mda

methods, data, analyses

Journal for Quantitative Methods and Survey Methodology

Volume 10, 2016 | 1

Edited by Annelies G. Blom, Edith de Leeuw,
Gabriele Durrant, Bärbel Knäuper

methods, data, analyses is published by GESIS – Leibniz Institute for the Social Sciences.

Managing editor: Sabine Häder
GESIS – Leibniz Institute for the Social Sciences
PO Box 12 21 55
68072 Mannheim
Germany
Tel.: + 49.621.1246282
E-Mail: mda@gesis.org
Internet: www.gesis.org/mda

Methods, data, analyses (mda) publishes research on all questions important to quantitative methods, with a special emphasis on survey methodology. In spite of this focus we welcome contributions on other methodological aspects. We especially invite authors to submit articles extending the profession's knowledge on the science of surveys, be it on data collection, measurement, or data analysis and statistics. We also welcome applied papers that deal with the use of quantitative methods in practice, with teaching quantitative methods, or that present the use of a particular state-of-the-art method using an example for illustration.

All papers submitted to mda will first be screened by the editors for general suitability and then double-blindly reviewed by at least two reviewers. Mda appears in two regular issues per year (June, December).

Please register for a subscription via http://www.gesis.org/en/publications/journals/mda/subscribe

Print: Bonifatius Druck GmbH Paderborn, Germany

ISSN 1864-6956 (Print)
ISSN 2190-4936 (Online)

Content

RESEARCH REPORTS

Editorial

Annelies G. Blom
School of Social Sciences, University of Mannheim

It is with great pleasure that I present my first issue as editor-in-chief of *methods, data, analyses.* I would like to take this occasion to express my gratitude to the previous editor-in- chief, Henning Best, for three laborious years of service, during which he has transitioned the originally German-language journal to an internationally-oriented open access journal in English. I would also like to thank the two departing Associate Editors, Marek Fuchs and Petra Stein, for their years of support of the journal, and Sabine Häder for her continued commitment as Managing Editor.

On purpose, the new editorial board has a distinctly international composition, with Edith de Leeuw from Utrecht University in the Netherlands, Bärbel Knäuper from the University of Montreal in Canada, Gabriele Durrant from the University of Southampton in the UK and myself, covering a broad range of survey methodological and statistical expertise. This way we aim to further foster the international readership of and contributions to the journal.

The journal's internationalisation of the past years is depicted below (grey line in Figure 1). While before 2014, when English was introduced as the sole language of publication, 13% of published articles were in English, today it is 100%. At the same time, the acceptance rate has been relatively stable over the years with on average 44% of submitted manuscripts accepted for publication after peer-review (black line in Figure 1).

methods, data, analyses has been among the first journals to follow an Open Science mission, where publicly-funded research findings are made freely available to the scientific community and general public, with its predecessor – ZUMA Nachrichten – already giving free access to the journal in pre-Internet times. Continuing in the Open Science tradition, making the journal even more attractive for submissions from both established academics and young scholars, who are under increas-

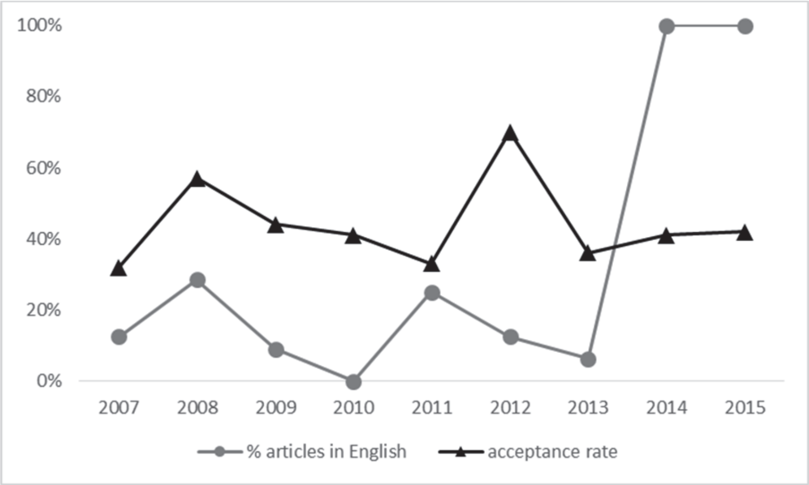


Figure 1 Proportion of English-language articles and acceptance rate

ing pressure to publish swiftly and in journals on the Citation Indices, is our key objective for the coming year.

One step towards this goal will be the introduction of an online manuscript submission and review system based on the Open Journal System. This will enable us to follow the review process more closely and professionally. Another step is the Online First publication of manuscripts, which comes into effect this spring. This allows the ongoing and timely publication of research in *methods, data, analyses* with short turn-over cycles of the submitted manuscripts. The DOI numbers for each article, which were introduced in the journal in 2013, will uniquely cross-identify the Online First and later printed journal publications.

In summary, this journal distinguishes itself from the great many closed-access journals as one of the few outlets for survey methodological and survey statistical work that can reach beyond privileged academic communities in Western universities. With its continued professionalization and internationalization we aim to contribute to scientific knowledge creation and dissemination. For these reasons, I look forward to shaping the future of the mda together with you, through your manuscript submissions, involvement as reviewers, discussions and recommendations!

methods, data, analyses | Vol. 10(1), 2016, pp. 5-24 DOI: 10.12758/mda.2016.001

Non-Observation Bias in an Address-Register-Based CATI/CAPI Mixed Mode Survey

Oliver Lipps

Swiss Centre of Expertise in the Social Sciences (FORS)

Abstract

Landline surveys suffer from an increasing risk of excluding a relevant share of the population. To analyze and correct telephone coverage issues, face-to-face surveys are often used, which contain questions about landline ownership and registration. Others use dual frame approaches and compare results from the landline with another mode. However, such surveys lack information about unobserved sample members.

In this article we analyze representation bias using a household survey with a sample drawn from a population register, where landline is used for households with a matched landline, and face-to-face for those without. We distinguish between the different components of nonobservation, including landline undercoverage, non-contact, and non-cooperation, by either incorporating face-to-face sample members or not, and by the fieldwork phases to recruit households and individuals. Our main interest is how biases from each of these components add up to a final representation bias in the responding sample. In addition, we analyze income and deprivation differences by either including face-to-face sample members or not.

The strongest representation bias in the telephone sample on the household level is caused by telephone undercoverage. The combined sample suffers much less from representation bias, which mostly stems from noncooperation. In terms of income and deprivation differences, our results show that the face-to-face sample is poorer than the telephone sample and needs to be considered for unbiased estimates. Based on these findings we offer some fieldwork recommendations to help reduce selection bias based on the different reasons for nonobservation.

Keywords: mixed mode, telephone number matching, paradata, coverage, contact, cooperation, representation bias

1 Bias from Nonobservation in Surveys with Telephone as the Main Mode

In addition to nonresponse, landline telephone surveys are increasingly challenged by undercoverage (e.g., Peytchev et al., 2011). This latter issue results from a dramatic increase in the proportion of "mobile-only" households (Mohorko et al., 2013; Sala & Lillini, 2014) and an increasing proportion of individuals who no longer wish to be listed in a public directory (Blumberg & Luke, 2014; De Vitiis & Righi, 2011; Ernst Stähli, 2012; Joye et al., 2012; Link & Fahimi, 2013; Sala & Lillini, 2014; Von der Lippe et al., 2011). Landline coverage rates depend on contexts and effort. For example, Brick et al. (2011) used commercial sources to match telephone numbers to a random sample of addresses in the US, and achieved a 57% telephone matching rate. In Switzerland, the Swiss Federal Statistical Office (SFSO) matches register-based samples against its own register of telephone numbers, which includes both publicly listed and unlisted landline numbers. SFSO matching rates of randomly sampled individuals reach an average of 76% (Joye, 2012). A comparable Swiss telephone survey, which is based on register-based samples but uses additional sources of telephone numbers such as commercial databases instead of unlisted landline numbers[1], reports a matching rate of 86% (Lipps & Kissau, 2012).

Undercoverage is compounded by the fact that people with or without a listed landline differ on the basis of socio-demographic information (Busse & Fuchs, 2012; Cobben & Bethlehem, 2005; Lipps & Kissau, 2012; Mohorko et al., 2013; Sala & Lillini, 2014). For example, there is evidence that people without a landline are more likely to be men, living alone, who are young and foreign (Lipps & Kissau, 2012;, Link et al., 2007; Schneiderat & Schlinzig, 2012), Consequently, landline surveys tend to overrepresent women, older people, those with a low or a high education level (students), and households without children (Sala & Lillini, 2014). In addition, there is evidence of substantive variables bias (Joye et al., 2012; Sala & Lillini, 2014) in landline surveys, including for example, an overrepresentation of people who are more satisfied with their lives (Mohorko et al., 2013). Others identify more homeowners (Sala & Lillini, 2014), fewer people who live below the poverty threshold (Safir & Goldenberg, 2008), fewer minority respondents (Holbrook et al., 2003), and a higher average household income (Gordoni, 2010; Holbrook et al., 2003; Schneiderat & Schlinzig, 2012).

[1] The SFSO does not provide unlisted telephone numbers to commercial survey agencies.

Direct correspondence to
Oliver Lipps, Swiss Centre of Expertise in the Social Sciences (FORS)
c/o University of Lausanne, CH – 1015 Lausanne
E-mail: oliver.lipps@fors.unil.ch

To analyze and correct bias from landline telephone undercoverage, some researchers use face-to-face surveys which contain questions on landline ownership and registration (e.g., Joye et al., 2012; Mohorko et al., 2013; Sala & Lillini, 2014). However, face-to-face surveys are expensive and if these surveys suffer from selective nonresponse the results then become questionable. For example, it is possible that households who own a listed landline are easier to reach by telephone than by face-to-face and may not be contacted using the face-to-face mode. In addition, telephone households may be more willing to participate (Sala & Lillini, 2014). Other researchers use experimental data including telephone and face-to-face samples both drawn independently at random (e.g., Holbrook et al., 2003). Nevertheless such experiments are also expensive and cannot replace large-scale social surveys. Other researchers also use a landline survey and, in addition, sample mobile-only members (e.g., Link et al., 2007; Lohr & Brick, 2014; Schneiderat & Schlinzig, 2012). Nonetheless interviews using mobile phones generally suffer from high nonresponse rates (Schneiderat & Schlinzig, 2012). In addition it still remains unknown whether the data quality of social science surveys via mobile phone is sufficient due to location issues, voice quality and net availability aspects, third party influence on socially desired answers (Kühne & Häder, 2012), or other factors affecting measurement errors (Lynn & Kaminska, 2012). Finally, extending a landline sampling frame to include mobile phones is not an easy task in European countries (Heckel & Wiese, 2012).

An alternative to analyzing and correcting bias from landline telephone undercoverage is to use additional survey modes to approach sample members without access to a landline (e.g., Cobben, 2009). However, knowledge about the extent to which sample representation can be improved due to the inclusion of additional survey modes for those without access to the primary mode is scarce. In the present research, we analyze bias from undercoverage and from nonresponse using a general population mixed mode survey, where the landline is the mode for households with a landline, and face-to-face for those without. The sample of this survey was drawn from a population register which includes basic socio-demographic variables, in addition to fully covering the population. Specifically, we analyze to what extent 1. the additional mode is able to decrease the number of errors from undercoverage in the telephone sample, 2. errors from the two main components of nonresponse, non-contact and non-cooperation, can be decreased by adding the face-to-face mode, 3. substantive variables are different in the telephone-only compared with the combined sample. As for 2., to distinguish non-contact and non-cooperation is not common in the literature (e.g., Peytchev et al., 2011; but see Cobben, 2009 and Olson, 2007), even though this distinction was previously noted over sixty years ago (Deming, 1947).

The article is organized as follows. First, we introduce the data and the socio-demographic frame variables. Next, we model bias in the frame variables according to the different reasons for nonobservation. We compare predicted probabilities from multivariate logit models distinguishing the telephone and the combined telephone/face-to-face sample, and the fieldwork phases of, first, recruiting households and, second, recruiting enumerated household members. Finally, we analyze income and deprivation differences when either including the face-to-face sample members or not. The final chapter concludes with sampling and fieldwork considerations.

2 Data

For this research we use survey and register data from Switzerland. The Swiss case is interesting, since the percentage of research turnover via the telephone is amongst the highest in Europe (Häder et al., 2012). Nevertheless we believe that our findings are generalizable to other countries where formerly high landline coverage rates are declining and also to surveys in which the face-to-face mode is used to contact households without a telephone. In addition the different language regions in Switzerland add variance: they not only have different landline coverage rates, but are characterized by different cultural backgrounds and behaviors. Finally, unlike most other (European) countries, a harmonized sampling frame is available based on population registers from which the Swiss Federal Statistical Office (SFSO) draws samples for specific surveys, including the Swiss Household Panel (SHP).

We use data from the SHP 2013 refreshment sample (SHP III). The SHP is a nationwide, annual panel survey, which started in 1999 with slightly more than 5,000 randomly selected households using the centralized telephone survey mode. Each year a letter announcing the survey is sent in advance to the sampled households. Then the household reference person, an adult with sufficient knowledge of the household, is asked to report the current household's composition in the grid questionnaire. Conditional to the completion of the household grid, all household members eligible for interview complete their individual questionnaires.

The SFSO drew the refreshment sample SHP III at random from the national register of individuals residing in Switzerland. The SHP III total sample comprises 11,110 persons aged 16 years and over, of which a random subsample of 9,048 persons was fielded.[2] All members registered in the same household as the sampled individuals can be identified via the household identifier. The register provides demographic information about all household members such as sex, age, nationality, civil status, and municipality, but no telephone numbers. These must be searched separately and matched to the sample. The SFSO matched the

[2] We dropped seven cases, among which were five who were surveyed using the web mode or could not be matched with call data, and two whose marital status was missing.

Table 1 Variables from sampling frame and categories used

Variable	Categories
Household size	1 person, 2 persons, 3 persons, 4 or more persons
Age of youngest child in household	No child, 0-6 years, 7-17 years
Language region	Swiss-German, French, Italian
Size of municipality of residence	more than 100,000 inhabitants, 20-100,000 inhabitants, 10-20,000 inhabitants, 5-10,000 inhabitants, 2-5,000 inhabitants, less than 2,000 inhabitants
Age group	16-30 years, 31-44 years, 45-58 years, 59-72 years, 73+ years
Nationality	Swiss or Swiss born, foreigners from one of the neighboring countries (sharing one of the Swiss national languages), other foreigners*
Civil status	single and never married (referred to as single), married (including separated), divorced, widowed
Sex	Women, Men

* See Lipps et al. (2013) for reasons why these two foreigner groups need to be distinguished in nonresponse analyses.

sample against its own register of telephone numbers. 7,396 (66.6%) households with publicly listed landline numbers were matched. After dropping the ineligible[3] households, we arrived at an analysis sample of 8,098 interview eligible households, of which 5,485 (67.7%) were from the telephone sample, and 2,613 from the face-to-face sample. All household members from the age of 16 years on were survey eligible in the first wave of the SHP III households considered here. Unlike the previous samples, members of the SHP III sample were not asked to fill out the individual questionnaire in their first wave, but were sent a biographical paper and pencil questionnaire with a pre-stamped envelope together with an unconditional incentive of 10 Swiss Francs.

In table 1 we depict the variables available from the sampling frame and the categories used in the analysis.

The reason for including language regions is that households living in the French or Italian speaking area of Switzerland have a lower landline coverage rate than those in the Swiss-German speaking part (see, e.g., Lipps & Pekari 2016). In addition we are interested in in-house effects: because the fieldwork for the Swiss-

[3] Address problems included empty or demolished houses, addresses of an institution or a secondary home, or matched telephone numbers that did not work, such as modems. Other ineligible sample members comprised of dead people or those having left the country (AAPOR, 2011).

