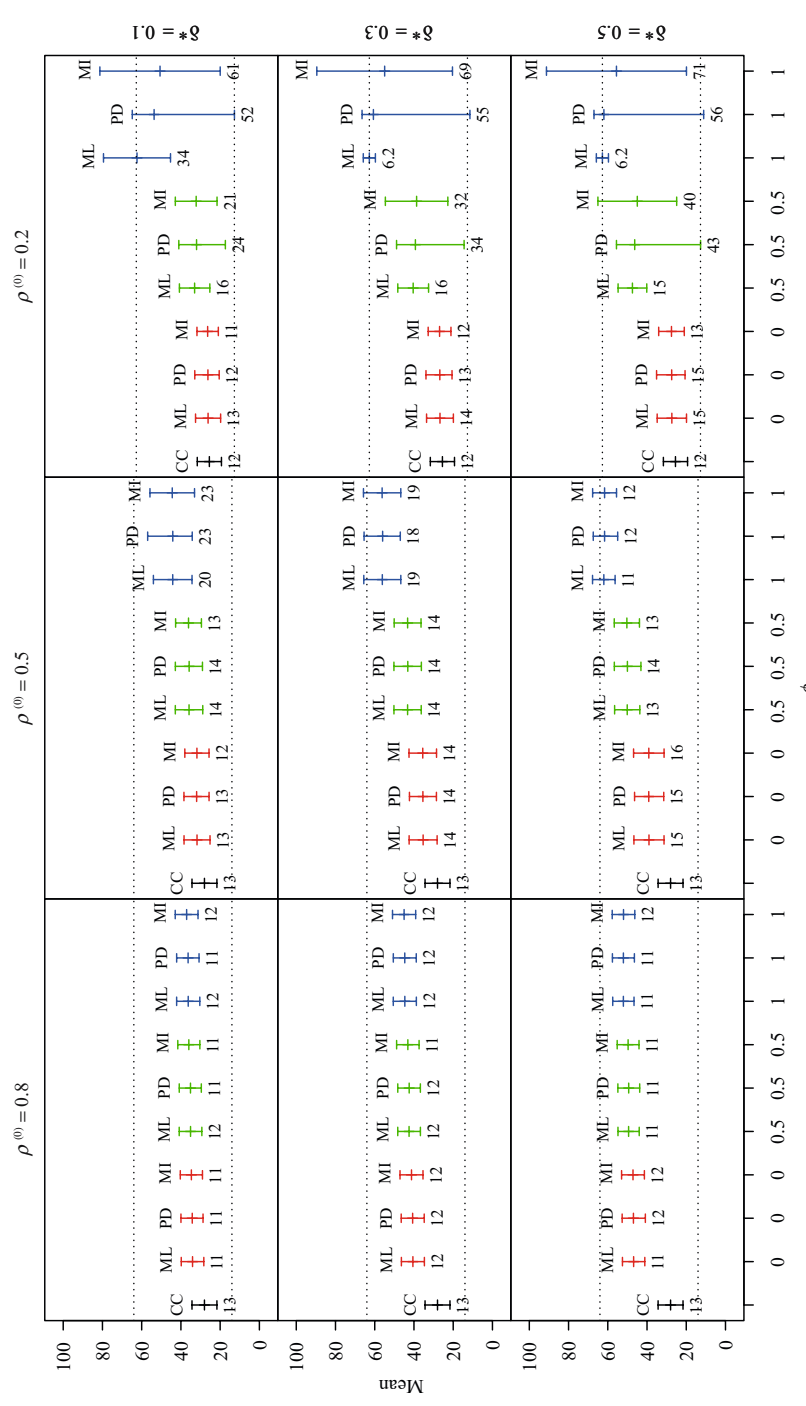

Fig. 1. 95% intervals for nine generated data sets ($n = 400$) for $\phi = (0, 0.5, 1)$. Numbers below intervals are the interval length. Dotted lines are the point estimates obtained by filling in all ones or all zeros for missing values. CC: Complete cases; ML: Maximum likelihood; PD: Posterior distribution; MI: 50 multiply imputed data sets.

and MI tend to be wider than those for ML. For both Bayesian methods (PD, MI) the interval width increases as we move from $\phi = 0$ to $\phi = 1$, with a marked increase in length when $\rho^{(0)} = 0.2$. The ML inference displays different behavior; its intervals actually get very small for the weak proxies and large d^*. This is caused by the unstable behavior of the ML estimate near the boundary of the parameter space. For weak proxies (small $\rho^{(0)}$), the ML estimate of $\sigma^{(1)}_{uu}$ can be zero if the nonrespondent proxy variance is smaller than the respondent variance (see Equation (7)). Since the ML estimate of the mean of Y is given by $\hat{\mu}^{(1)}_y = \Phi\left(\hat{\mu}^{(1)}_u / \sqrt{\hat{\sigma}^{(1)}_{uu}}\right)$, a zero value for $\hat{\sigma}^{(1)}_{uu}$ causes $\hat{\mu}^{(1)}_y$ to be exactly 0 or 1 depending on the sign of $\hat{\mu}^{(1)}_u$. The large sample variance will then be small since the estimate of $\sigma^{(1)}_{uu}$ is zero, and interval widths will be small relative to the PD or MI intervals.

Since the outcome is binary, we obtain a natural upper and lower bound for the mean of Y by filling in all missing values with zeros or all with ones. These bounds are sometimes referred to as Manski bounds (Manski 2016) and are shown with dotted lines in Figure 1. For strong proxies, even with a large deviation this upper bound is not reached, suggesting that the overall mean would not be this extreme even in the worst-case MNAR scenario, where missingness depends entirely on Y through U. However, for the weakest proxy ($\rho^{(0)} = 0.2$) we see that even for the smallest deviation ($\delta^* = 0.1$) the intervals for PD and MI cover or nearly cover these bounds. This is due to the weak information about Y contained in the proxy. The PD intervals are highly skewed and the MI intervals are exaggerated in length. The posterior distribution of μ_y is bimodal, with modes at each of the two bounds obtained when missing values are all zeros or all ones. Thus the posterior interval essentially covers the entire range of possible values of μ_y. Similarly, for MI the imputed data sets have imputed values that are either all zeros or all ones. This causes very large variance and thus large intervals, and since by construction the intervals are symmetric for MI, they are even larger than the posterior intervals from PD. As previously discussed, the ML method gives extremely small intervals for the weak proxies, with the point estimate at the upper bound.