German speaking part on the one hand and the French and the Italian speaking part on the other were conducted by different centers (of the same survey agency) there may be different results. Note that in the SHP, a household is defined as all people living together for a longer time span, having at least one common meal per week, and – perhaps most importantly – for whom the flat/house in question is their principal residence.

3 Modeling and Results

Using the fielded eligible households, for all frame variable characteristics we analyze the proportion of households still present after each recruitment step. We use the characteristics of the sampled individual to represent *individual* frame variables (age, nationality, marital status and sex)[4] on the household recruitment level. We distinguish bias due to unmatched telephone numbers, noncontact, and noncooperation, the latter two separated by the telephone matched sample alone and the telephone/face-to-face sample combined. We tested the dependency of subsequent models (e.g., cooperation can only be analyzed for people who are contacted) using probit models with a sample selection (Heckman selection models; see Cobben (2009) for its application to components of nonresponse). The estimated correlation between matching, contact and cooperation is significant on a 5%-level, but not on a 1%-level. Given our large sample sizes, we use independent logit models. In the following tables 2 and 4, we list predicted probabilities. Compared with beta-coefficients or odds ratios, predicted probabilities are comparable across models and easier to interpret (Mood, 2010).

3.1 Household Grid Level

In table 2 we depict average predicted probabilities from each step of nonobservation during the household recruitment phase. As a reading example, we find a telephone matching probability of 50.0% if every household in the data was treated as if they contained one-person (upper left figure). The probability of being in the sample after being asked to cooperate (and therefore the conditional response rate) would be 17.9% in the telephone sample, if the sample members were treated as if they were foreigners from a country other than a neighboring country. We describe significant (1%-level) differences between the categories of a variable when appropriate, but don't depict significance levels in table 2 due to readability.

[4] In only three households (with 12 individuals of age 16 years or older), different communication languages are recoded for at least two household members. We therefore treat language as a household variable.

Table 2 Predicted probabilities during the household recruitment phase

[average predicted probabilities from logit model]	Teleph. match	Contact Teleph.	Contact All	Coop. Teleph.	Coop. All
1 Person	0.500	0.483	0.826	0.288	0.415
2 Persons	0.630	0.622	0.895	0.378	0.489
3 Persons	0.789	0.767	0.927	0.472	0.527
4+ Persons	0.856	0.839	0.950	0.583	0.606
no children in household	0.698	0.674	0.886	0.397	0.480
youngest child in HH 0-6 years old	0.527	0.518	0.884	0.343	0.517
youngest child in HH 7-17 years old	0.645	0.648	0.916	0.428	0.538
Language Swiss-German	0.686	0.671	0.894	0.403	0.493
Language French	0.666	0.640	0.899	0.392	0.499
Language Italian	0.604	0.575	0.805	0.395	0.468
Municipality size >100K	0.658	0.642	0.867	0.395	0.471
Municipality size 20-100K	0.670	0.654	0.879	0.416	0.497
Municipality size 10-20K	0.666	0.645	0.901	0.394	0.501
Municipality size 5-10K	0.679	0.661	0.894	0.380	0.472
Municipality size 2-5K	0.679	0.661	0.901	0.405	0.510
Municipality size <2K	0.716	0.696	0.905	0.415	0.504
16-30 years old	0.413	0.410	0.814	0.283	0.452
31-44 years old	0.506	0.478	0.826	0.309	0.450
45-58 years old	0.720	0.698	0.902	0.430	0.515
59-72 years old	0.859	0.843	0.945	0.535	0.566
73+ years old	0.906	0.894	0.974	0.467	0.486
Native Swiss or born in Switzerland	0.716	0.697	0.903	0.437	0.519
from a neighbor. country	0.588	0.578	0.848	0.333	0.463
from another country	0.493	0.472	0.850	0.179	0.334
single	0.691	0.664	0.887	0.389	0.486
married	0.675	0.665	0.900	0.414	0.507
widowed	0.711	0.694	0.911	0.390	0.467
divorced	0.621	0.607	0.872	0.382	0.481
Women	0.691	0.672	0.900	0.405	0.496
Men	0.663	0.645	0.881	0.394	0.490
McFadden Pseudo R-squared	0.209	0.208	0.127	0.083	0.036
Mean value all households	0.677	0.659	0.890	0.400	0.493

Data: SHP III (2013 refreshment sample, N (households) = 8,098).

Also, we focus more on effect sizes than significance levels because the latter depends heavily on sample sizes. To give an example, there is a conditional matching probability of 69.1% for single households (N=2,702), of 67.5% for married households (N=3,748), and of 71.1% for widowed households (N=689) (see column “Teleph. match”, rows distinguishing marital status). Nevertheless, although the matching probability difference between single and married households (1.6% points) is smaller than that between married and widowed households (3.6% points), the former difference is significant while the latter is not.

For each nonobservation step, we define the representation bias of each socio-demographic group by the ratio of its predicted probability to the mean probability. These biases are shown in table 3. For example, the conditional matching probability of 50.0% of a one-person household over the sample mean of 67.7% (=0.739) gives an underrepresentation of 26.1%. In addition, we define as the nonobservation-specific representation bias the standard deviation of the representation bias across the groups (last row).[5] All telephone samples have a higher representation bias than the combined samples. By far the highest representation bias is provided in the first step by the unmatched telephone numbers (0.157). (Additional) bias from noncontact plays no role, and from noncooperation a minor one (0.181). In the combined samples, bias from noncontact and (additional) bias from noncooperation are similar and amount to 0.042 and 0.092, respectively.

In the following, we discuss the relevant representation biases of the frame variables, distinguished by the different steps of nonobservation.

Landline telephone matching (column “Teleph.match”)
Overall, 67.7% of all fielded households can be matched with a landline number (table 2). The larger the household, the higher the match probability. One-person households have a 26.1% underrepresentation and four or more person households a 26.4% overrepresentation. Households without children and those with children from 7 years on are well represented, while those with small children are underrepresented by 22.2%. Concerning language, Italian speakers are underrepresented among the telephone matched households by 10.8%, which is in line with experiences made by the SFSO (e.g., Joye, 2012). Households in small municipalities (<2,000 inhabitants) are slightly overrepresented. The older the household the easier it can be matched with a listed telephone number, with the youngest group underrepresented by 39.0%, and the oldest group overrepresented by 33.8%. Native Swiss or people born in Switzerland are easier to match than foreigners from a neighboring country who are in turn easier to match than other foreigners. Finally, widowed households are easier to match than divorced.

[5] Not to be confused with the R (representativity)-indicator, which is defined for all sample members, see e.g., Schouten et al. (2009). The R-indicator is defined as 1 - 2 * the standard deviation of the response probabilities. For convenience we use the standard deviation of the representation bias across the socio-demographic groups.

Table 3 Representation bias during the household recruitment phase

[average predicted probabilities / mean value]	Teleph. match	Contact Teleph.	Contact All	Coop. Teleph.	Coop. All
1 Person	0.739	0.733	0.928	0.720	0.842
2 Persons	0.931	0.944	1.006	0.945	0.992
3 Persons	1.165	1.164	1.042	1.180	1.069
4+ Persons	1.264	1.273	1.067	1.458	1.229
no children in household	1.031	1.023	0.996	0.993	0.974
youngest child in HH 0-6 years old	0.778	0.786	0.993	0.858	1.049
youngest child in HH 7-17 years old	0.953	0.983	1.029	1.070	1.091
Language Swiss-German	1.013	1.018	1.004	1.008	1.000
Language French	0.984	0.971	1.010	0.980	1.012
Language Italian	0.892	0.873	0.904	0.988	0.949
Municipality size >100K	0.972	0.974	0.974	0.988	0.955
Municipality size 20-100K	0.990	0.992	0.988	1.040	1.008
Municipality size 10-20K	0.984	0.979	1.012	0.985	1.016
Municipality size 5-10K	1.003	1.003	1.004	0.950	0.957
Municipality size 2-5K	1.003	1.003	1.012	1.013	1.034
Municipality size <2K	1.058	1.056	1.017	1.038	1.022
16-30 years old	0.610	0.622	0.915	0.708	0.917
31-44 years old	0.747	0.725	0.928	0.773	0.913
45-58 years old	1.064	1.059	1.013	1.075	1.045
59-72 years old	1.269	1.279	1.062	1.338	1.148
73+ years old	1.338	1.357	1.094	1.168	0.986
Native Swiss or born in Switzerland	1.058	1.058	1.015	1.093	1.053
from a neighbor. country	0.869	0.877	0.953	0.833	0.939
from another country	0.728	0.716	0.955	0.448	0.677
single	1.021	1.008	0.997	0.973	0.986
married	0.997	1.009	1.011	1.035	1.028
widowed	1.050	1.053	1.024	0.975	0.947
divorced	0.917	0.921	0.980	0.955	0.976
Women	1.021	1.020	1.011	1.013	1.006
Men	0.979	0.979	0.990	0.985	0.994
Standard deviation	0.157	0.160	0.042	0.181	0.092

Data: SHP III (2013 refreshment sample, N (households) = 8,098).

Noncontact (column “Contact Teleph.” and “Contact All”)
65.9% of all telephone fielded households can be successfully contacted (column “Contact Teleph.” in table 2). Because of the high telephone contact rate (65.9/67.7=.973; CON1 according to AAPOR 2011), there is not much room for a large bias change due to the uncontacted telephone households. None of the groups change bias by more than 3% points. The only groups that change by more than 2% points are households with older children, who decrease their under-representation by 3.0% points, and 31-44 years old households, who increase their under-representation by 2.2% points.

Adding the face-to-face survey mode boosts the proportion of contacted households from 65.9% to 89.0% (column “Contact All” in table 2), which is a likely reason for the much smaller standard deviation of the representation bias (0.042) compared to the contacted telephone sample (0.160). For example, the underrepresentation of one-person households is reduced to 7.2%, of foreigners from another than a neighboring country to 4.5%, and of young households to 8.5%. Conversely, the overrepresentation of large households decreases to 6.7%, and of older households to 9.4%.

Noncooperation (column “Coop. Teleph.” and “Coop. All”)
40.0% of the eligible telephone sample members participate in the survey (column “Coop. Teleph.” in table 2), which corresponds to a cooperation rate of 60.7% (=40.0/65.9; COOP1 according to AAPOR 2011). Substantial changes compared with the biases in the telephone contacted sample concern household size, age, and nationality in particular. Large households increase their overrepresentation by 18.5% points. Households without children decrease their underrepresentation (3.0% points) and are well represented in the sample of cooperating telephone households. While households with small children also decrease their underrepresentation (7.2% points), households with older children are now overrepresented. Italian speakers decrease their underrepresentation by 11.5% points and are now well represented. As for municipality sizes, there are small and nonlinear changes due to noncooperation. With respect to age groups, while young adults decrease their underrepresentation by 8.6% points and households between 31 and 44 years by 4.8% points, households between 59 and 72 years increase their overrepresentation by 5.9% points. Older people decrease it by 18.9% points. Native Swiss or people born in Switzerland increase their overrepresentation by 3.5% points, while foreigners from a neighboring country increase their underrepresentation by 4.5% points and other foreigners by 26.9% points. Widowed households decrease their overrepresentation by 7.8% points and are now well represented.

In the combined eligible sample (column “Coop. All” in table 2), 49.3% of the households participate (cooperation rate COOP1 55.4%). Small households see further losses due to noncooperation (8.6% points), while households with four or more persons increase their overrepresentation by 16.2% points. Households with small children increase their (formerly well) representation by 5.5% points, and households with older children by 6.2% points. Language and municipality size play a minor role. As for age, households between 59-72 years increase their overrepresentation by 8.6 % points, and older households decrease it by 10.8% points and are now well represented. Native Swiss or people born in Switzerland increase their overrepresentation by 3.8% points, while foreigners from a country other than a neighboring country increase their underrepresentation by 27.8% points. Finally, widowed households decrease their overrepresentation by 7.7% points and are now slightly underrepresented.

3.2 Person Level

We now turn to representation bias in terms of individual frame variables due to selective losses of individuals in households with a completed grid questionnaire. All enumerated individuals from the age of 16 years on are eligible for an interview. The 3,989 cooperating households report a total of 8,056 persons, of whom 7,826 were interview eligible and fielded. Similar to the household recruitment phase, we list predicted probabilities and representation bias for contact and cooperation, in table 4 and table 5, respectively.

85.1% of all enumerated interview eligible individuals can be contacted by telephone (column “Contact Teleph.” in table 4). Similar to the household recruitment phase, the older the individual the easier it is to obtain contact. Contacted young adults are underrepresented by 7.5%, contacted older individuals overrepresented by 11.5%. As in the household recruitment phase, native Swiss or people born in Switzerland are easier to contact than foreigners from a neighboring country, who in turn are easier to contact than other foreigners. The latter are underrepresented by 22.4%. Unlike during the household recruitment phase, the widowed are more difficult to contact, but still slightly easier than divorced people. If face-to-face sample members are included (column “Contact All” in table 4), the individual specific contact rate boosts to 98.3%, which again leaves little room for representation bias.

Considering cooperation rates by telephone (column “Coop. Teleph.”), people aged 73 years and over cooperate less than other age groups and change their overrepresentation from noncontact into underrepresentation. Foreigners from a neighboring country increase their underrepresentation by 4.6% points, other foreigners by 13.0% points. As for marital status, singles increase their underrepresentation by 3.8% points, and the divorced by 6.3% points. Including the face-to-face sample members increases the individual specific cooperation rate to 78.8% (table 4). Older people from 59 years on improve cooperation relatively less in the combined sample. The strongest improvements come from foreigners and especially those from a country other than a neighboring country, and younger and single people.

Table 4 Predicted probabilities during person recruitment phase

[average predicted probabilities from logit model]	Contact Teleph.	Contact All	Coop. Teleph.	Coop. All
16-30 years old	0.787	0.984	0.675	0.851
31-44 years old	0.770	0.981	0.640	0.833
45-58 years old	0.874	0.977	0.689	0.778
59-72 years old	0.949	0.989	0.709	0.749
73+ years old	0.948	0.988	0.603	0.652
Native Swiss or born in Switzerland	0.882	0.985	0.702	0.796
from a neighbor. country	0.767	0.972	0.573	0.755
from another country	0.660	0.977	0.433	0.729
single	0.834	0.977	0.631	0.762
married	0.874	0.987	0.705	0.806
widowed	0.807	0.966	0.636	0.765
divorced	0.776	0.980	0.569	0.739
Women	0.858	0.985	0.680	0.793
Men	0.843	0.981	0.659	0.781
McFadden Pseudo R-squared	0.119	0.019	0.032	0.020
Mean value all people	0.851	0.983	0.670	0.788

Data: SHP III (2013 refreshment sample, N (individuals) = 7,826).

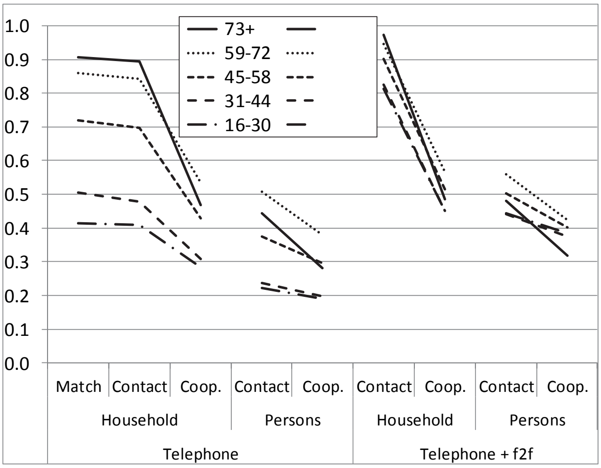
Looking at the representation bias in table 5, we note that the highest representation bias is due to noncontact errors in the telephone sample (0.087). (Additional) bias from noncooperation in the telephone sample is small (0.107). Bias from noncontact in the combined sample is again negligible (0.006) while (additional) bias from noncooperation is considerable (0.059).