5.2. *Confidence Coverage, Normally Distributed Proxy*

We now assess coverage properties for each of the three estimation methods when the PPM model is correct. We generated data using the same set-up as Subsection 5.1. We fixed $\delta^* = 0.3$ and varied $\rho^{(0)} = \{0.8, 0.5, 0.2\}$ and $n = \{100, 400\}$ for a total of six scenarios. We generated 1,000 replicate data sets for each scenario and applied the binary PPM model using each of $\phi = \{0, 0.5, 1\}$ to each one, assuming the value of ϕ is the true value. This led to a total of eighteen hypothetical populations, and for each we computed the actual coverage of a nominal 95% interval and median interval length. We also calculated the relative empirical bias for each estimator. Treating the assumed value of ϕ as correct is unrealistic, but coverages are clearly not valid when the value of ϕ is misspecified, and uncertainty in choice of ϕ is captured by the sensitivity analysis.

A total of six estimators for the mean of the binary outcome Y and its variance were obtained for each replicate data set. These included the unmodified ML, unmodified Bayesian (PD), and MI estimators, as well as the three modified estimators described in Subsection 4.1: the modified ML estimator (MML) and two modifications to the

Bayesian estimator (PD-redraw, PD-predprob). Confidence intervals for the modified ML estimator were based on 500 bootstrap samples. Posterior intervals for all three PD methods were based on 1,000 draws from the Gibbs sampler as the chains were quick to converge.

5.2.1. Results

Table 1 displays the average empirical relative bias, nominal coverage, and median CI width for the eighteen populations. For the smaller sample size ($n = 100$; Table 1a), all methods suffer from slight undercoverage, even when the proxy is strong. Undercoverage is worst in the populations with the weakest proxy ($\rho^{(0)} = 0.2$) and with $\phi = 1$, where all the methods are negatively biased. With 50% nonresponse, these small samples have only 50 observed data points, and estimation of the distribution of the latent variable is challenging. With this small sample size, confidence intervals are slightly wider for the modified ML than for unmodified ML, and thus coverage is better for the modified ML procedure. No method displays consistently better performance for the small sample size, though the larger interval lengths of PD-redraw and MI yield slightly closer to nominal coverage.

More differences between the methods emerge with the larger sample size ($n = 400$; Table 1b). All methods perform well when the proxy is strong ($\rho^{(0)} = 0.8$), though the second modification to the Bayesian method that resamples the predicted probabilities (PD-predprob) shows a small amount of undercoverage. As expected, the interval widths for the first modification to the Bayesian method (PD-redraw) are wider than the unmodified Bayesian method (PD), with PD-redraw actually overcovering for several populations, most notably when $\phi = 0.5$ or 1 and with weaker proxies. As was evident in the previous simulation, when $\rho^{(0)} = 0.2$ and $\phi = 1$ the confidence interval length for the standard ML procedure is much smaller than any of the other methods, and this leads to undercoverage. In this worst-case scenario, the modified ML, PD-redraw, and MI procedures have nominal coverage despite having some bias and have the largest median confidence interval widths.

5.3. Confidence Coverage, Non-Normally Distributed Proxy

The final objective of the simulation study was to assess the performance of the methods when the normality assumption of the proxy was violated. Complete data were generated and missing values created using a selection model with a variety of missing data mechanisms. The sample size was fixed at $n = 400$, and three different distributions for a single covariate Z were selected: (a) Normal(0, 1), (b) Gamma(4, 0.5), (c) Exponential(1). These distributions were chosen to evaluate the effect of both moderate skew (Gamma) and severe skew (Exponential). We note that the selection model implies marginal normality, whereas the PPM model assumes conditional normality, so even with a normally distributed covariate the distributional assumptions of the PPM model are violated.

Data were generated as follows. For each of the three Z distributions the covariate z_i, $i = 1,..., n$ was generated. Then, for each of $\rho = \{0.8, 0.5, 0.2\}$ the latent u_i was generated from $[u_i|z_i] \sim N(a_0 + a_1 z_i, 1)$, with $a_1 = \rho/\sqrt{1-\rho^2}$ so that $\text{Corr}(U, X) = \rho$, where the

Table 1. Average relative empirical bias, 95% interval coverage and median interval length for eighteen artificial populations with $\delta* = 0.3$ and $\rho^{(0)} = \{0.8, 0.5, 0.2\}$ for (a) $n = 100$ and (b) $n = 400$. ML: Maximum likelihood; PD: Posterior distribution (Bayesian method); PD-redraw: Modification 1 to PD; PD-predprob: Modification 2 to PD; MI: 20 multiply imputed data sets. Results over 1,000 replicates.

(a) $n = 100$

		$\rho^{(0)} = 0.8$			$\rho^{(0)} = 0.5$			$\rho^{(0)} = 0.2$		
ϕ	Method	Relative bias (%)	Coverage (%)	CI width	Relative bias (%)	Coverage (%)	CI width	Relative bias (%)	Coverage (%)	CI width
0	ML	−0.4	**92.9**	0.24	0.4	**92.7**	0.27	0.5	**93.4**	0.27
	Modified ML	−0.1	94.5	0.26	0.5	**93.2**	0.28	0.5	93.8	0.28
	PD	0.3	**93.3**	0.23	1.2	93.7	0.26	1.7	94.2	0.26
	PD-redraw	0.1	**93.4**	0.24	1.0	96.3	0.30	1.5	*97.2*	0.31
	PD-predprob	−0.2	**90.9**	0.22	0.7	**91.9**	0.25	1.2	**93.3**	0.25
	MI	−0.3	**92.2**	0.25	0.6	**93.4**	0.27	1.3	93.8	0.27
0.5	ML	−0.5	**93.6**	0.24	0.0	**92.2**	0.28	−6.7	**80.4**	0.29
	Modified ML	−0.1	94.5	0.25	0.1	**92.6**	0.28	−6.7	**81.5**	0.29
	PD	−0.3	93.7	0.24	−0.5	94.8	0.29	−9.6	95.4	0.37
	PD-redraw	−0.5	**93.0**	0.25	−0.8	*96.5*	0.33	−9.9	*97.3*	0.41
	PD-predprob	−0.6	**92.1**	0.23	−0.8	**93.1**	0.28	−9.7	95.0	0.36
	MI	−0.6	93.8	0.25	−1.2	**93.1**	0.30	−10	**93.2**	0.41
1	ML	−0.3	**93.3**	0.26	0.8	**76.2**	0.37	−18	**75.4**	0.30
	Modified ML	0.1	94.7	0.29	0.8	**93.4**	0.43	−18	**82.2**	0.65
	PD	−0.5	93.8	0.26	−4.3	95.3	0.36	−25	**81.8**	0.49
	PD-redraw	−0.7	93.9	0.27	−4.1	*96.6*	0.38	−25	**86.2**	0.52
	PD-predprob	−0.7	**92.0**	0.26	−4.1	94.7	0.36	−25	**82.4**	0.50
	MI	−0.4	**93.6**	0.28	−5.1	93.9	0.38	−27	**86.6**	0.60

Bolded values are below 1.96 simulation standard errors.
Italicized values are above 1.96 simulation standard errors.

Table 1. Continued.

(b) $n = 400$

		$\rho^{(0)} = 0.8$			$\rho^{(0)} = 0.5$			$\rho^{(0)} = 0.2$		
ϕ	Method	Relative bias (%)	Coverage (%)	CI width	Relative bias (%)	Coverage (%)	CI width	Relative bias (%)	Coverage (%)	CI width
0	ML	0.0	**93.3**	0.12	−0.1	95.4	0.14	0.0	95.7	0.14
	Modified ML	0.1	96.2	0.13	−0.1	95.3	0.14	0.0	95.7	0.14
	PD	0.1	95.3	0.12	0.1	94.6	0.13	0.3	95.1	0.13
	PD-redraw	0.0	95.4	0.13	0.0	*97.7*	0.15	0.3	*98.6*	0.16
	PD-predprob	0.0	**93.4**	0.11	−0.1	93.9	0.13	0.2	94.0	0.13
	MI	−0.1	95.1	0.13	0.0	95.3	0.14	0.2	95.4	0.14
0.5	ML	0.0	94.8	0.12	−0.1	95.1	0.14	−0.6	94.0	0.15
	Modified ML	0.1	95.4	0.13	−0.1	95.6	0.14	−0.6	94.5	0.15
	PD	0.0	95.1	0.12	−0.1	95.7	0.14	−1.3	*97.1*	0.21
	PD-redraw	−0.1	95.5	0.13	−0.2	*97.8*	0.16	−1.4	*98.5*	0.23
	PD-predprob	−0.1	94.1	0.11	−0.2	94.9	0.14	−1.3	*96.8*	0.21
	MI	−0.1	95.1	0.13	−0.3	95.8	0.15	−3.0	*97.2*	0.23
1	ML	0.0	94.4	0.13	1.2	**92.1**	0.22	−5.0	**90.7**	0.16
	Modified ML	0.1	95.3	0.13	1.3	**92.9**	0.21	−5.0	94.7	0.38
	PD	0.0	94.7	0.13	0.1	95.8	0.20	−8.3	**91.0**	0.32
	PD-redraw	−0.1	94.8	0.13	0.3	*96.7*	0.21	−8.2	93.9	0.33
	PD-predprob	−0.1	**93.1**	0.12	0.2	95.6	0.19	−8.2	**90.8**	0.32
	MI	0.0	94.9	0.13	0.4	94.6	0.21	−11	95.3	0.37

Bolded values are below 1.96 simulation standard errors.
Italicized values are above 1.96 simulation standard errors.

Table 2. Average relative empirical bias, 95% interval coverage and median interval length for eighteen artificial populations with n = 400 with covariate distributions (a) Normal, (b) Gamma, and (c) Exponential. ML: Maximum likelihood; PD: Posterior distribution (Bayesian method); PD-redraw: Modification 1 to PD; PD-predprob: Modification 2 to PD; MI: 20 multiply imputed data sets. Results over 1000 replicates.

(a) $Z \sim$ Normal(0,1)

		MAR $\Pr(M = 1\|Z, U) = f(Z)$			MNAR $\Pr(M = 1\|Z, U) = f(U)$		
ρ	Method	Relative bias (%)	Coverage (%)	CI width	Relative bias (%)	Coverage (%)	CI width
0.8	ML	−0.1	*96.4*	0.12	0.3	95.8	0.13
	Modified ML	0.0	96.2	0.12	0.4	96.1	0.14
	PD	0.3	95.9	0.12	0.3	95.6	0.13
	PD-redraw	0.1	96.3	0.12	−0.1	95.6	0.14
	PD-predprob	0.0	94.9	0.11	0.0	94.6	0.13
	MI	0.0	96.2	0.12	0.3	95.7	0.14
0.5	ML	0.0	95.0	0.13	0.5	**91.7**	0.28
	Modified ML	0.0	95.0	0.13	0.6	*96.6*	0.31
	PD	0.4	94.5	0.13	1.6	94.8	0.24
	PD-redraw	0.2	*97.6*	0.15	1.5	*96.5*	0.25
	PD-predprob	0.1	**93.1**	0.12	1.5	93.8	0.24
	MI	0.1	95.2	0.13	1.5	94.9	0.25
0.2	ML	−0.1	94.9	0.13	4.3	**64.7**	0.49
	Modified ML	−0.1	95.1	0.14	4.4	96.2	0.54
	PD	0.2	94.2	0.13	6.5	*98.7*	0.34
	PD-redraw	0.1	*98.6*	0.16	6.4	*99.1*	0.35
	PD-predprob	0.1	**93.2**	0.12	6.4	*98.6*	0.34
	MI	0.1	94.7	0.14	5.8	*97.7*	0.36