When interpreting these findings, we must account for the fact that the final representation bias of the person groups is the sum of all the biases from the recruitment phases, i.e., from the household recruitment phase the bias from matching, noncontact, and noncooperation, and from the (subsequent) person recruitment phase the bias from noncontact and noncooperation. We illustrate this in figure 1 below using the example of old households aged 73 years or over.

Table 5 Representation bias during the person recruitment phase

[average predicted probabilities / mean value]	Contact Teleph.	Contact All	Coop. Teleph.	Coop. All
16-30 years old	0.925	1.001	1.007	1.080
31-44 years old	0.905	0.998	0.955	1.057
45-58 years old	1.027	0.994	1.028	0.987
59-72 years old	1.115	1.006	1.058	0.951
73+ years old	1.114	1.005	0.900	0.827
Native Swiss or born in Switzerland	1.036	1.002	1.048	1.010
from a neighbor. country	0.901	0.989	0.855	0.958
from another country	0.776	0.994	0.646	0.925
single	0.980	0.994	0.942	0.967
married	1.027	1.004	1.052	1.023
widowed	0.948	0.983	0.949	0.971
divorced	0.912	0.997	0.849	0.938
Women	1.008	1.002	1.015	1.006
Men	0.991	0.998	0.984	0.991
Standard deviation	0.087	0.006	0.107	0.059

Data: SHP III (2013 refreshment sample, N (individuals) = 7,826).



Data: SHP III (2013 refreshment sample, N (households) = 8,098, N (individuals) = 7,826).

Figure 1 Representation of households / persons by age due to reasons for nonobservation

Old households aged 73 years or over (solid line in figure 1) are overrepresented among the telephone matched households (+33.8%, see table 3), the telephone contacted households (+35.7%), and also among the contacted total sample (+9.4%). However, as they refuse more often these households are only slightly overrepresented among the telephone responding households (+16.8%), and well represented among the total responding households (-1.4%). Next, on the person recruitment level (table 5), old people are overrepresented among the telephone contacted people (+11.4%), while (still) well represented among the total contacted people (-1.2%). But after being asked to participate, they are underrepresented among the telephone respondents (-10.0%), and especially among the total respondents (-17.3%).

4 Deprivation: Telephone Versus Combined Sample

In this section we evaluate whether it is worth adding face-to-face households to the telephone households in terms of substantive variables, using deprivation as an example. We analyze regression coefficients from multivariate regressions with and without taking into account responding households without a telephone. We account for education level, age, the number of children under the age of 18 years in the household the number of adults in the household, and working status (full-time, part-time, retired, other). To this end, we analyze four deprivation variables:

- logarithm of household gross income (mean: 11.32)
- home ownership (49.8% of all households)
- a deprivation index, constructed as the number of items which the household cannot afford (car for private use (3.7%); savings into 3rd pillar (11.8%), dentist (2.7%), fresh fruit or vegetables (1%), and a room of one's own (2.1%))
- whether households were in arrears with their payments during the past 12 months (11.7%)

Table 6 shows regression coefficients of the face-to-face main- and interaction coefficients of the four regression models. Individual-level characteristics (education and working status) are taken from the household reference person. The models are controlled for the main effects of the interacted variables.

The sample sizes for income are smaller than the total respondent sample due to missing information. To test the effect of the survey mode on missing income, by means of a chi^2 test we find that these two variables are not significantly correlated (5% level). Although missing income is possibly affected in addition by mode

Table 6 Regression coefficients of face-to-face (F2F) dummies.

[beta-coefficients]	Ln Income (OLS)	Owner (logit)	Deprivat. (poisson)	Arrears (logit)
F2F main effect	-0.275**	-0.878*	0.982**	0.990*
F2F * education (11 categories)	-0.008	-0.031	-0.024	0.035
F2F * age (continuous)	0.002	0.020	0.010	0.012
F2F * number of children in household	-0.015	-0.151	0.008	-0.026
F2F * number of adults in household	0.034	-0.064	-0.009	-0.026
F2F * full-time employed	0.181**	0.453	-0.627**	-0.672*
F2F * part-time employed	0.060	0.627	-0.016	0.128
F2F * retired	-0.044	-0.411	-0.101	0.205
N	3,290	3,971	3,972	3,972

Data: SHP III (2013 refreshment sample, N (households) = 3,989 (740 f2f)). ** p<0.01, * p<0.05

selection effects, which we can only control for the variables at hand[6], this leads us to believe that mising income is independent of the mode. We find that the face-to-face households have a lower income, are less likely to be owners of their house, and suffer both from more deprivation and more payment arrears. The differences are substantial. For example, after controlling for all other variables in the model, face-to-face households have an 18.5% points lower probability of owning their house than telephone households. These findings are in line with the literature. As for interaction terms, only full-time employment plays a role: full-time employed face-to-face households have the same income as full-time employed telephone households (the sum of the face-to-face main effect and the face-to-face full-time interaction effect is statistically insignificant). The same is true for these households in terms of the deprivation index and the arrears. These results show that face-to-face households are poorer than telephone households on average, but that this does not hold for households with a full-time employed reference person.

[6] For example, in a logit model regressing missing income on the survey mode, the coefficient of the survey mode hardly changes if the (negative) effect of education is also accounted for.

5 Summary and Discussion

Some surveys add a second mode to the landline to reduce issues from undercoverage while nonresponse remains a problem. To reduce bias from nonobservation, the idea of a responsive fieldwork design has recently been put forward (Groves & Heeringa, 2006): differences between observed and nonobserved sample members can be reduced by adjusting fieldwork efforts. Knowing the reason for nonobservation by mode facilitates fieldwork decisions. For example, higher noncooperation from a certain population group in mode A may be acceptable if this group exhibits higher coverage and contact rates in mode B.

In this paper we analyze socio-demographic representation bias on the basis of the different reasons for nonobservation, using a mixed-mode survey where the landline is used for households with a listed number and face-to-face otherwise. Some findings stand out in our analysis. People from one-person households and those with small children at home, young adults, and foreigners are more difficult to match, while the opposite is true especially for those living in large households and in particular older people. Additional bias from noncontact is small. Existing bias tends to increase when trying to obtain cooperation, with the exception of young households, who cooperate more often and older households, who cooperate less often. Adding the face-to-face mode largely decreases the bias. Still, the underrepresentation of one-person households and foreigners increases with each step. During the recruitment of eligible individuals in cooperating households, noncontact can be largely decreased by adding the face-to-face sample. Otherwise, existing bias from the household recruitment phase remains constant, with – again – the exception of older people, who are easier to contact by telephone but cooperate to a lesser extent in both samples. Foreigners from a country other than a neighboring country (and thus not sharing one of the survey languages) are both difficult to contact and to convince to participate, especially in the telephone sample.

We model income and deprivation of responding households for the telephone and the combined telephone / face-to-face sample. The result shows that the telephone respondents are richer on average and suffer from less deprivation, which also proves the importance of including the face-to-face mode in terms of substantive survey variables.

To optimize fieldwork, our findings imply that different socio-demographic groups should be treated differently according to their selective reason for dropping-out. First, more effort should be invested for groups with a low matching probability (one-person households, households with young children, Italian speaking households, young households, and foreigners). It may be an idea to use additional data sources (Lipps et al., 2015), manual researches, postcards asking for contact information (e.g., Lipps & Kissau, 2012), or less sensitive algorithms to match names. With respect to obtaining cooperation, more older and foreigner households, especially those from countries not sharing one of the survey languages, fall out of the sample. These groups should be treated with special care. One idea would be to use ethnic or bilingual interviewers (Kappelhof, 2015; Laganà et al., 2013), to approach them face-to-face to facilitate communication in a foreign language, or to provide an extra incentive. In the face-to-face sample, making contact is more difficult with one-person households, with Italian speakers, in large municipalities, and with foreigners from a neighboring country. This is probably due to no one being at home at typical calling times. More calls at different times and on different weekdays can be attempted with these households. Among the cooperating households, again there is underrepresentation in one person households, those without children, Italian speakers, those in large municipalities, older people, and the married[7]. A further idea could be to use more successful interviewers to visits these households, or again to offer incentives.

During the person recruitment phase in the contacted telephone sample, young people are underrepresented, as well as foreigners and especially those from countries not sharing one of the survey languages, and divorced people. Noncontact has a small effect on bias in the total sample. As for noncooperation in the telephone sample, refusals are more prevalent among people aged 73 or over, and foreigners from countries not sharing one of the survey languages. In the face-to-face sample, people aged 59-72 years refuse more often. To reduce bias caused in the person recruitment phase, similar measures to the “critical” households during the household recruitment phase should be taken, with perhaps more “person-tailored” measures.

We here note some limitations of this paper. Evidently, bias can only be analyzed for the representativity of the socio-demographic variables available from the population register. While these variables reflect household at-home patterns and are suitable for analyzing noncontact, non-cooperation depends on social participation and interest in societal well-being (Stoop, 2005). Because socio-demographic variables are “correlates, not causes of the survey participatory behavior” (Groves & Couper, 1996, p. 81), other register variables could be matched to sample members. While this was successfully done in Northern European countries (e.g., Nordberg et al., 2001), experiences from other countries are in their infancy and still restricted to specific domains like employment (e.g., De Gregorio et al., 2014).

In addition, the composition of the samples during fieldwork of course depends on the effort made in the previous steps, including the sources used to match telephone numbers and also the algorithm used to match telephone numbers. Similarly, effects from one mode depend on effort from another mode. As far as they go the results are therefore not easily generalizable. Our research is just one example to be used to shed light on the characteristics of sample members lost at the different

[7] Note that married people are overrepresented from the other reasons for nonobservation.

steps during the survey recruitment phases in a mixed mode survey, and to show which steps require special care to keep socio-demographic representation bias at a reasonable level. More comparable mixed-mode surveys are needed to assess the fieldwork quality in the different modes, to find an optimal resource allocation for the modes, and to balance selective losses due to the different reasons for nonobservation.

References

AAPOR (The American Association for Public Opinion Research). (2011). Standard Definitions: *Final Dispositions of Case Codes and Outcome Rates for Surveys*. 7th edition. http://www.aapor.org/AAPORKentico/AAPOR_Main/media/MainSiteFiles/Standard-Definitions2011_1.pdf (accessed Feb 16, 2015).

Blumberg, S. & Luke, J. (2014). *Wireless substitution: Early release of estimates from the National Health Interview Survey*. July–December 2013. accessed 9AUG2014: http://www.cdc.gov/nchs/data/nhis/earlyrelease/wireless201407.pdf.

Brick, J., Williams, D., & Montaquila, J. (2011). Address-based sampling for subpopulation surveys. *Public Opinion Quarterly*, nfr023. doi: 10.1093/poq/nfr023.

Busse, B. & Fuchs, M. (2012). The components of landline telephone survey coverage bias. The relative importance of no-phone and mobile-only populations. *Quality & Quantity,* 46(4), 1209-1225.

Cobben, F. (2009). *Nonresponse in sample surveys: methods for analysis and adjustment*. PhD Thesis, Statistics Netherlands.

Cobben, F. & Bethlehem, J. (2005). Adjusting Undercoverage and Nonresponse Bias in Telephone Surveys. *Discussion paper 0506, Statistics Netherlands*, Voorburg/Heerlen.

De Gregorio, C., Filipponi, D., Martini, A., & Rocchetti, I. (2014). A comparison of sample and register based survey: the case of labour market data. (Q2014 Conference). Vienna: Statistik Austria.

Deming, W. (1947). Some Criteria for Judging the Quality of Surveys. *The Journal of Marketing,* 12(2), 145-157.

de Leeuw, E. (2005). To mix or not to mix data collection modes in surveys. *Journal of Official Statistics,* 21(2), 233-255.

De Vitiis, C. & Righi, P. (2011). Evaluations on list undercoverage bias and possible solutions: the case of ISTAT CATI survey» Trips, holidays and daily life. *Rivista di statistica ufficiale*, 13(2-3), 5-19.

Ernst Stähli, M. (2012). Telephone Surveys in Switzerland: Spotlight. In M. Häder, S. Häder & M. Kühne (Eds.), *Telephone Surveys in Europe: Research and Practice* (pp. 25-36): Springer.

Gordoni, G., Oren, A., & Shavit, Y. (2010). Coverage bias in telephone surveys in Israel. *Field Methods*, 1525822X10387573.

Groves, R. & Couper, M. (1996). Contact-Level Influences on Cooperation in Face-to-Face Surveys. *Journal of Official Statistics,* 12, 63-83.

Groves, R. & Heeringa, S. (2006). Responsive Design for Household Surveys: Tools for Actively Controlling Survey Errors and Costs, *Journal of the Royal Statistical Society - Series A*, 169(3), 439-457.

Häder, S., Häder, M., & Kühne, M. (2012). Introduction: Telephone Surveys in Europe. In M. Häder, S. Häder & M. Kühne (Eds.), *Telephone Surveys in Europe: Research and Practice* (pp. VII-XIII): Springer.

Heckel, C. & Wiese, K. (2012). Sampling Frames for Telephone Surveys in Europe. In M. Häder, S. Häder & M. Kühne (Eds.), *Telephone Surveys in Europe: Research and Practice* (pp. 103-119): Springer.

Joye, C. (2012). SRPH-Castem. FORS – SFSO workshop, June 21. Neuchâtel.

Joye, D., Pollien, A., Sapin, M., & Ernst Stähli, M. (2012). Who Can Be Contacted by Phone? Lessons from Switzerland. In M. Häder, S. Häder & M. Kühne (Eds.), *Telephone Surveys in Europe: Research and Practice* (pp. 85-102): Springer.

Kappelhof, J. W. (2015). Face-to-Face or Sequential Mixed-Mode Surveys Among Non-Western Minorities in the Netherlands: The Effect of Different Survey Designs on the Possibility of Nonresponse Bias. *Journal of Official Statistics*, 31(1), 1-30.

Kühne, M. & Häder, M. (2012). Telephone Surveys via Landline and Mobile Phones: Mode Effects and Response Quality. In M. Häder, S. Häder & M. Kühne (Eds.), *Telephone Surveys in Europe: Research and Practice* (pp. 229-246): Springer.

Laganà, F., Elcheroth, G., Penic, S., Kleiner, B., & Fasel, N. (2013). National minorities and their representation in social surveys: which practices make a difference? *Quality & Quantity*, 47(3), 1287-1314.

Link, M., Battaglia, M., Frankel, M., Osborn, L., & Mokdad A. (2007). Reaching the U.S. cell phone generation. Comparison of cell phone survey results with an ongoing landline telephone survey. *Public Opinion Quarterly*, 71(5), 814–839

Lipps, O. & Kissau, K. (2012). Nonresponse in an Individual Register Sample Telephone Survey in Lucerne (Switzerland). In M. Häder, S. Häder & M. Kühne (Eds.), *Telephone Surveys in Europe: Research and Practice* (pp. 187-208): Springer.

Lipps, O., Laganà, F., Pollien, A., & Gianettoni, L. (2013). Under-representation of foreign minorities in cross-sectional and longitudinal surveys in Switzerland. In J. Font & M. Méndez (Eds.), *Surveying Ethnic Minorities and Immigrant Populations: Methodological Challenges and Research Strategies* (pp. 241-267): Amsterdam University Press.

Lipps, O. & Pekari, N. (2016). Sample representation and substantive outcomes using web with and without incentives compared to telephone in an election survey. *Journal of Official Statistics*, 32(1).

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

Lohr, S. & Brick, J. (2014). Allocation for Dual Frame Telephone Surveys with Nonresponse. *Journal of Survey Statistics and Methodology*, smu016. doi: 10.1093/jssam/smu016

Lynn, P. & Kaminska, O. (2012). Factors affecting Measurement Error in Mobile Phone Interviews. In M. Häder, S. Häder, & M. Kühne (Eds.), Telephone Surveys in Europe: Research and Practice (pp. 211-228): Springer.

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(1), 67-82.