Bolded values are below 1.96 simulation standard errors.
Italicized values are above 1.96 simulation standard errors.

proxy $X = a_0 + a_1 Z$. Note that in this simulation ρ is the correlation in the entire sample, not just among respondents. The binary outcome Y was then created as $y_i = 1$ if $u_i > 0$, with values of a_0 chosen so that $E[Y] = 0.3$. The missing data indicator m_i was generated according to a logistic model,

$$\text{logit}(\Pr(m_i = 1|u_i, z_i)) = \gamma_0 + \gamma_Z z_i + \gamma_U u_i, \tag{11}$$

and values of y_i were deleted when $m_i = 1$. The two different missingness mechanisms selected were MAR, with $\gamma_Z = 0.5$, $\gamma_U = 0$, and extreme MNAR, with $\gamma_Z = 0$, $\gamma_U = 0.5$. Aside from the discrepancy of marginal versus conditional normality, these two mechanisms correspond to ϕ values of 0 and 1, respectively. For both scenarios, values of γ_0 were selected to induce approximately 50% missingness.

The process of generating $\{z_i, u_i, y_i, m_i\}$, and inducing missingness was repeated 1,000 times for each of the eighteen populations (3 Z distributions × 3 ρ values × 2 missingness mechanisms). The same six estimators for the mean of the binary outcome Y and its variance were obtained for each of the eighteen data sets as in the previous simulation. For the MAR mechanism, ϕ was taken to be zero, and for MNAR $\phi = 1$.

Table 2. Continued.

(b) $Z \sim$ Gamma(4,0.5)

		MAR $\Pr(M = 1 \mid Z, U) = f(Z)$			MNAR $\Pr(M = 1 \mid Z, U) = f(U)$		
ρ	Method	Relative bias (%)	Coverage (%)	CI width	Relative bias (%)	Coverage (%)	CI width
0.8	ML	7.4	**87.7**	0.13	11	**79.6**	0.12
	Modified ML	2.3	94.2	0.12	9.3	**84.3**	0.12
	PD	7.8	**86.1**	0.12	11	**78.9**	0.12
	PD-redraw	0.1	95.5	0.12	4.0	**93.5**	0.12
	PD-predprob	0.0	**92.9**	0.11	4.0	**91.2**	0.11
	MI	−0.1	94.7	0.12	4.2	95.3	0.13
0.5	ML	2.1	94.2	0.13	7.7	**86.7**	0.18
	Modified ML	0.8	95.2	0.13	6.8	**90.1**	0.21
	PD	2.5	**92.8**	0.13	8.1	**86.8**	0.17
	PD-redraw	0.4	*97.7*	0.15	4.9	**92.2**	0.18
	PD-predprob	0.3	**92.5**	0.12	4.9	**89.3**	0.17
	MI	0.2	94.5	0.13	5.2	**92.6**	0.19
0.2	ML	0.3	95.5	0.14	−2.2	**64.8**	0.33
	Modified ML	0.1	95.4	0.14	−2.3	94.2	0.46
	PD	0.6	94.8	0.13	4.8	*97.5*	0.30
	PD-redraw	0.3	*98.7*	0.16	3.8	*98.4*	0.31
	PD-predprob	0.2	93.7	0.12	3.9	*97.6*	0.30
	MI	0.4	94.8	0.14	4.0	96.2	0.32

Bolded values are below 1.96 simulation standard errors.
Italicized values are above 1.96 simulation standard errors.

5.3.1. Results

When Z is normally distributed, results are similar to the previous simulation, as seen in Table 2a. All methods have negligible bias across all scenarios, except when $\rho = 0.2$ under MNAR. For this population, there is a small bias but all methods except unmodified ML still achieve nominal coverage, and in fact many show higher than nominal coverage. The consistently best performing methods are Modified ML, PD-redraw, and MI, which achieve close to nominal coverage in all scenarios. PD-predprob shows undercoverage when the proxy is strong under MAR. As was previously seen, the unmodified ML procedure has intervals that are too short under MNAR with a weak proxy ($\rho = 0.2$), and thus exhibits very poor coverage. Modified ML fixes this problem, since the bootstrap is used for variance estimation instead of the large-sample approximation, though the intervals are more than 50% longer than with other methods.

Table 2b shows results for the slightly skewed proxy, when Z has a Gamma distribution. In general, all estimation methods perform better with weaker proxies when the covariate is skewed. The methods that rely the most on the underlying normality assumption of the PPM analysis, unmodified ML and PD, show bias for the stronger proxies under both missingness mechanisms and hence tend to undercover. When missingness is at random, as

Table 2. Continued.

(c) $Z \sim$ Exponential(1)

		MAR $\Pr(M = 1 \mid Z, U) = f(Z)$			MNAR $\Pr(M = 1 \mid Z, U) = f(U)$		
ρ	Method	Relative bias (%)	Coverage (%)	CI width	Relative bias (%)	Coverage (%)	CI width
0.8	ML	15	**71.7**	0.13	20	**42.0**	0.11
	Modified ML	4.5	**91.8**	0.12	16	**60.8**	0.11
	PD	15	**69.6**	0.13	20	**42.3**	0.11
	PD-redraw	−0.6	94.4	0.12	4.6	**91.5**	0.11
	PD-predprob	−0.6	**90.9**	0.10	4.7	**88.1**	0.10
	MI	−0.7	94.2	0.12	4.8	**93.5**	0.12
0.5	ML	3.1	93.9	0.14	13	**71.5**	0.14
	Modified ML	0.9	94.7	0.13	11	**83.0**	0.19
	PD	3.6	**93.2**	0.13	13	**71.5**	0.14
	PD-redraw	−0.8	*96.7*	0.14	5.4	**93.2**	0.15
	PD-predprob	−0.9	**92.7**	0.12	5.4	**89.2**	0.14
	MI	−0.9	94.0	0.13	5.6	**93.3**	0.16
0.2	ML	−0.6	95.0	0.14	0.2	**66.0**	0.24
	Modified ML	−0.8	94.9	0.14	0.1	94.5	0.41
	PD	−0.1	95.1	0.13	5.7	95.0	0.26
	PD-redraw	−0.9	*98.1*	0.15	3.2	*97.2*	0.27
	PD-predprob	−0.9	**93.6**	0.12	3.3	96.3	0.26
	MI	−0.9	94.5	0.13	3.4	96.0	0.28

Bolded values are below 1.96 simulation standard errors.
Italicized values are above 1.96 simulation standard errors.

before the best performers are Modified ML, PD-redraw, and MI, with some overcoverage by PD-redraw. The other PD modification (PD-predprob) slightly undercovers for stronger proxies. The more difficult populations are under MNAR. For both $\rho = 0.8$ and $\rho = 0.5$ all methods exhibit some bias, with the unmodified ML and PD methods showing the most bias. Subsequently, almost all methods fail to achieve nominal coverage. The exception is MI, which is at nominal coverage for all but one scenario. Under MNAR, for the weakest proxy ($\rho = 0.2$) the unmodified ML again shows undercoverage, while Modified ML corrects this problem. However, it does so with very large confidence intervals relative to the PD and MI methods, which reach nominal coverage.

Results for Z having an Exponential distribution are displayed in Table 2c. The results are similar to the Gamma case, with larger biases and lower coverage rates across all populations. For both mechanisms, coverage actually increases for all methods as ρ decreases (the exception being ML under MNAR with $\rho = 0.2$). Under MNAR, it is difficult for any estimation method to reach nominal coverage except when the proxy is weak. In that scenario, the modified ML, both modified PD, and MI methods reach nominal coverage, while the unmodified ML and PD methods remain biased and have poor coverage. Of course, confidence intervals are very wide, especially Modified ML, which has nearly 50% longer intervals.

Overall, the best performing method is MI, which achieves nominal or just under nominal coverage for all three distributions of Z, including the severely skewed Exponential, and under both missingness mechanisms with all strengths of proxies. This result is not unexpected. Even though MI uses the fully parametric PPM model to generate posterior draws of the parameters, these draws are subsequently used to impute the missing Y values via the conditional distribution of $[U|X, M = 1]$. Even if the proxy is not normally distributed, the conditional distribution of the latent variable given the proxy is normal by definition, and so MI is the least sensitive to departures from normality in the proxy.

The one other method that does reasonably well in most scenarios is the first modification to the Bayesian draws, PD-redraw. As with MI, this method conditions on the proxy and draws the latent U and thus outperforms the unmodified Bayesian method that relies entirely on the joint normality of U and the proxy X. PD-redraw achieves at or near nominal coverage for strong proxies across all levels of skewness, but exhibits overcoverage for weaker proxies. This is to be expected, since in this modification the latent U for respondents are redrawn unconditional on the observed Y, which is effectively imputing the observed Y, and certainly has the potential to add unnecessary variability, as was noted in Subsection 4.1.

6. Application to the 2015 Ohio Medicaid Assessment Survey

We now return to the missing income data for the 2015 OMAS (RTI International 2015) and apply the PPM model to assess the potential nonresponse bias in the estimated proportion of people at various income levels. Overall, 22.2% of subjects ($n = 9,511$) were missing the categorical income variable, and there were low levels of missingness for other variables important for the construction of design weights (e.g., age, gender, race). In order to focus on the potential bias in the income variable, we used the publicly available data that already had these other variables singly imputed (using a weighted sequential hot deck procedure).

We created two dichotomized income variables to analyze separately with the PPM model. Low income was defined as 138% or lower of the Federal Poverty Level (FPL), and high income was defined as 300% or more of the FPL. These two binary variables were analyzed separately using the PPM model; see Section 7 for a discussion of the extension of the model to ordinal outcomes. Covariates that were fully observed (or completed by single-imputation) and used in the analysis were stratum, region, household composition (number of adults, number of children), respondent age, respondent gender, respondent race, education level, insurance type, an indicator for fair/poor health status, and the sample weight. Probit regression was used to estimate the proxy, with the final models chosen based on the cases with income observed, using backwards selection starting from a model that contained all second-order interactions. Data and code for implementing the PPM model using the 2015 OMAS data is available at http://github.com/randridge/PPMA.