Mohorko, A., de Leeuw, E., & Hox, J. (2013). Coverage bias in European telephone Surveys: Developments of landline and mobile phone coverage across countries and over time." Survey Methods: Insights from the Field. DOI:10.13094/SMIF-2013-00002.

Nordberg, L., Pentillä, I., & Sandstöm, S. (2001). A study on the effects of using interview versus register data in income distribution analysis with an application to the Finnish ECHP Survey in 1996. Statistics Finland working paper, (1).

Olson, K. (2007). An Investigation of the Nonresponse Error - Measurement Error Nexus. University of Michigan, Ann Arbor.

Peytchev, A., Carley-Baxter, L., & Black, M. (2011). Multiple Sources of Nonobservation Error in Telephone Surveys: Coverage and Nonresponse. Sociological Methods & Research 40(1), 138-168.

Sala, E. & Lillini, R. (2014). The impact of unlisted and no-landline respondents on non-coverage bias. The Italian case (No. 2014-16). Institute for Social and Economic Research.

Safir A. & Goldenberg K. (2008). Consumer Expenditure Survey Program: Telephone Effects in the Consumer Expenditure Quarterly Interview Survey. US Bureau of Labor Statistics internal paper

Schneiderat, G. & Schlinzig, T. (2012). Mobile-and landline-Onlys in dual-frame-approaches: effects on sample quality. In M. Häder, S. Häder, & M. Kühne (Eds.), Telephone Surveys in Europe: Research and Practice (pp. 121-143): Springer.

Schouten, B., Cobben, F., & Bethlehem, J. (2009). Indicators for the representativeness of survey response. Survey Methodology, 35(1), 101-113.

Stoop, I. (2005). The Hunt for the Last Respondent: Survey Nonresponse in Sample Surveys. The Hague: Social and Cultural Planning Office.

Von der Lippe, E., Schmich, O., & Lange, C. (2011). Advance letters as a way of reducing non-response in a national health telephone survey: Differences between listed and unlisted numbers. Survey Research Methods 5(3), 103-116.

methods, data, analyses | Vol. 10(1), 2016, pp. 25-46 DOI: 10.12758/mda.2016.002

Ballpoint Pens as Incentives with Mail Questionnaires – Results of a Survey Experiment

Marcus Heise

Martin-Luther-Universität Halle-Wittenberg

Abstract

The results of meta-analyses carried out in studies designed to examine the effectiveness of different types of incentives routinely applied in numerous Anglo-American survey research projects to secure higher response rates have led to the following general conclusion: monetary incentives (i.e. cash) perform better than non-monetary incentives (e.g. small-sized gifts). Comparatively few such studies have been conducted in Germany and they cover only a rather limited range of monetary or money-related incentives. The current paper seeks to go beyond such limitations by testing the assumption that, in the case of surveys covering rather more intimate and morally relevant issues, less expensive non-monetary incentives might be quite effective in increasing the response rate. This study was carried out within the context of a larger research project ("Self-Expressive Forms and Functions of Personal Conscience in Every-Day Life") conducted at the University of Halle-Wittenberg and based on a random sample of 4000 people drawn from the city registry in Halle (Saale). These individuals were then randomly assigned to a control group (without an incentive) or a test group (presented with a ballpoint pen, i.e. a non-monetary incentive), each made up of 2000 people. Our data analysis showed that the gift of a ballpoint pen affected the willingness to respond, the speed of the response, and the completeness of the surveys that were returned. Furthermore, no negative effects were detected on the composition of the sample that was obtained.

Even though the effect of the non-monetary incentive was revealed to be fairly small in comparison with the effect of monetary incentives observed in other studies, the use of small in-kind incentives can be advantageous in certain survey designs. Inexpensive, non-monetary incentives may serve as a possible substitute for follow-up contact in study-designs that face a variety of limitations such as budget-restrictions or regulations on data protection.

Keywords: mail questionnaire, experiment, non-monetary incentive, response rate, response speed

1 Introduction: Monetary or Non-monetary Incentives with Mail Questionnaires?

The completeness of the responses that were returned and the response rate are central quality markers of postal surveys. The Total Design Method (TDM, Dillman, 1978; Dillman, 2000) recommends sending small monetary or non-monetary incentives with the survey in order to increase willingness to respond on the part of the subjects. Such incentives have been in use in conjunction with surveys since the 1930's (see Armstrong, 1975, p. 116 or Wotruba, 1966, p. 398) and practically no other issue has received as much attention in studies on survey methodology (for an overview see Hippler, 1988, p. 245). A vast number of methodological experiments have been conducted, especially in the Anglo-American area, to investigate the effects of monetary and in-kind incentives on the willingness to respond to surveys.[1] These studies conclude that incentives not only increase the general response readiness of subjects, but that they also positively impact on response speed, the make-up of the sample thus obtained, and the completeness of responses to open and closed questions (for an overview, see Berger, 2006).

To achieve these effects, it is beneficial if the incentive is sent together with the mail questionnaire. Even though incentives that are paid out upon successful completion of the survey can have positive effects on the response rate (see e.g. Singer et al. 1999, p. 223), unconditional incentives are generally considered to be comparatively more effective in increasing the odds of response (Church 1993; Auspurg & Schneck, 2014; an example in which conditional incentives outperform unconditional incentives is given by Castiglioni, Pforr, & Krieger, 2008). Meta-analyses from the Anglo-American language area demonstrated that monetary incentives are much better at increasing the return rate for postal surveys than non-monetary incentives (Church, 1993, p. 75; Fox, Crask & Kim, 1988, p. 485; Goodstadt, Chung, Kronitz, & Cook, 1977; Simmons & Wilmot, 2004, p. 3; Yu & Cooper 1983, p. 40; for telephone and personal interviews, see Singer et al., 2000).

[1] The meta-analysis in Singer, van Hoewyk, Gebler, Raghunathan, and McGonagle (1999, p. 219) identified more than 1000 reviews that dealt with the topics of incentives, survey experiments, and response rates.

Direct correspondence to
Marcus Heise, Institut für Medizinische Epidemiologie, Biometrie und Informatik (Sektion Allgemeinmedizin), Medizinische Fakultät, Martin-Luther-Universität Halle-Wittenberg, 06097 Halle (Saale)
E-mail: Marcus.heise@medizin.uni-halle.de

Acknowledgements: The author would like to thank two anonymous reviewers as well as Steffen Jakobs, Christian Rademacher, Sylvia Terpe and Helmut Thome for their valuable suggestions. Translation: Adam Smant

Non-monetary incentives of low financial value proved to be particularly ineffective in American surveys.

Comparatively few reviews on this topic exist in the German language area, where literature research carried out by the author revealed that only the effects of monetary or near-monetary incentives had been examined. In these studies, money was either sent directly with the mail questionnaire (see Becker, Imhof, & Mehlkop, 2007; Blohm & Koch, 2013; Börsch-Supan, Krieger, & Schröder, 2013; Castiglioni, Pforr, & Krieger, 2008; Fick & Diehl, 2013; Mehlkop & Becker, 2007; Stadtmüller, 2009) or incentives were used that had a clear corresponding monetary value, such as phone cards or stamps (see Arzheimer & Klein, 1998; Diekmann & Jann, 2001; Harkness, Mohler, Schneid & Christoph, 1998; Porst, 1999; Reuband, 1999). Schröder et al. (2013), commissioned by the German Socio-Economic Panel, showed that conditional monetary incentives were more efficient than a lottery ticket in increasing the response rate. The work carried out by Anja Göritz and colleagues provides evidence for the effects of prepaid cash (van Veen, Göritz, Sattler, 2015), cash lotteries (Göritz & Luthe, 2013; Göritz, 2006) and promised cash (Göritz, Wolff & Goldstein, 2008) in web surveys.

No German studies were found that examine the effects of non-monetary incentives without a clear corresponding monetary value. However, Singer et al. (1999, p. 219) and Harkness et al. (1998, p. 205) point out that the expected reciprocity of a monetary incentive is culturally variable. Because of this, it is questionable whether the superiority of monetary incentives over in-kind incentives, as demonstrated in American and Canadian studies, can be transferred to other cultures. However, this superiority is implied in the literature for the German language area: Stadtmüller and Porst (2005, p. 8) recommend incentives with a clear monetary value as opposed to non-monetary incentives. Mehlkop and Becker (2007, p. 14) agree with this recommendation based on the fact that, in comparison with monetary incentives, in-kind incentives like ballpoint pens are valued differently by different subjects and therefore only appeal to certain groups (see also Little & Engelbrecht, 1990). The current study is the first of its kind that investigates the effects of a non-monetary incentive offered to a sample drawn from the population of a German city, challenging these assumptions and addressing this rather under-researched aspect. In the context of the diminishing response rates in surveys (Auspurg & Schneck, 2014; Börsch-Supan, Krieger & Schröder, 2013), the paper addresses an important issue. The results suggest that sending a ballpoint pen along with the mail questionnaires can yield meaningful effects on response behavior.

2 Theoretical Background, Past Research on Non-monetary Incentives and Hypotheses

2.1 Theoretical Background

The positive effects of monetary or in-kind incentives on an individual's readiness to help have been discussed in several disciplines, for example in economics (see Falk 2007) and in psychology, where this mechanism is called the "Feeling-Good-Effect" (Levin & Isen 1975). In sociological methodological research, the effects of incentives on response behavior in postal surveys are usually based on four theoretical approaches:

Gouldner's (1960, pp. 171-175) approach on the effect of incentives is based on a universal norm of reciprocity that urges a person to help others from whom this person has received help or material goods.

Dillman's Social Exchange Theory (2000) stresses, with a reference to Blau (1964) and Thibaut and Kelly (1959), that a subject's trust in the researcher is the most important factor in increasing their readiness to participate.

According to the rational choice perspective (e.g. Singer, 2011), respondents will decide to answer a survey when the perceived benefits outweigh the costs of participation in the survey. While the benefits of participation can be related to intrinsic motives of the respondents, to the perceived usefulness of the survey (to oneself or others), as well as to incentives associated with participation, the perceived costs can include the time required to answer the questions or can be privacy-related.

The Leverage-Saliency-Theory (Groves & McGonagle, 2001; Groves, Singer & Corning, 2000) additionally emphasizes that interviewer behavior can influence the saliency of specific benefits and costs of survey participation, dropping the assumption that the effects of specific aspects of a survey are constant across respondents.

2.2 Response Rate

Against this theoretical background, it is to be expected that a ballpoint pen used as an in-kind incentive could have a positive impact on the response rate. As stated above, there are no published methodological experiments describing the effects of a ballpoint pen as an incentive on a German sample. In the USA, Houston & Jefferson (1975) studied the effects of ballpoint pens as incentives in an American sample of vehicle buyers, as well as the difference between personalized and non-personalized letters in a 2x2 design. The ballpoint pen increased the response rate in the personalized group by 6% and by a remarkable 31% in the non-personalized group. In an American postal survey, Hansen (1980, p. 79) compared the effect of a monetary incentive with that of a ballpoint pen: while the response rate in people who had received a quarter along with the questionnaire reached 39%, the response rate for the group that had received a ballpoint pen of the same value was only 22%, and the response rate for the control group (with no incentives) was 14%. In two Dutch studies (Nederhof, 1983), ballpoint pens increased the response rate from 20.6% to 31.8% and then from 27.3% to 33.8% after the first mailing. However, after further reminder letters were sent, the differences were balanced out to become non-significant, leading to the hypothesis that non-monetary incentives like ballpoint pens only influence response behavior for a short time and show no long-lasting effects.

These results are, however, based on non-German samples. Furthermore, these study designs used non-personalized letters and, at most, one follow-up mailing. Nonetheless, in the context of these findings, the following two hypotheses can be put forward:

H_{1a}: Survey participation is higher if the respondents receive a ballpoint pen together with the questionnaire.

H_{1b}: The increase in the response rate caused by the incentive is only short-lived and diminishes with the number of days after its mailing.

2.3 Sample Composition

The thesis that in-kind incentives are especially likely to be valued differently by various socio-demographic groups (Mehlkop & Becker, 2007) raises the question on how non-monetary incentives might affect sample composition. Incentives could either lead to the over-representation of certain groups or might encourage otherwise under-represented respondents to take part in the survey, thereby reducing non-response bias (Singer & Ye, 2013). Divergent results are found in the literature on the effects of incentives on sample composition: for example, Arzheimer (1998, p. 24), Nederhof (1983, p. 106) and Stadtmüller (2009, p. 180) explicitly deny a gender-specific incentive effect. In contrast, Harkness et al. (1998, p. 216) state that women over the age of 65 were especially likely to respond to a lottery-incentive. Mehlkop & Becker (2007) found slightly (but not significantly) stronger effects of a monetary incentive on women. Investigating the effect of a monetary incentive, Baron et al. (2008) were more likely to contact respondents with a higher socio-economic status. More recently, Blohm and Koch (2013) and Martin et al. (2014) have concluded that the value of monetary incentives does not have a substantial effect on sample composition. In contrast, Börsch et al. (2013) and Medway (2012) found that a monetary incentive significantly affected the age composition of the sample that was collected. Furthermore, in the study by Börsch et al. (2013), retired respondents and respondents without university degrees were less likely to react to the monetary incentive. Simmons and Wilmot (2004) provide evidence that incentives can influence the composition of the achieved sample with respect to ethnic affiliation, income and education. A systematic review from Singer and Ye (2013) concludes that few studies have found significant effects of incentives on sample composition. However, no results on the effect of ballpoint pens on sample composition were found in the literature.

H_{2a}: The distribution of socio-demographic characteristics differs between the samples collected in the incentive and the control conditions.

H_{2b}: Sending a ballpoint pen does not influence the composition of the sample that is collected.

2.4 Response Speed

While it is clear that sending monetary incentives also has a positive impact on response speed in the German language area (see Becker et al., 2007; Berger 2006; Diekmann 2001; Stadtmüller 2009), there are differing results on the effects of ballpoint pens. Houston and Jefferson (1975, p. 400) found a significantly higher cumulative response rate in the trial group, that ceased one week after the questionnaire was mailed. Nederhof (1983, p. 106) also confirmed that receiving a ballpoint pen led to increased response speed among the subjects. This may be explained by the fact that a ballpoint pen has a direct relationship to the questionnaire in that it is an instrument that can be used for the completion of the survey. This may strengthen the subject's motivation to begin right away with answering the questions presented to them. Conversely, Hansen (1980, p. 81) indicated a clear decrease in response speed, with a ballpoint pen nearly doubling the time (from an average of 8 days to an average of 15 days) required to complete the questionnaire. These findings lead to the formulation of two competing hypotheses:

H_{3a}: Sending a ballpoint pen together with the questionnaire increases the response speed.

H_{3b}: Sending a ballpoint pen together with the questionnaire slows down the response speed or has no effect.

2.5 Item Non-response

The question on whether or not a non-monetary incentive with low monetary value could have a similar effect on the response rates for postal surveys as monetary incentives (see Stadtmüller, 2009, p. 167) is especially important as reasons can be found to consciously decide to send such items instead of money. Several studies have shown that the response effect of incentives increases linearly with their monetary value (Church, 1993, p. 73; Furse & Stewart 1982, p. 377; James & Bolstein, 1990, p. 351; Jobber, Saunders, & Vince-Wayne, 2004, p. 23; Singer et al., 1999, p. 223; Yu & Cooper, 1983), at least until they approach certain thresholds

(Armstrong, 1975, p. 115; Berger, 2006; Fox et al., 1988, p. 485; Linsky, 1975, p. 8; Martin, Abreu, & Winters, 2001, p. 274; Mizes, Fleece & Roos, 1984, pp. 797-799; Warriner et al., 1996, p. 549). However, it has also been shown that incentives that are too valuable can demotivate subjects for further studies (Lynn, 2001), lead subjects to a "quid-pro-quo" thought process over time (Martin et al., 2001, p. 280), or provoke a reactive response behavior (Hansen, 1980). Data quality can suffer in cases where the individual views the money sent to them as being unwarranted or pushy (see Barón et al., 2008, p. 11; Trussell & Lavrakas, 2004, p. 361). This is explained by Stadtmüller and Porst (2005, p. 5) via an interpretive process in which a subject addressed in this way no longer views the incentive as a symbolic gesture, but instead as a form of financial pre-payment, calling for participation in an economic exchange instead of an exchange based on a cultural norm of reciprocity (see Trussell & Lavrakas, 2004, p. 364).

Following this "case for smaller incentives" (Stadtmüller, 2009, p. 170), one may assume that a ballpoint pen could prove to be quite an advantageous incentive within the context of postal surveys that deal with rather complex or intimate questions. In comparison to monetary incentives, such incentives are less obtrusive and therefore less likely to produce a negative response on the part of the subject. A ballpoint pen as an incentive, so the assumption, will retain its symbolic meaning and thus, due to its low financial value, not run the risk of being seen as a form of payment (see Singer et al. 1999, p. 222).

Conflicting hypotheses are found in the existing literature on the effect of incentives on the readiness to respond to more or fewer questions. James and Bolstein (1990), Houston and Ford (1976), Shettle and Mooney (1999), as well as Wotruba (1966), describe a positive effect of monetary incentives on the completeness of answers in American studies. Singer and Ye (2013) conclude that further research is needed on this question. Stadtmüller (2009, p. 182) found no evidence for a positive effect of incentives on data quality in a German sample. In contrast, Davern, Rockwood, Sherrod, & Campbell (2003, p. 140) suspect that incentives can animate undecided subjects to participate in a superficial manner and to refrain from answering certain questions. With respect to nonmonetary incentives, Hansen (1980, 81) concurs, stating that the ballpoint pen sent in his study had a negative impact on the quality and completeness of responses to open-answer style questions. Furthermore, sensitive items may be more susceptible to incentive effects than nonsensitive items (Medway, 2012). With regard to sensitive questions, Tzamourani and Lynn (1999) showed that a monetary incentive increased non-response, while Medway (2012) and Krenzke et al. (2005) could not confirm a negative effect on data quality.

H_{4a}: Sending a ballpoint pen together with the questionnaire has no effect on item non-response.

H_{4b}: Sending a ballpoint pen together with the questionnaire has a negative impact on data quality and increases item non-response.

2.6 Cost-effectiveness

A relevant question addresses the additional costs associated with the use of an incentive in relation to the gain in response rate. Depending on the nature and size of the incentive, the cost per completed interview can increase (see for instance Börsch et al., 2013) or decrease (see for instance Jobber et al., 2004; Medway, 2012). Sending a hard object like a ballpoint pen can be associated with additional mailing cost, thus increasing the cost per completed interview. However, in comparison with phone cards or a banknote, a ballpoint pen is an inexpensive gift, which makes it especially well-suited for surveys with a large sample size.

H_{5a}: Sending a ballpoint pen together with the questionnaire increases the cost per completed interview.

H_{5b}: Sending a ballpoint pen together with the questionnaire decreases the cost per completed interview.

3 Method and Design of the Experiment

The experiment was carried out within the context of the research project "Self-Expressive Forms and Functions of Personal Conscience in Every-Day Life"[2] conducted at the Martin-Luther University of Halle-Wittenberg. The sample of 4000 subjects was taken, using a stratified randomization approach, from the registry of inhabitants of the city of Halle (Saale), which is home to 230,000 residents. The twelve page questionnaire consisted of 118 closed and five open-ended questions that dealt with personal experiences of shame and guilt in everyday life, moral values and pangs of conscience. In the preliminary test, respondents took between 45 and 90 minutes to complete all the questions. R*espondent burden was fairly high,* given the scope of the questionnaire and the intimate and emotionally stressful nature of the questions. *The study was carried out from May 2012 to September 2012*. In total, 1166 respondents aged 17 to 94 years (mean: 48.5 years; SD: 18.8) answered the survey.

The sample of 4000 subjects was randomly partitioned into a control group and a test group, each comprising 2000 subjects. The members of the test group received a plastic ballpoint pen (worth 21 eurocents) along with the questionnaire.

[2] The project was funded by the German Research Foundation (DFG (TH 260/7-1)) from April 2011 to April 2014 and was led by Prof. Dr. Helmut Thome.

The internet address of the project was printed in one color on the pen. In addition, all subjects were informed of monetary prizes totaling 1,500 Euro that were to be raffled off among the respondents who returned their questionnaires.[3]

The design of the survey was limited to a single follow-up action (possible skewing of the sample discussed by Hippler, 1985, p. 50). The subjects had to be assured of absolute anonymity as the mail questionnaire was characterized by several time-consuming and particularly intimate questions. Therefore, the questionnaires were not numbered and did not display any other identification markers. Because of this, it was impossible to know which subjects had already completed the questionnaire and which subjects needed a reminder. In order to assign the returned questionnaires to the experimental conditions, a marker was used that was not visible to the respondents: the incentive group received a questionnaire with the headline printed in bold, while the headline was underlined in the control group. Due to financial considerations, only one follow-up letter was sent out four weeks after the original questionnaires had been included in the gross samples.

Our survey design deviated in a further point from the TDM recommendations and all studies known to us that assess the effects of incentives on response behavior: the ballpoint pen sent with the questionnaire was not explicitly referred to as a small "thank you" gift. This was done to avoid provoking any adverse reactions – especially with respect to the topic of the survey "Moral and conscience in everyday life".

3.1 Results: Effects of the Ballpoint Pen on Response Rate

The effect of the ballpoint pen on the readiness to participate in the survey is summarized in Table 1. According to the AAPOR standard definition, RR2, the response rate was calculated by dividing the number of returned surveys (complete and partial) by the sum of returned surveys, refusals, non-contacts and all cases of unknown eligibility. The response rate reached 26.2% in the control group and 30.9% in the test group (recipients of the ballpoint pen). This difference of 4.7 percentage points between the response rate in the control and test groups proved to be statistically significant (C=0.052; p<0.05).

It is worth mentioning that the pen only affected the willingness to respond in subjects who responded before the reminder was sent out: there was no discernible difference in the response behavior between the control and test group after the reminder was sent. The increase in the response rate caused by the ballpoint pen was therefore only short-term.

[3] Both the control and test groups were assured of participation in the raffle and the effects of these potential prizes are therefore ignored in the discussion of our results.

Table 1 Number of returned questionnaires and response rates in a postal survey conducted in the city of Halle / Saale on the theme "Morality and Conscience in Life Today" for the control group and the test group that received an ballpoint pen (absolute values; response rate in parentheses)

	Gross Sample	Absentees	Eligibles	Participation before reminder	Participation after reminder	Net Sample
Control Group	2000	54	1946	442 (22.1 %)	83 (4.15 %)	525 (26.25 %)
Experimental Group: Ballpoint Pen	2000	53	1947	538 (26.9 %)	81 (4.05 %)	619 (30.95 %)

Contingency coefficient (for the total inquiry period) C = 0.052

3.2 Results: Effects of the Ballpoint Pen on Sample Composition

Figure 1 shows the response rate in the control and test groups according to age and gender. While male subjects between 62 to 76 years showed the most prominent reaction to the incentive, the ballpoint pen affected women between 32 and 61 years of age most strongly: the response rate increased by about 6 percentage points (from 26.8% to 33.3%) in women aged 32 to 46 and by about 9 percentage points (from 30.0% to 38.9%) in women aged 47 to 61.

In summary, Figure 1 displays three results: Firstly, the effect of the ballpoint pen on the willingness to participate in the postal survey varies dependent on age. However, this relationship is not a monotone function. Secondly, an interaction between age and gender on the effectiveness of the incentive is visible, even though these effects are not significant and cannot be interpreted contextually. The incentive in this study, for example, shows a comparatively weak effect on women over the age of 62. Thirdly, Figure 1 suggests that women were more likely than men to be motivated by the ballpoint pen to take part in the survey. This hypothesis was tested using logistical regression and the results are presented in Table 2. Women completed the survey significantly more frequently than men, both in the control and incentive group. The fourth column in Table 2 shows that women tended to respond more strongly to the incentive than men. Furthermore, the ballpoint pen was especially effective in raising the response rate among respondents aged 32 to 46 years and 62 to 76 years. However, none of these interaction effects proved to be significant.

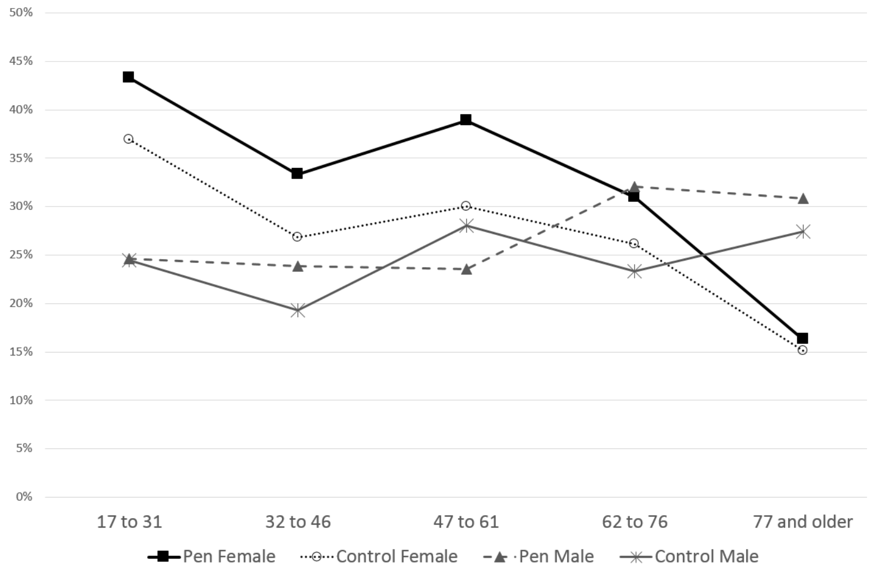


Figure 1 Response rate in the control and test group based on age and gender

Table 3 compares the composition of the sample collected with the results of a representative survey of local residents from the City of Halle / Saale (Harm & Jaeck, 2013), taking religious affiliation, education and employment status into consideration. Data analysis is limited to 1137 cases as no information was available about the non-responders in the current study. Respondents with a religious affiliation are slightly over-represented in the incentive condition, whereas the distribution within the control condition comes closer to the results reported by Harm and Jaeck (2013). Regarding the school-leaving certificate, respondents with a university entrance exam are noticeably over-represented in both experimental conditions, which might be due to the topic of the survey. In comparison to the reference study, the percentage of respondents with a technical baccalaureate, a university entrance exam, a Master's certification and a university degree is slightly higher in the incentive condition. Furthermore, unemployed respondents are marginally under-represented in the sample that received a ballpoint pen. In summary, the use of the ballpoint pen as an incentive did not substantially alter the composition of the sample. Furthermore, none of the differences between the control and incentive conditions proved to be significant in bivariate analyses.

Table 2 Logistic regression of response behavior on gender and age in the incentive and control condition (betas, standard errors in parenthesis)

	Control	Incentive	Incentive vs. Control
Gender			
Male (Reference)			
Female	0.217* (0.104)	0.384*** 0.099)	0.167 (0.143)
Age			
17 to 31 (Reference)			0.065 (0.167)
32 to 46	-0.410* (0.158)	-0.264 (0.150)	0.146 (0.218)
47 to 61	-0.095 (0.145)	-0.136 (0.142)	-0.041 (0.202)
62 to 76	-0.313* (0.151)	-0.141 (0.145)	0.172 (0.209)
77 and older	-0.678** (0.200)	-0.692 (0.193)	-0.014 (0.279)
Constant	-0.914*** (0.120)	-0.849*** (0.117)*	
n	2000	2000	
Cox & Snell Pseudo- R²	0.01	0.014	

Note: #: p<0,1; *: p < 0,05; **: p < 0,01; ***: p < 0,001

Table 3 Composition of the sample collected compared with a survey of local residents from the city of Halle / Saale (Harm & Jaeck, 2013)

	Survey of local residents 2012	Current study	Current study: incentive condition	Current study: control condition
Religious affiliation				
none	79.6%	78.7	77.4	80.1
Catholic	5.1%	5.4	5.9	5.2
Protestant	13.1%	13.9	14.4	13.2
Christian Congregational Chapel	1.5%	1.2	1.5	1
Non-Christian	0.8%	0.7	0.8	0.6
n	2780	1134	611	523
School leaving certificate				
In school education	0.3%	0.6%	0.3%	1.0%
Without a school-leaving qualification	1.2%	1.0%	1.1%	0.8%
Lower secondary school qualification	15.3%	12.0%	11.7%	12.3%

	Survey of local residents 2012	Current study	Current study: incentive condition	Current study: control condition
Secondary education (ISCED level 2)	38.0%	31.9%	30.9%	33.0%
Technical baccalaureate	13.7%	13.0%	13.4%	12.6%
University entrance exam	31.4%	41.5%	42.5%	40.4%
n	2729	1136	614	522
Vocational training				
None, or still in vocational training	10.4%	10.5%	10.0%	11.1%
Completed vocational training	45.2%	39.6%	39.0%	40.0%
Master certification	5.1%	8.9%	9.6%	8.1%
Technical college degree	14.9%	16.5%	16.5%	16.5%
University degree	24.4%	24.5%	25.0%	24.0%
n	2824	1134	613	521
Employment status				
Full-time	36.6%	35.7%	35.5%	35.9%
Part-time	8.4%	9.0%	9.3%	8.6%
Student	8.2%	12.0%	11.4%	12.8%
In vocational training	1.1%	2.0%	2.4%	1.5%
Irregularly employed	0.6%	2.5%	3.0%	2.1%
Unemployed	5.2%	4.2%	3.7%	4.8%
Retired / on leave	36.6%	30.3%	30.8%	29.6%
Military service or alternative service	0.2%	0.6%	0.5%	0.6%
Housewife / househusband	0.7%	0.8%	0.7%	1.0%
Parental leave	1.0%	1.3%	1.5%	1.1%
Not employed for other reasons	1.4%	1.7%	1.5%	1.9%
n	2861	1137	614	523

3.3 Results: Effects of the Ballpoint Pen on Response Speed

Figure 2 shows the cumulative survival odds for the return of the survey, which illustrates the effect of the ballpoint pen on response speed.

After the ninth day, the graph for the trial group approaches the abscissa more rapidly than for the control group and the effect of ballpoint pen thus led to the questionnaire being returned more quickly. The corresponding Log-Rank Test[4] showed

[4] This study used the method of calculation described by Bland & Altmann (2004).

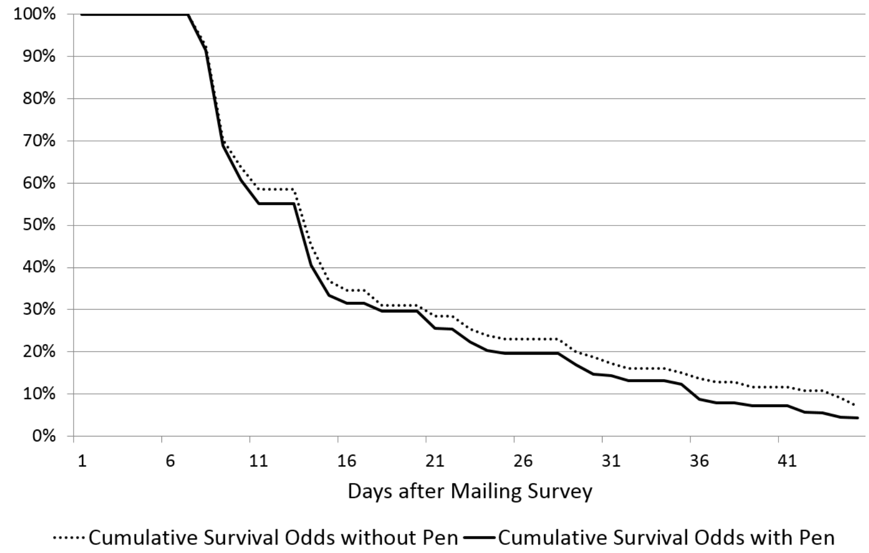


Figure 2 Cumulative survival odds (in percent) for the return of the survey with and without ballpoint pen

this effect to be significant (p<0.05). In this study, the inclusion of the ballpoint pen shortened the average time to the return of the questionnaire from 19.1 to 17.9 days.