The ML estimates of the strength of the proxies were $\hat{\rho}^{(0)} = 0.71$ for low income, and $\hat{\rho}^{(0)} = 0.71$ for high income, reflecting a moderate to strong amount of information in the auxiliary data for predicting income. The standardized difference between the mean of the proxy X for the entire sample compared to that of the respondents, $d^* = \left(\bar{x} - \bar{x}^{(0)}\right)/\sqrt{S_{xx}^{(0)}}$, was $d^* = 0.066$ for low income and $d^* = -0.064$ for high income. Estimates of the proportion of low and high income and 95% intervals for each of $\phi = (0, 0.5, 1)$ were

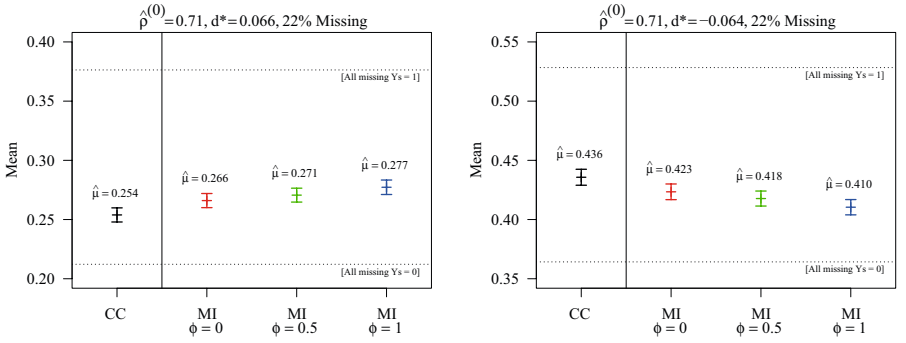

Fig. 2. Estimates of the proportion (a) low income and (b) high income based on 2015 Ohio Medicaid Assessment Survey data. CC: Complete cases; MI: 20 multiply imputed data sets using the proxy pattern mixture model with $\phi = 0$ (MAR), $\phi = 0.5$ (intermediate MNAR), and $\phi = 1$ (extreme MNAR)

obtained using the MI estimation procedure with $K = 20$ data sets. The burn-in period was 20 draws due to quick convergence and imputation occurred on every hundredth iteration. Since OMAS has a complex survey design, we used design-based estimators of the proportion using the survey weights via the "survey" routines in R, which estimate variances using Taylor series linearization (Lumley 2004).

Estimated proportions and 95% confidence intervals are displayed in Figure 2, with Manski bounds (obtained by filling in all zeros or all ones) denoted by dotted lines. For low income, the complete case (respondent) estimate underestimates the percent of subjects relative to the MAR and MNAR estimates, with the amount of underestimation increasing as ϕ increases. For high income, using the complete cases yields an estimate of the percent of high income earners that is too large, and as we move from MAR to MNAR the estimated proportion decreases. Since the proxies are relatively strong ($\hat{\rho}^{(0)}$), and the bias in the proxy (as measured by d^*) is relatively small, the confidence interval widths do not get drastically larger for $\phi = 1$.

While the potential impact of MNAR nonresponse on the estimated proportions do not seem very large, they are potentially large when put into context. Recall that 2015 was the first year OMAS was conducted after the Medicaid expansion in Ohio. Estimating the proportion of adults whose incomes are below 138% FPL is necessary to quantify the number of adults who are newly eligible for Medicaid. The difference between 26.6% low income and 27.7% low income could have large policy implications.

7. Discussion

In this article, we have extended the previously developed normal PPM analysis to handle binary data, which are ubiquitous in sample surveys. As with a continuous outcome, the new method integrates the three key components that contribute to nonresponse bias: the amount of missingness, differences between respondents and nonrespondents on characteristics that are observed for the entire sample, and the relationship between these fully observed covariates and the survey outcome of interest. The analysis includes, but does not make the assumption, that missingness is MAR, allowing the user to investigate a

range of MNAR mechanisms and the resulting potential for nonresponse bias. For the binary case, it is common to investigate what the estimates would be if all nonresponding units were zeros (or ones), and in fact the binary PPM analysis produces these two extremes (Manski bounds) when the proxy is weak, that is, when good predictors of Y among respondents do not exist. With even just moderately strong covariate information, as in the OMAS application, our new PPM method produces intervals that are considerably shorter than the Manski bounds, thus reducing the size of the potential nonresponse bias.

In the binary PPM model, ϕ is a sensitivity parameter and there is no information in the data with which to estimate it. Thus, we have proposed a sensitivity analysis using $\phi = \{0, 0.5, 1\}$ to produce separate estimates; comparing these estimates paints a picture of the potential nonresponse bias across a range of missingness mechanisms. An alternative approach would be to put a beta prior on f, as was done for the normal PPM model by Little et al. (2019). While this would produce a "single answer", the result is liable to be sensitive to how this prior is specified. If, for example, a noninformative Uniform(0,1) prior is used, the resulting posterior distribution interval for the mean of Y will effectively span the range from the estimate at $\phi = 0$ to the estimate at $\phi = 1$. This approach would produce a single interval and may be preferred by some practitioners.

There are some drawbacks to the binary PPM model, relative to the normal PPM on which it is based. The ease of implementation of the PPM model is lost in the binary case; closed-form ML estimates are no longer available and Bayesian and MI methods require iteration using Gibbs sampling. However, the ML solutions are good starting points for the Gibbs sampler and only very short burn-in periods were required in the examples we considered. An additional level of complexity in the binary case is the effect of skewed proxies. The normal PPM model estimates are relatively robust to departures from bivariate normality in the proxy and outcome, but the binary model relies more heavily on normality – even slight deviations away from normality in the latent variable lead to biased results. To address this weakness, we introduced modified estimators that perform better when the normality assumption is violated, while maintaining good performance when the normality assumption holds.

We have described three different estimation methods for the binary PPM model: ML, fully Bayesian (PD), and MI. In our investigations MI is consistently the best of these approaches. Unlike ML and Bayesian methods, it does not require a modification to handle skewed proxies, and complex design features like design weights are readily incorporated in the complete-data component of the MI combining rules. However, the ML and Bayesian methods do have one potential advantage: they can be applied in scenarios where microdata are available for respondents only (not available for the nonrespondents). If one has only the sample mean and covariance matrix of Z (the auxiliary variables) for nonrespondents, the unmodified ML and Bayesian methods could be used. In fact, if one only has the summary mean and covariance matrix of Z for the population, along with microdata for the respondents only, as in poststratification, the nonrespondent mean and covariance could be "backed out" of the population values, thus allowing ML and Bayesian estimation of the binary PPM model.

The binary PPM model can be extended to ordinal outcomes. Suppose instead of a binary outcome, we observe a partially missing ordinal outcome Y, where Y_i takes one of J ordered values, $1, \ldots, J$. As with the binary case, we assume there is an underlying latent

continuous variable U, related to the observed Y through the rule that $Y = j$ if $\gamma_{j-1} < U < \gamma_j$ for $j = 1, \ldots, J$, with $\gamma_0 = -\infty$ and $\gamma_J = \infty$. This latent structure allows an extension of probit regression to ordinal outcomes (e.g., Agresti 2002, chap. 7), such that $\Pr(Y \leq j|Z, M = 0) = \Pr(U \leq \gamma_j) = \Phi(\gamma_j + \alpha Z)$. For the PPM we take the proxy $X = \hat{\alpha} Z$, and apply the proxy pattern-mixture Model (4) to the joint distribution of the proxy X and latent U. Resulting ML estimates of the parameters $\mu_u^{(1)}$ and $\sigma_{uu}^{(1)}$ have the same form as in the binary case. The ML estimates of the parameters of the distribution of X for respondents and nonrespondents are the usual estimators. This leaves $\mu_u^{(0)}, \sigma_{uu}^{(0)}$, $\rho^{(0)}$ and $\gamma = \{\gamma_j\}$ to be estimated. Without loss of generality we take $\mu_u^{(0)} = 0$ and $\sigma_{uu}^{(0)} = 1$ and obtain ML estimates for the correlation $\rho^{(0)}$ and cutpoints γ. This reduces to the problem of estimating the polyserial correlation between the ordinal Y and continuous X, first considered by Cox (1974). As with the binary case, there is no closed-form solution and an iterative solution is required. The Bayesian and MI estimation methods follow as direct extensions of the binary case.

In future work we plan to extend PPM analysis to domain estimation, an important issue in practice. When the domain indicator is fully observed (for example, region in the OMAS data), application of the PPM model is straightforward; the domain indicator can be included in the model that creates the proxy, or the entire PPM method can be applied separately within each domain. If using multiple imputation, the product of the sampling weights and the domain indicator should be included as covariates in constructing the proxy (i.e., as part of Z) so MI variance estimates are unbiased (Kim et al. 2006). The more complex case is when the domain indicator and outcome are jointly missing, for example, income and current Medicaid status in the OMAS data. The methods of Little and Wang (1996), who extend the bivariate pattern-mixture model to the multivariate case when there are two patterns of missingness, may be useful in this scenario.

8. References

Agresti, A. 2002. *Categorical Data Analysis*. New York: Wiley.

Albert, J.H. and S. Chib. 1993. "Bayesian Analysis of Binary and Polychotomous Response Data." *Journal of the American Statistical Association* 88: 669–679. DOI: https://doi.org/10.1080/01621459.1993.10476321.

Andridge, R., A.M. Noone, and N. Howlader. 2017. "Imputing estrogen receptor (ER) status in a populationbased cancer registry: a sensitivity analysis." *Statistics in Medicine* 36: 1014–1028. DOI: https://doi.org/10.1002/sim.7193.

Andridge, R.R. and R.J.A. Little. 2011. "Proxy Pattern-Mixture Analysis for Survey Nonresponse." *Journal of Official Statistics* 27: 153–180. DOI: https://doi.org/10.1214/15-AOAS878SUPP.

Andridge, R.R. and K.J. Thompson. 2015. "Assessing nonresponse bias in a business survey: proxy pattern-mixture analysis for skewed data." *The Annals of Applied Statistics* 9: 2237–2265. DOI: https://doi.org/10.1214/15-AOAS878.

Barnhart, W.R., D. Ellsworth, A.C. Robinson, J. Myers, R.R. Andridge, and S.M. Havercamp. 2019. "Caregiving in the shadows: National analysis of health outcomes and intensity and duration of care among those who care for people with mental illness

and for people with developmental disabilities." *Disability and Health Journal* 3: 100837. DOI: https://doi.org/10.1016/j.dhjo.2019.100837.

Brick, J.M. and D. Williams. 2013. "Explaining Rising Nonresponse Rates in Cross-Sectional Surveys." *The ANNALS of the American Academy of Political and Social Science* 645: 36–59. DOI: https://doi.org/10.1177/0002716212456834.

Cox, N.R. 1974. "Estimation of the Correlation Between a Continuous and a Discrete Variable." *Biometrics* 30: 171–178. DOI: https://doi.org/10.2307/2529626.

Curtain, R., S. Presser, and E. Singer. 2005. "Changes in Telephone Survey Nonresponse over the Past Quarter Century." *Public Opinion Quarterly* 69: 87–98. DOI: https://doi.org/10.1093/poq/nfi002.

Heckman, J.J. 1976. "The Common Structure of Statistical Models of Truncation, Sample Selection, and Limited Dependent Variables and a Simple Estimator for Such Models." *The Annals of Economic and Social Measurement* 5: 475–492.

Hedeker, D., R.J. Mermelstein, and H. Demirtas. 2007. "Analysis of binary outcomes with missing data: missing = smoking, last observation carried forward, and a little multiple imputation." *Addiction* 102: 1564–1573. DOI: https://doi.org/10.1111/j.1360-0443.2007.01946.x.

Higgins, J.P.T., I.R. White, and A.M. Wood. 2008. "Imputation methods for missing outcome data in meta-analysis of clinical trials." *Clinical Trials* 5: 225–239. DOI: https://doi.org/10.1177/1740774508091600.

Jackson, D., I.R. White, D. Mason, and S. Sutton. 2014. "A general method for handling missing binary outcome data in randomized controlled trials." *Addiction* 109: 1286–1993. DOI: https://doi.org/10.1111/add.12721.