3.4 Results: Effects of the Ballpoint Pen on Non-response to the Item Using Open Answer Examples

In this survey, the ballpoint pen had no observable effect on response behavior to closed questions. However, the questionnaire contained five particularly intimate open-answer style questions, where subjects were asked to describe stirrings of conscience or situations where they felt indignation, shame, or guilt. Table 4 includes the results of a multinomial regression that predicted the number of complete answers to these questions. Model 1 shows no significant effect through the incentive for the women's reference group. In the control condition, men were more likely than women to answer only a portion of the open questions in the questionnaire. The significant interaction effect between incentive and gender shows that men who received a ballpoint pen had an increased tendency to refuse to answer any of the five open-answer questions. This gender-specific negative effect of the incentive on the readiness to respond to open-answer questions is also demonstrated when corrected for the effects of education and employment in Model 2 and for the effects of class and age in Model 3.

Table 4 Effects of incentive and sociodemographic variables on completeness of open-answer responses - Effect coefficients of a multinomial logistic regression (betas, standard errors in parenthesis; reference: 'all open-answer questions within the questionnaire completed')

	Model 1		Model 2		Model 3	
	partially answered	not answered	partially answered	not answered	partially answered	not answered
Intercept	0.173 (0.129)	-0.649* (0.162)	0.494* (0.189)	-0.345 (0.238)	0.233 (0.143)	-0.802*** (0.192)
Male (Ref: Female)	0.48* (0.201)	0.046 (0.265)	0.466* (0.208)	0.05 (0.283)	0.468* (0.215)	0.048 (0.296)
Pen (Ref.: No pen)	0.3 (0.174)	-0.087 (0.228)	0.300 (0.18)	-0.092 (0.244)	0.344 (0.188)	-0.005 (0.261)
Interaction: Male*Pen	-0.114 (0.284)	0.921* (0.360)	-0.117 (0.294)1	1.009** (0.384)	-0.134 (0.302)	0.85* (0.402)
Apprentice / Student (Reference: employment)			-0.494 (0.253)	-1.909*** (0.489)		
Not employed (Reference: employment)			0.361* (0.161)	0.499* (0.204)		
Lower secondary school (Reference: high school diploma)			0.318 (0.302)	0.874** (0.330)		
High school diploma + university entrance exam (Reference: High school diploma)			-0.631** (0.242)	-0.410 (0.313)		
University degree (Reference: high school diploma))			-0.581*** (0.171)	-0.997*** (0.221)		
Age (z-score)					0.522*** (0.079)	0.879*** (0.106)
Class (z-score)					-0.302*** (0.077)	-0.593*** (0.1)
n	1137		1115		1038	
Cox & Snell Pseudo- R^2	0.024		0.115		0.138	

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

The category "not employed" includes unemployed, irregularly employed, retired, and individuals on leave.

The class variable was handled metrically based on Winkler (1998) and calculated based on statements about education and employment, with possible values between 4 and 13. Since no information was collected on income, this concept of class is incomplete and should only be seen as an approximation of socio-economic status.

Table 5 Costs for printing and mailing based on incentive condition

	Total	Incentive	Control
Pre notification: print	611.09 €	305.55 €	305.55 €
Pre notification: mailing	1,194.47 €	597.24 €	597.24 €
Questionnaire: print	2,492.17 €	1,246.09 €	1,246.09 €
Questionnaire: mailing	2,182.57 €	1,091.29 €	1,091.29 €
Mailing of freepost and pre-addressed envelopes	1,682.00 €	909.96 €	772.04 €
Reminder-letter: print	629.92 €	314.96 €	314.96 €
Reminder-letter: mailing	1,155.79 €	577.90 €	577.90 €
Thank you letter and preliminary results: print	810.51 €	405.26 €	405.26 €
Thank you letter and preliminary results: mailing	1,158.49 €	579.25 €	579.25 €
Raffle	1,500 €	750.00 €	750.00 €
Incentive: material	417.03 €	417.03 €	
Incentive: mailing	0.00 €	0.00 €	
Total	13,834.04 €	7,194.50 €	6,639.54 €
Response rate (RR2)		30.95 %	26.25 %
Cost per complete questionnaire	12.09 €	11.62 €	12.65 €

3.5 Results: Effects of the Ballpoint Pen on Cost-effectiveness

Table 5 shows that the ballpoint pen reduced the cost per completed questionnaire from 12.65 Euro to 11.62 Euro. This decrease in cost per completed questionnaire is caused by the higher response rate in the incentive condition (30.95 vs. 26.25%). From an economic point of view, the additional costs of 417 Euro for the ballpoint pen were redeemed. Note that the inclusion of the ballpoint pen did not cause additional costs for mailing in this study.

4 Summary of Results

In this study, the use of a ballpoint pen increased the response rate by 4.7 percentage points. This effect of the incentive proved to be statistically significant (C=0.052; p<0.05). In comparison with other experiments in the German language area that had worked with monetary or money-like incentives, the ballpoint pen thus had a relatively weak effect on the subjects' willingness to respond. However, the relationship between the effect of an incentive on the response rate and its cost also needs to be considered: Harkness et al. (1998) increased unit non-response by 5% (from 29.3% to 34.3%) by sending four stamps, each with a value of one German

Mark, while Stadtmüller (2009) increased the response rate by 13% (from 30% to 42.7%) by sending a one Euro coin, and Becker et al. (2007) reported an increase of 24% (from 39% to 63%) by sending a ten Franc bill.[5] Therefore, in view of its comparatively low financial value, the ballpoint pen had a surprisingly strong impact on the response rate.

Further analysis revealed a gender-specific effect, i.e. women were more likely than men to react to the in-kind incentive. This is consistent with the findings of Mehlkop & Becker (2007) as well as Harkness et al., (1998, p. 213), where slightly (but not significantly) stronger effects of a monetary incentive on women were demonstrated. In contrast, Arzheimer (1998, p. 24), Nederhof (1983, p. 106) and Stadtmüller (2009, p. 180) explicitly deny a gender-specific incentive effect. Baumgartner and Rathbun (1997) and Groves et al. (2000, p. 304) attribute these contradictory results to the influence of a third variable, "Interest in the Survey Topic". Applied to our survey, it is possible that the female sample contained more "undecided" subjects that, through an incentive, could be motivated to participate, whereas the male sample contained more "decided" subjects who had no interest in the survey and could not be swayed by the non-monetary incentive. This hypothesis cannot be tested as there is no further information available about the interest of the non-responders in the survey topic. Apart from the gender-specific effect, the ballpoint pen had no meaningful influence on sample composition: looking at the variables religious affiliation, education, vocational training and employment status, the incentive did not substantially alter the composition of the sample that was collected.

Consistent with the findings of Houston and Jefferson (1975) and Nederhof (1983), the ballpoint pen reduced the time required before the questionnaires were returned. The hypothesis that a ballpoint pen has an adverse effect on response speed (Hansen, 1980, p. 81) was not supported by this study. In addition, economic considerations favor the inclusion of a ballpoint pen: the "break-even-equation" proposed by Jobber et al. (2004, p. 23) calculates the cost efficiency of printing and sending the survey, the reminder, and the incentive. In this study, the cost per returned questionnaire without an incentive was 12.65 Euro, while costs were reduced to 11.62 Euro per returned survey with the incentive and the associated increased return rate.

The ballpoint pen had no observable effect on response behavior to closed questions. However, the incentive caused subjects to refuse to answer all of the sensitive open-answer questions more frequently in the male population, while this effect was not seen in the female population. These results seem to support the interpretation that uninterested men refused to answer the personal questions because they

[5] These results are, however, only partially comparable to this study due to the fact that the design, sample, number of reminders, special theme and structure of the questionnaire differed in comparison to the other experiments.

felt pressured into completing the survey after receiving the incentive. The fact that this effect of the ballpoint pen on non-response to an item was only shown in relation to the open-answer questions could be based on the relative ease with which closed questions can be answered. Our results are therefore in line with the hypothesis put forward by Medway (2012), stating that sensitive items may be more susceptible to incentive effects than non-sensitive items.

5 Conclusions

This survey experiment demonstrated a significant effect of a ballpoint pen on unit non-response in a postal survey on a German sample. Although this effect on the response rate is small in comparison with the effects of monetary incentives identified in other studies, the use of in-kind incentives can be advantageous in certain survey designs: the results show that sending a ballpoint pen along with a postal survey can lead to faster response times, an effect that is short-lived and, in our study, had ceased by the time the reminder was sent. This result suggests that non-monetary incentives with a low value can be a sensible substitute for follow-ups. In study designs where financial limitations or privacy protection do not permit a reminder to be sent, sending a small gift in the form of a non-monetary incentive could be a valid alternative. The provocation of reactive response behavior is a clear disadvantage to the use of incentives. In our study, we found evidence that the quality of answers to especially intimate and complicated questions suffered through the inclusion of the incentive and that this negative effect was gender-specific. These results cannot be generalized as the sample of this study is made up of the inhabitants of Halle / Saale and the sponsor of the survey was an academic institution (the University of Halle-Wittenberg). The effect of in-kind incentives on other populations or in surveys with a different sponsorship is therefore hard to predict and should be subject of further research. Nonetheless, our results suggest the use of ballpoint pens as a cost-effective means of increasing the response rate to postal surveys. However, this gain should be weighed against the risk of lowering the validity of answers incentivized by gifts – at least when it comes to time consuming, sensitive and morally relevant questions.

References

Armstrong, J. S. (1975). Monetary Incentives in Mail Surveys. *The Public Opinion Quarterly*, 39(1), 111-116.

Arzheimer, K. & Klein, M. (1998). Die Wirkung materieller Incentives auf den Rücklauf einer schriftlichen Panelbefragung. *ZA-Information / Zentralarchiv für Empirische Sozialforschung*, 43, 6-31.

Auspurg, K. & Schneck, A. (2014). What difference makes a difference?: A meta-regression approach on the effectiveness conditions of incentives in self-administered surveys. MAER-Net. MAER-Net 2014 Athens Colloquium, Athen. Retrieved from http://meta-analysis2014.econ.uoa.gr/fileadmin/metaanalysis2014.econ.uoa.gr/uploads/Schneck_Andreas.pdf [June 2015]

Barón, J. D., Breunig, R. V., Cobb-Clark, D., Gørgens, T., & Sartbayeva, A. (2008). *Does the Effect of Incentive Payments on Survey Response Rates Differ by Income Support History?* [IZA Discussion Paper No. 3473]. Bonn: Forschungsinstitut zur Zukunft der Arbeit.

Baumgartner, R. & Rathbun, P. (1997). *Prepaid Monetary Incentives and Mail Survey Response Rates*. Paper presented at the annual conference of the American Association of Public Opinion Research, Norfolk: VA.

Becker, R., Imhof, R., & Mehlkop, G. (2007). Die Wirkung monetärer Anreize auf den Rücklauf bei einer postalischen Befragung und die Antworten auf Fragen zur Delinquenz. *Methods, Data, Analyses,* 1 (2), 131-159.

Berger, F. (2006). Zur Wirkung unterschiedlicher materieller Incentives in postalischen Befragungen: ein Literaturbericht. *ZUMA Nachrichten Nr. 58*, 30, 81-100.

Bland, J. M. & Altman, D. G. (2004). The logrank test. *BMJ*, 328 (1May), 1073.

Blau, P. M. (1964). *Exchange and power in social life*. New York: John Wiley & Sons.

Blohm, M. & Koch, A. (2013). Respondent Incentives in a National Face-to-Face Survey: Effects on Outcome Rates, Sample Composition and Fieldwork Efforts. *Methods, Data, Analyses*, 7 (1), 89-122.

Börsch-Supan, A., Krieger, U., & Schröder, M. (2013). Respondent incentives, interviewer training and survey participation. SHARE Working Paper. 12-2013. München.

Castiglioni, L, Pforr, K., & Krieger, U. (2008). The effect of incentives on response rates and panel attrition. Results from a controlled experiment. *Survey Research* Methods 2(3), 151-158.

Church, A. H. (1993). Estimating the Effect of Incentives on Mail Survey Response Rates: A Meta-Analysis. *The Public Opinion Quarterly*, 57 (1), 62-79.

Davern, M., Rockwood, T. H., Sherrod, R., & Campbell, S. (2003). Prepaid monetary incentives and data quality in face-to-face interviews. Data from the 1996 survey of income and program participation incentive experiment. *The Public Opinion Quarterly*, 67 (1), 139-147.

Diekmann, A. & Jann, B. (2001). Anreizformen und Ausschöpfungsquoten bei postalischen Befragungen. Eine Prüfung der Reziprozitätshypothese. *ZUMA-Nachrichten,* 48 (25), 18-27.

Dillman, D. A. (2000). *Mail and internet surveys. The tailored design method.* New York: John Wiley & Sons.

Dillman, D. A. (1978). *Mail and telephone surveys: The total design method.* New York: John Wiley & Sons.

Falk, A. (2007). Gift Exchange in the Field. *Econometrica*, 75 (5), 1501-1511.

Fick, P. & Diehl, C. (2013). Incentivestrategien bei Minderheitenangehörigen: Ergebnisse eines Methodenexperimentes. *Methods, Data, Analyses,* 7 (1), 59-88.

Fox, R. J., Crask, M. R., & Kim, J. (1988). Mail Survey Response Rate: A Meta-Analysis of Selected Techniques for Inducing Response. *The Public Opinion Quarterly*, 52 (4), 467-491.

Furse, D. H. & Stewart D. W. (1982). Monetary Incentives versus Promised Contribution to Charity: New Evidence on Mail Survey Response. *Journal of Marketing Research*, 19 (3), 375-380.

Göritz, A. S. & Luthe, S. C. (2013). Lotteries and study results in market research online panels. International Journal of Market Research, 55(5), 611-626.

Göritz, A. S., Wolff, H.-G. & Goldstein, D. G. (2008). Individual payments as a longer-term incentive in online panels. *Behavior Research Methods*, 40(4), 1144-1149.

Göritz, A. S. (2006). Cash lotteries as incentives in online panels. *Social Science Computer Review*, 24(4), 445-459

Goodstadt, M. S., Chung, L., Kronitz, R., & Cook, G. (1977). Mail Survey Response Rates: Their Manipulation and Impact. *Journal of Marketing Research*, 14 (3), 391-395.

Gouldner, A. W. (1960). The Norm of Reciprocity: A Preliminary Statement. *American Sociological Review*, 25 (2), 161-178.

Groves, R. M. & McGonagle, K. A. (2001). A Theory-Guided Interviewer TrainingProtocol Regarding Survey Participation. *Journal of Official Statistics*, 17 (2), 249-266.

Groves, R. M., Singer, E., & Corning, A. (2000). Leverage-Saliency Theory of Survey Participation. Description and Illustration. *The Public Opinion Quarterly*, 64 (3), 299-308.

Hansen, R. A. (1980). A Self-Perception Interpretation of the Effect of Monetary and Nonmonetary Incentives on Mail Survey Respondent Behavior. *Journal of Marketing Research*, 17 (1), 77-83.

Harkness, J., Mohler, P., Schneid, M., & Christoph, B. (1998). Incentives in Two German Mail Surveys 1996/97 & 1997, In A. Koch & R. Porst (Eds.), *Nonresponse in Survey Research* [ZUMA-Nachrichten Spezial 4.] (pp. 201-218). Mannheim: ZUMA.

Harm, K. & Jaeck, T. (2013). Bürgerumfrage Halle 2012. *Der Hallesche Graureiher,* 2013-2. Institut für Soziologie der Martin-Luther-Universität Halle-Wittenberg. Halle/Saale.

Hippler, H.-J. (1988). Methodische Aspekte schriftlicher Befragungen: Probleme und Forschungsperspektiven. *Planung und Analyse*, 6/88, 244-248.

Hippler, H.-J. (1985). Schriftliche Befragung bei allgemeinen Bevölkerungsstichproben: Untersuchungen zur Dillmanschen ‚Total Design Method'. *ZUMA-Nachrichten*, 16 (9), 39-56.

Houston, M. J. & Ford, N. M. (1976). Broadening the Scope of Methodological Research on Mail Surveys. *Journal of Marketing Research*, 13 (4), 397-403.

Houston, M. J. & Jefferson, R. W. (1975). The Negative Effects of Personalization on Response Patterns in Mail Surveys. *Journal of Marketing Research*, 12 (1), 114-117.

James, J. M. & Bolstein, R. (1990). The Effect of Monetary Incentives and Follow-Up Mailings on the Response Rate and Response Quality in Mail Surveys. *The Public Opinion Quarterly*, 54 (3), 346-361.

Jobber, D., Saunders, J., & Vince-Wayne, M. (2004). Prepaid Monetary Incentive Effects on Mail Survey Response. *Journal of Business Research*, 57 (1), 21-25.

Krenzke, T., Mohadjer, L., Ritter, G., & Gadzuk, A. (2005). Incentive effects on self- report of drug use and other measures of response quality in the Alcohol and Drug Services study. *Journal of Economic and Social Measurement*, 30 (2,3), 191-217.

Levin, P. F. & Isen, A. M. (1975). Further Studies on the Effect of Feeling Good on Helping. *Sociometry*, 38 (1), 141-147.

Linsky, A. S. (1975). Stimulating Responses to Mailed Questionnaires: A Review. *The Public Opinion Quarterly*, 39 (1), 82-101.

Little, E. L. & Engelbrecht. E. G. (1990). The use of incentives to increase mail survey response rates in a business environment: A field experiment. *Journal of Direct Marketing*, 4 (4), 46-49.

Lynn, P. (2001). The impact of incentives on response rates to personal interview surveys: Role and perceptions of interviewers. *International Journal of Public Opinion Research*, 13 (3), 326-336.

Martin, S., Helmschrott, S., & Rammstedt, B. (2014): The Use of Respondent Incentives in PIAAC: The Field Test Experiment in Germany. *Methods, Data, Analyses*, 8 (2), 223-242.

Martin, E., Abreu, D. & Winters, F. (2001). Money and motive: Effects of incentives on panel attrition in the survey of income and program participation. *Journal of Official Statistics*, 17(2), 267-284.

Medway, R. (2012). Beyond response rates. The effect of prepaid incentives on measurement error (Dissertation). University of Maryland, College Park, MD. Retrieved from http://drum.lib.umd.edu/bitstream/1903/13646/1/Medway_umd_0117E_13833.pdf; June 2015

Mehlkop, G. & Becker, R. (2007). Zur Wirkung monetärer Anreize auf die Rücklaufquote in postalischen Befragungen zu kriminellen Handlungen: theoretische Überlegungen und empirische Befunde eines Methodenexperiments. *Methods, Data, Analyses,* 1 (1), 5-24.

Mizes, S. J., Fleece, L. E., & Roos, C. (1984). Incentives for Increasing Return Rates: Magnitude Levels, Response Bias and Format. *The Public Opinion Quarterly*, 48 (4), 794-800.

Nederhof, A. J. (1983). The effects of material incentives in mail surveys: Two studies. *The Public Opinion Quarterly*, 47 (1), 103-111.

Porst, R. (1999). Thematik oder Incentives? Zur Erhöhung der Rücklaufquoten bei postalischen Befragungen. *ZUMA Nachrichten* Nr. 45, 23, 72-87.

Reuband, K.-H. (1999). Telefonkarten als Incentives für nicht-kooperative Zielpersonen in postalischen Befragungen. Auswirkungen auf die Teilnahmebereitschaft und die Zusammensetzung der Befragten. *Planung & Analyse*, 99 / 3, 63-66.

Schröder, M., Saßenroth, D., Körtner, J., Kroh, M., & Schupp, J. (2013). Experimental evidence of the effect of monetary incentives on cross-sectional and longitudinal response. Experiences from the Socio-Economic Panel (SOEP). *SOEPpapers on Multidisciplinary Panel Data Research,* 603. Deutsches Institut für Wirtschaftsforschung (DIW). Berlin.

Shettle, C. & Mooney, G. (1999). Monetary Incentives in U.S. Government Surveys. *Journal of Official Statistics*, 15 (2), 231-250.

Simmons, E. & Wilmot, A. (2004). Incentive Payments on Social Surveys: A Literature Review. *Survey Methodology Bulletin*, 3, 1-11.

Singer, E., & Ye, C. (2013). The use and effects of incentives in surveys. *The ANNALS of the American Academy of Political and Social Science*, 645(1), 112-141.

Singer, E. (2011). Toward a benefit-cost theory of survey participation: evidence, further tests, and implications. *Journal of Official Statistics*, 27(2), 379-392

Singer, E., van Hoewyk, J., & Maher, M. P. (2000). Experiments with Incentives in Telephone Surveys. *Public Opinion Quarterly*, 64 (2), 171-188.

Singer, E., van Hoewyk, J., Gebler, N., Raghunathan, T., & McGonagle, K. (1999). The effect of incentives on response rates in interview-mediated surveys. *Journal of Official Statistics*, 15(2), 217-230.

Stadtmüller, S. & Porst, R. (2005). *Zum Einsatz von Incentives bei postalischen Befragungen* [ZUMA How-to-Reihe, Nr. 14]. Mannheim: Universität Mainz & Zentrum für Umfragen, Methoden und Analysen.

Stadtmüller, S. (2009). Rücklauf gut, alles gut? Zu erwünschten und unerwünschten Effekten monetärer Anreize bei postalischen Befragungen. *Methods, Data, Analyses*, 3 (2), 167-185.

Thibaut, J. W. & Kelley, H. (1959). The social psychology of groups. New York: John Wiley & Sons.

Trussell, N. & Lavrakas, P. J. (2004). The influence of incremental increases in token cash incentives on mail survey response. Is there an optimal amount? *The Public Opinion Quarterly*, 68(3), 349-367.

Tzamourani, P. & Lynn, P. (1999). The effect of monetary incentives on data quality – Results from the British Social Attitudes Survey 1998 experiment. Centre for Research into Elections and Social Trends Working Paper.

Van Veen, F., Göritz, A. S., & Sattler, S. (2015). Response effects of prenotification, prepaid cash, prepaid vouchers, and postpaid vouchers: An experimental comparison. *Social Science Computer Review,* 34(3). doi: 10.1177/0894439315585074

Warriner, K., Goyder, J., Gjertsen, H., Hohner, P., & McSpurren, K. (1996). Charities, no; lotteries, no; cash, yes. *The Public Opinion Quarterly*, 60 (4), 542-562.

Winkler, J. (1998). Die Messung des sozialen Status mit Hilfe eines Index in den Gesundheitssurveys der DHP, In W. Ahrens, B. M. Bellach & K. H. Jöckel (Eds), *Messung soziodemographischer Merkmale in der Epidemiologie* [Schriften des Robert Koch-Institut 1/98] (pp. 69-74) München: Robert-Koch Institut.

Wotruba, T. R. (1966). Monetary Inducements and Mail Questionnaire Response. *Journal of Marketing Research*, 3 (4), 398-400.

Yu, J. & Cooper, H. (1983). A Quantitative Review of Research Design Effects on Response Rates to Questionnaires. *Journal of Market Research*, 20 (1), 36-44.

methods, data, analyses | Vol. 10(1), 2016, pp. 47-72 DOI: 10.12758/mda.2016.003

Privacy Concerns in Responses to Sensitive Questions. A Survey Experiment on the Influence of Numeric Codes on Unit Nonresponse, Item Nonresponse, and Misreporting

Felix Bader[1, 2], *Johannes Bauer*[1, 3], *Martina Kroher*[4]& *Patrick Riordan*[1]

1 *Department of Sociology, Ludwig-Maximilians-Universität München*
2 *Mannheim Centre for European Social Research (MZES)*
3 *Institute for Employment Research (IAB), Nuremberg*
4 *Institute for Sociology, Leibniz Universität Hannover*

Abstract

Paper-and-pencil surveys are a widely used method for gaining data. Numeric codes printed on the questionnaire are often a prerequisite for the use of scan software, which, in turn, permits a fast and efficient entering of the data from such surveys. However, printed numbers used for optical mark recognition on a questionnaire can provoke concerns about anonymity that may lead to unit nonresponse, item nonresponse, and misreporting.

To test this, we conducted an experiment in a mail survey on group-focused enmity, printing a scanner code on half of the questionnaires. Our results show no significant deviation concerning unit nonresponse. We find a higher item nonresponse and misreporting bias towards socially desirable answers in sensitive questions if the questionnaire is marked with a code. The influence of biased responses on regression results is minor. If the numeric code is brought to the respondents' attention in the cover letter, regression coefficients might be affected. Therefore we conclude that researchers should trade off these small biases against the usefulness of the code. From a methodological perspective, we recommend not to make a statement concerning the numeric code in the cover letter.

Our results are of relevance for researchers conducting paper-and-pencil surveys as well as for those analyzing data sets from these surveys. While this article analyzes biases caused by scanner codes, the results are potentially transferable to printed identification numbers used in panel studies, in survey experiments, or to match paradata or context data.

Keywords: questionnaire design, scanner codes, sensitive questions, tailored design, unit nonresponse, item nonresponse, misreporting

1 Introduction

In this study, we analyze the effects of numeric codes printed on paper questionnaires and their influences on respondent answers. These codes are often used in scan software for a fast and efficient entering of data. While we are primarily concerned with paper-and-pencil mail and interviewer surveys, with mixed-mode surveys becoming more prevalent the following results are also relevant for other survey types. This study focuses on scanner codes, but there are other potential purposes of such codes. They might be useful to identify respondents in panel studies or recognize treatment groups in survey experiments. Another application is the adding of paradata or context data. Regional identification numbers printed on the questionnaire, for example, can help to evaluate regional nonresponse and append geodata and other context data from external sources.

In paper-and-pencil surveys the respondent either completes the questionnaire herself or with an interviewer and the answers are filled in manually on a paper questionnaire. Subsequently this data needs to be digitalized in some way in order to enable researchers to efficiently analyze it. The digitalization can be done by manual input or by utilizing scan software.

The obvious advantages of scanning questionnaires is the considerable amount of time saved compared to the arduous procedure of manual data input and the higher quality of the produced data set. Also, when different people collaborate in the manual data processing, structural errors like different coding of missings and filter questions can become a problem. Consequently, scanning is advantageous in many respects and part of multiple recent and past surveys.

However, a problem when capturing data optically may arise due to the use of numeric codes, barcodes or QR codes, which are printed on the questionnaire. Prevalent software like Readsoft, TeleForm, EvaSys, and at least 20 other popular tools utilize optical mark recognition, which uses a printed code to identify a stored master form when processing the data. Nine member institutes of the German ADM-Sampling-System for Face-to-Face Surveys regularly apply paper-and-pencil surveys. An informal survey revealed that roughly half of these institutes use and recommend printed scanner codes.

Direct correspondence to
Johannes Bauer, Department of Sociology, LMU Munich, Konradstr. 6, 80801 Munich
E-Mail: johannes.bauer@lmu.de

Acknowledgements: We are grateful to the participants of the course in research practice "Methods of quantitative empirical social research 1" in 2013 who helped preparing the survey documents, generated the addresses, and delivered the questionnaires. Many thanks to Christian Ganser who provided valuable advice and support during all phases of the project. We also thank two anonymous reviewers for their helpful comments.

While scanner codes do not permit an identification of respondents, survey participants might feel that their anonymity is jeopardized by such numbers and react to a perceived breach of their anonymity in one of three ways: First, they might not participate in the survey at all (unit nonresponse). Second, they might participate, but decline answering sensitive questions (item nonresponse). Third, they might take part and answer even sensitive questions, but their answers might be biased by social desirability (misreporting). Misreporting is a general problem in surveys, especially in surveys with sensitive topics (Preisendörfer & Wolter, 2014; Wolter, 2012).

In this paper we present a survey experiment conducted to examine whether such codes lead to any of the three mentioned reactions by respondents. We delivered paper questionnaires on political opinions and group-focused enmity (GFE) to a random sample of the resident population of Munich. Respondents randomly received one of three versions: a) questionnaire without numeric code (these were carefully input by hand), b) questionnaire with numeric code but without specific statement concerning this code in the cover letter, and c) questionnaire with numeric code and a specific statement in the cover letter truthfully explaining the purpose of the code and that it cannot be used to jeopardize anonymity. The latter procedure is proposed by Dillman (2007) if numeric codes are inevitable. We compare unit nonresponse, item nonresponse, and answers to sensitive questions of these three groups as well as differences in results of typical GFE-regressions to determine the influence of such codes.

In the following section we outline the theoretical arguments relevant to our research question and discuss previous related research (section 2). In section 3 we present our data and the experimental design, before reporting central findings in section 4. We discuss these findings and their applicability to other topics and conclude the paper with some implications for practical research in section 5.

2 Theoretical and Empirical State of Research

There is an extensive theoretical and empirical literature on the factors influencing the response of survey participants. In this section we briefly outline the basic theoretical argument underlying our hypotheses and discuss empirical work directly related to the effects of survey design, sensitive questions and respondents' concern for the anonymity of their answers. For brevity's sake, this outline is limited to self-administered mail surveys, since this is the mode in question for this study. Regarding answers to sensitive questions, self-administered surveys tend to lead to less biased answers than interview-based surveys (Bradburn, Sudman, & Wansink, 2004; Richman, Kiesler, Weisband, & Drasgow, 1999; Stocké, 2004).

2.1 Tailored Design

The literature on the "tailored design method" (Dillman, 2007) has shown that in order to maximize response in mail surveys it is necessary to pay attention to every detail of the questionnaire, the cover letter, and all other elements submitted to the respondents (Babbie, 2013; de Leeuw & Hox, 2008; de Leeuw, Hox, & Huisman, 2003; Dillman, 1991, 2007, 2008). Although the efficiency of these elements has not been empirically settled (e.g. de Rada, 2005; Edwards et al., 2002) it is a practical default to assume that potential responders will react to various aspects of the survey materials.[1]

In the widely used model of rational respondents, survey participation and truthful answers hinge on the respondents' sense of benefits outweighing costs (Dillman, 1991, 2007; Tourangeau, Rips, & Rasinski, 2000). Warwick and Lininger have outlined factors on the one hand contributing to survey response and on the other hand preventing respondents from giving information as early as 1975 (see also Lessler & Kalsbeek, 1992). They describe how individuals are willing to share their experiences with interested listeners, as long as the questions are not sensitive and participation is not too costly in any other way.

Respondents can have major concerns about the anonymity of their data. This is why virtually all survey researchers make a statement on anonymity or confidentiality at some point, usually in the cover letter. Responders will feel even more anonymous, when there is no possible way to breech their anonymity: "If you make it a white envelope without any marks on it other than the address the questionnaire has to be sent to, and if the questionnaire does not contain any visible form of numbering, the respondent will feel freer to respond honestly to the questions" (Lensvelt-Mulders, 2008, p. 470). As described above, automated optical data capture using scanners makes some form of coding on questionnaires necessary. We aim at exploring whether these codes influence the responses.