Kim, J.K., J.M. Brick, W.A. Fuller, and G. Kalton. 2006. "On the Bias of the Multiple-Imputation Variance Estimator in Survey Sampling." *Journal of the Royal Statistical Society B* 68: 509–521. DOI: https://doi.org/10.1111/j.1467-9868.2006.00546.x.

Little, R.J.A. 1993. "Pattern-Mixture Models for Multivariate Incomplete Data." *Journal of the American Statistical Association* 88: 125–134. DOI: https://doi.org/10.2307/2533148.

Little, R.J.A. 1994. "A Class of Pattern-Mixture Models for Normal Incomplete Data." *Biometrika* 81: 471–483. DOI: https://doi.org/10.1093/biomet/81.3.471.

Little, R.J.A. and D.B. Rubin. 2019. "Statistical Analysis with Missing Data." 3rd edition. Wiley: New York.

Little, R.J.A. and Y. Wang. 1996. "Pattern-Mixture Models for Multivariate Incomplete Data with Covariates." *Biometrics* 52: 98–111. DOI: https://doi.org/10.1080/01621459.1993.10594302.

Little, R.J.A., B.T. West, P.S. Boonstra, and J. Hu. 2019. "Measures of the Degree of Departure from Ignorable Sample Selection." *Journal of Survey Statistics and Methodology*. DOI: https://doi.org/10.1093/jssam/smz023.

Liublinska, V. and D.B. Rubin. 2014. "Sensitivity analysis for a partially missing binary outcome in two-arm randomized clinical trial." *Statistics in Medicine* 33: 4170–4185. DOI: https://doi.org/10.1002/sim.6197.

Lumley, T. 2004. "Analysis of Complex Survey Samples." *Journal of Statistical Software* 9: 1–19.

Magder, L.S. 2003. "Simple approaches to assess the possible impact of missing outcome information on estimates of risk ratios, odds ratios, and risk differences." *Controlled Clinical Trials* 24: 411–421. DOI: https://doi.org/10.1016/s0197-2456(03)00021-7.

Manski, C.F. 2016. "Credible Interval Estimates for Official Statsitics with Survey Nonresponse." *Journal of Econometrics* 191: 293–301. DOI: https://doi.org/10.1016/j.jeconom.2015.12.002.

Muthen, B., T. Asparouhov, A.M. Hunter, and A.F. Leuchter. 2011. "Growth Modeling With Nonignorable Dropout: Alternative Analyses of the STAR∗D Antidepressant Trial." *Psychological Methods* 16: 17–33. DOI: https://doi.org/10.1037/a0022634.

Nandram, B. and J.W. Choi. 2002a. "A Bayesian Analysis of a Proportion Under Non-Ignorable Non-Response." *Statistics in Medicine* 21: 1189–1212. DOI: https://doi.org/10.1002/sim.1100.

Nandram, B. and J.W. Choi. 2002b. "Hierarchical Bayesian Nonresponse Models for Binary Data from Small Areas with Uncertainty about Ignorability." *Journal of the American Statistical Association* 97: 381–388. DOI: https://doi.org/10.1198/016214502760046934.

Nandram, B., G. Han, and J.W. Choi. 2002. "A Hierarchical Bayesian Nonignorable Nonresponse Model for Multinomial Data from Small Areas." *Survey Methodology* 28: 145–156.

Nandram, B., N. Liu, J.W. Choi, and L. Cox. 2005. "Bayesian Non-response Models for Categorical Data from Small Areas: An Application to BMD and Age." *Statistics in Medicine* 24: 1047–1074. DOI: https://doi.org/10.1002/sim.1985.

Olsson, U., F. Drasgow, and N.J. Dorans. 1982. "The Polyserial Correlation Coefficient." *Psychometrika* 47: 337–347. DOI: https://doi.org/10.1007/BF02294164.

R Core Team. 2017. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. Available at: https://www.R-project.org (accessed June 2020).

RTI International. 2015. *2015 Ohio Medicaid Assessment Survey Methodology Report*. Technical report, RTI International. Available at: http://grc.osu.edu/sites/default/files/inline-files/12015OMASMethReptFinal121115psg.pdf (accessed June 2020).

Rubin, D.B. 1976. "Inference and Missing Data" (with Discussion). *Biometrika* 63: 581–592. DOI: https://doi.org/10.1093/biomet/63.3.581.

Rubin, D.B. 1977. "Formalizing Subjective Notions About the Effect of Nonrespondents in Sample Surveys." *Journal of the American Statistical Association* 72: 538–542. DOI: https://doi.org/10.1080/01621459.1991.10475033.

Rubin, D.B. 1978. "Multiple Imputation in Sample Surveys." A Phenomenological Bayesian Approach to Nonresponse. In Proceedings of the Survey Research Methods Section, American Statistical Association (San Diego, CA): 20–34. DOI: https://doi.org/10.1002/9780470316696.

Stasny, E.A. 1991. "Hierarchical Models for the Probabilities of a Survey Classification and Nonresponse: An Example from the National Crime Survey." *Journal of the American Statistical Association* 86: 296–303. DOI: https://doi.org/10.1080/01621459.1991.10475033.

Sullivan, D. and R. Andridge. 2015. "A hot deck imputation procedure for multiply imputing nonignorable missing data: The proxy pattern-mixture hot deck."

Computational Statistics and Data Analysis 82: 173–185. DOI: https://doi.org/10.1016/j.csda.2014.09.008.

Tate, R.F. 1955a. "Applications of Correlation Models for Biserial Data." *Journal of the American Statistical Association* 50: 1078–1095. DOI: https://doi.org/10.1080/01621459.1955.10501293.

Tate, R.F. 1955b. "The Theory of Correlation Between Two Continuous Variables When One is Dichotomized." *Biometrika* 42: 205–216. DOI: https://doi.org/10.2307/2333437.

Received August 2018
Revised May 2019
Accepted October 2019