2.2 Sensitive Questions, Privacy, and Nonresponse

Whether respondents give truthful answers to sensitive questions or not, is a classic issue in survey methodology (Barton, 1958; Benson, 1941; Hyman, 1944) and numerous more recent studies have shown that respondents tend to underreport socially undesired behavior and to overreport socially desired behavior (Barnett, 1998; Beyer & Krumpal, 2010; Kreuter, Presser, & Tourangeau, 2008; Lee, 1993; Tourangeau et al., 2000). Regarding the domain of misreporting in surveys, research on sensitive questions is a very important matter (for reviews see Krumpal, 2011;

[1] We acknowledge that survey response is even more complex than outlined here. A brief review on the psychology of survey response can be found in Schwarz (2008).

Lee, 1993; Lensvelt-Mulders, 2008; Tourangeau & Yan, 2007). Such misreporting to sensitive questions has serious consequences. The prevalence estimates of the sensitive topics are systematically biased and valid analyses on relationships between independent variables and the sensitive behavior cannot be conducted (Bernstein, Chadha, & Montjoy, 2001; Ganster, Hennessey, & Luthans, 1983). Researchers have developed several specific techniques to reduce misreporting on sensitive questions such as wording, framing or randomized response, but empirical research shows that the success of these measures is limited (see Preisendörfer & Wolter, 2014).

We suspect that respondents will be more concerned about their privacy when sensitive topics are involved. Tourangeau and Yan (2007, p. 859) define sensitive questions as "questions that trigger social desirability concerns [and] […] those that are seen as intrusive by the respondents or that raise concerns about the possible repercussions of disclosing the information". Following this definition, our questionnaire encompasses questions with various degrees of sensitivity ranging from low-sensitivity questions on socio-demographics to very sensitive items on group-focused enmity. These items directly point at hostility toward certain groups such as disabled persons, homosexuals, immigrants, Muslims, Jews, homeless, and long-term unemployed. Conforming to such hostile items is clearly socially undesirable, given Western norms. Respondents may experience them as intrusive and fear for their reputation should their attitudes be revealed. More generally, items on political and ethical subjects are considered sensitive (Lensvelt-Mulders, 2008; Stocké, 2007).

As pointed out above, respondents can react to sensitive questions in different ways. First, they can answer truthfully, even if they are aware of their attitude being socially undesirable. Second, they might be reluctant to participate in such a survey (unit nonresponse) or decline answering sensitive questions (item nonresponse). Finally, they might answer sensitive questions but react by adjusting their answers according to what they suspect is socially desirable (misreporting). Survey research has repeatedly shown that sensitive questions increase nonresponse and lead to biased answers.[2] If numeric codes on the questionnaire are perceived as a potential breach in anonymity, all three effects should be enhanced.

The problem of nonresponse can in part be lessened by "a very specific privacy statement" (Lensvelt-Mulders, 2008, p. 467) to attenuate the respondents' concerns for anonymity or confidentiality. On the other hand it is often stated that this privacy statement should not be blatant since "too much emphasize on privacy protection can harm the bond of trust between the respondent and the researcher, resulting in higher nonresponse rates" (Lensvelt-Mulders, 2008, pp. 467f.; see also de Leeuw et al., 2003). However, other research finds only weak effects of different

[2] A detailed discussion of reasons for and consequences of respondents' possible reactions to sensitive questions can be found in Tourangeau et al. (2000).

versions of such statements and thus question the extent to which responders actually read confidentiality statements (see, e.g. Tourangeau et al., 2000).

2.3 Previous Empirical Studies

A classic study done by Singer (1978) finds that assuring confidentiality significantly reduces item nonresponse. In their meta-analysis of experimental studies on the effect of confidentiality assurances Singer, von Thurn, and Miller (1995) find a weak but robust positive effect only when sensitive topics were being surveyed. In a large study by Dillman, Singer, Clark, and Treat (1996) there was no difference between different versions of the confidentiality assurance. Ong and Weiss (2000) have shown that assuring anonymity or confidentiality has a strong impact on revealing socially undesirable information. It has also been documented that the sensitivity of the surveyed topic has an impact on the willingness to participate (Couper, Singer, Conrad, & Groves, 2008; Edwards et al., 2002).

Even more closely related to our research, Yang and Yu (2011) examined the effect of personal identifiers on questionnaires. They find that numerical and barcode identifiers on the cover page of a questionnaire both advance nonresponse and reduce socially undesirable answers. They also speculate about sensitive topics being most prone to these adverse effects. On the other hand, there is a number of experiments in surveys on sensitive topics that find no significant effect of a numeric code on unit nonresponse (Campbell & Waters, 1990; Reuband, 1999, 2006, 2015) or on unit nonresponse and reporting behavior (King, 1970; Wildman, 1977). In all these studies, the codes were actual identifiers, i.e. they were unique for every respondent.

2.4 Hypotheses

In conclusion, the review of the literature allows us to formulate these hypotheses:

H1: Numeric codes on questionnaires lead to higher unit nonresponse.

H2: Numeric codes on questionnaires lead to higher item nonresponse in sensitive questions.

H3: Numeric codes on questionnaires lead to answers to sensitive questions biased towards social desirability.

H4: If at least one of the Hypotheses H1-H3 is supported, the results of regressions might be biased.

H5.1-4: An explicit privacy statement in the cover letter addressing the nature of the numeric code might either attenuate or raise unit nonresponse (1), item nonresponse (2), misreporting (3), and biases in regression results (4).

3 Methods

3.1 Survey Design

In February and March 2013, 3,725 paper-and-pencil questionnaires were distributed to randomly selected households in Munich. The sample was generated following a recommendation for drawing local household samples in Bauer (2014). First we randomly selected 712 street sections from a list of 77,218 street sections in Munich. Each had the same probability to be selected. 12,130 households in the selected street sections were manually counted and listed. Based on this list a sample was drawn.

An envelope with the cover letter, the questionnaire, and a stamped return envelope was deposited in the mailboxes of all selected households.[3] The envelope as well as the cover letter and the questionnaire contained a letterhead from the university and were distributed without respondent names, so all respondents were in fact anonymous. The cover letter explained the topic of the survey as “better understanding of political and social developments”, asked for the participation of the household member over 18 who had their birthday most recently and emphasized the confidentiality and anonymity of the responses. Furthermore, small bag of jelly babies was added as a little incentive.[4] After two weeks, all households received a reminder postcard. Our methodical proceeding is guided by the tailored design method (Dillman, 2007). In total we received 1,138 questionnaires, which results in a response rate of 30.6%.

As the topic of the study is the analysis of group-focused enmity (GFE) in Munich, the questionnaire covered demographics, housing and neighborhood conditions, and social trends as well as societal and political opinions with an emphasis on GFE. Like in other studies on GFE (Heitmeyer, 2002a, 2002b; Zick et al., 2008) many questions were sensitive since they focus on topics of socially unde-

[3] The questionnaire (in German) can be found in the online appendix http://www.ls4.soziologie.uni-muenchen.de/forschung/zusatzinfos/sens_quest/.

[4] Studies suggest that, while it is more effective to use money as an incentive, small presents also have a positive effect (Church, 1993; Edwards et al., 2002; Fick & Diehl, 2013). Given financial restrictions, we decided to give sweets.

sired prejudices and discrimination.[5] The sensitivity of the questions on GFE was particularly suitable for this experiment, since concerns about potentially jeopardized anonymity are expected to influence answering behavior if true answers are discomforting.

3.2 Survey Experiment

To examine the effects of numeric codes on response behavior, we divided the sample into three groups (Table 1). Half of the questionnaires had a code written on the bottom of each page, the other half did not.[6] Half of those questionnaires with code received a cover letter, which explained the numeric codes to respondents: “The digits on the bottom of the questionnaire only facilitate electronic processing and are identical on each questionnaire of this project. They do not permit personal identification” (translated from German).[7] All households within the same street section received one type of questionnaire.

3.3 Data Analysis

All three treatment groups were analyzed regarding unit nonresponse, item nonresponse, and answers to sensitive questions.[8] Since researchers are typically interested in regression results, the groups were also compared in regressions on GFE by interacting all independent variables with the group dummies.

We focused on questions about GFE in the topics of attitudes towards people with disabilities, long-term unemployed persons, homeless people, homosexuals, Muslim and other immigrants, and attitudes towards cultural heterogeneity, anti-Semitism, and National Socialism. All questions from these topics ask respondent to express an attitude towards certain groups. The items range from 1 to 5 and are standardized to ensure that comparisons across items are not dependent on the variance within each item. Very sensitive questions have a lower variance, since

[5] For the question wording see Appendix A. Details on data collection and results concerning GFE in Munich can be found in Steinbeißer, Bader, Ganser, & Schmitt (2013). To test the sensitivity of GFE-items, 80 students from the University of Hanover (not the students from Munich who had helped preparing the project) were asked to rate the sensitivity of GFE and some other questions. According to our expectations, GFE-items were rated to be much more sensitive.

[6] In previous studies (Campbell & Waters, 1990; King, 1970; Reuband, 1999; 2006; 2015; Wildman, 1977; Yang & Yu, 2011) the codes were located only on the cover page of the questionnaire.

[7] The different versions of the cover letter in German can be found in the online appendix http://www.ls4.soziologie.uni-muenchen.de/forschung/zusatzinfos/sens_quest/.

[8] The data are available for scientific purposes from the authors. The R-code can be found at http://www.ls4.soziologie.uni-muenchen.de/forschung/zusatzinfos/sens_quest/.

Table 1 Variants of the questionnaire in the survey experiment

Treatment Group	Number of Questionnaires	Proportion
Without Numeric Code	1,863	50%
With Numeric Code, Without Notice	931	25%
With Numeric Code and Notice	931	25%
Total	3,725	100%

the truthful answers of only a small fraction of respondents result in the choice of a socially undesirable category. Even if these persons react stronger to a presumed violation of anonymity by the numeric code, smaller effects would be observed, as only a small percentage of respondents is affected. Standardizing the GFE-items allows to analyze variables relative to their own variance.

To test if respondents' reaction to numeric codes depends on the sensitivity of the item, we use a set of less sensitive questions for comparison. The non-sensitive questions cover life-satisfaction and finance, trust in institutions, and good neighborhood. As these questions deal with an individual's personal situation without aiming at morally relevant sentiments towards groups of human beings, we are confident that the urge to give socially acceptable answers is much weaker in these topics than in GFE.

Following Angrist and Pischke (2009) we applied multivariate methods, although we had conducted an experiment, for two reasons: First, there are significant differences between the treatment groups in age, employment status and city district of residence. We consider this an indicator of the randomization procedure, clustered by street sections, not producing a perfectly randomized split. This is why we control for regional and socio-demographic characteristics. Second, the variance in GFE is very large compared to the effect of numeric codes. This masks the effects of the printed codes. Therefore it is necessary to shrink the unexplained variance in GFE by conditioning on suitable explaining variables.

Table 2 gives an overview over mean and standard deviation for all GFE-scales and the used demographic variables. All GFE-items, the non-sensitive items, and their means can be found in the Appendix.

Table 2 Mean for GFE-scales and used demographic variables[a]

Unstandardized GFE-Scales [0,1]	Mean (Standard Deviation)
Xenophobia	0.245 (0.178)
Islamophobia	0.491 (0.250)
Anti-Semitism	0.244 (0.235)
Attitudes Towards Unemployed	0.596 (0.085)
Attitudes Towards Homosexuals	0.225 (0.261)
Attitudes Towards National Socialism	0.142 (0.126)

	Demographic Variables	Mean (Standard Deviation)
	Female	0.536
	German	0.933
	Age	49.3 (17.3)
	Monthly Net Income per Capita (in 1000)	1.722 (1.147)
Religion	Catholic	0.405
	Lutheran-protestant	0.196
	Other	0.031
	None	0.367
Education	No/Junior High School	0.128
	Middle School	0.188
	Adv. Tech. College Qualif.	0.053
	University Qualification	0.130
	University Degree	0.500
Employment	Full-time	0.514
	Regular Part-time	0.119
	Marginal Part-time	0.083
	None	0.283
Ever Registered as Unemployed		0.337

[a] For metric variables standard deviation in parentheses.

4 Results

4.1 Unit Nonresponse

The response rates in the three treatment groups are very similar and the differences are clearly not significant (see Figure 1 and Table 3). The 3-sample test for equality of proportions (Wilson, 1927) gives a χ^2-value of 0.314 and a p-value of 0.855.

It is save to conclude that the codes with or without notice have a minimal or no effect on unit nonresponse.

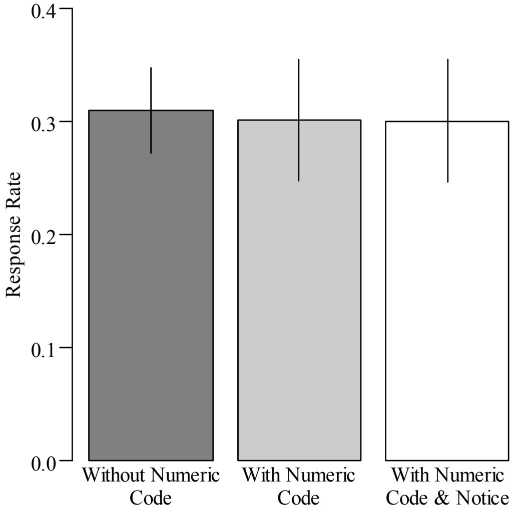


Figure 1 Response rate (percent and 95%-confidence interval) by treatment groups

Table 3 Response by treatment groups

Treatment Group	Number of responses	Response rate
Without Numeric Code	577	31.0%
With Numeric Code, Without Notice	281	30.2%
With Numeric Code and Notice	280	30.1%
Total	1,138	30.6%

4.2 Item Nonresponse

Overall, given the sensitivity of the survey, item nonresponse is pretty low. On average, a sensitive question on GFE was not answered by 13.6 of the 1,138 respondents (1.19%).

Each dot in Figure 2 represents the nonresponse to an item from the questionnaire. For example, the dots marked by arrows represent the item nonresponse to the question “To what extent do you agree with the following statement: The customs and habits of Islam feel creepy to me” (translated from German). In the group without numeric code, the nonresponse rate of this item is 1.04%, while it is higher for respondents who received a questionnaire with numeric code (1.78%) or with numeric code and notice in the cover letter (1.79%).

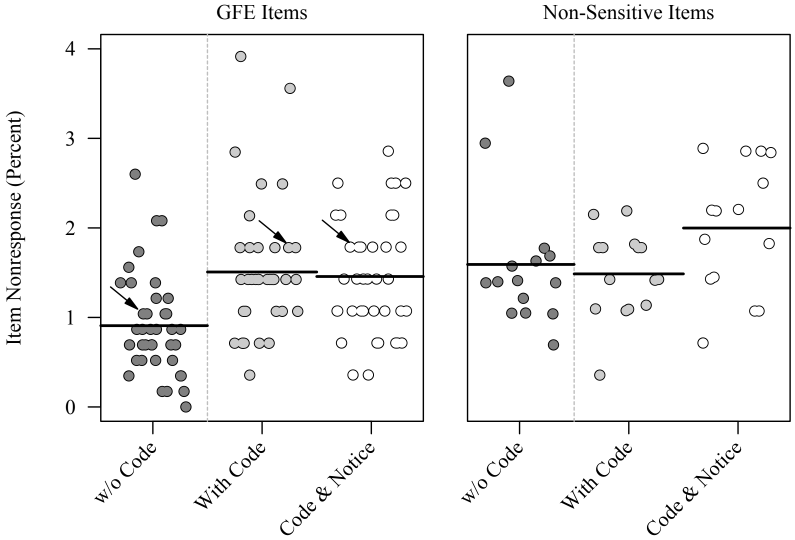


Figure 2 Item nonresponse rates (percent) for GFE-items and non-sensitive items. Black lines represent the average percentage of item nonresponse in each group. Arrows point to the example item in section 4.2 (skepticism about customs and habits of Islam).

Table 4 Average item nonresponse rates for GFE-items and non-sensitive items (p-values in parantheses)

Treatment Group	GFE			Non-Sensitive Items		
Without Numeric Code	0.91%			1.59%		
With Numeric Code, Without Notice	1.51%	(0.059)[a]	(0.032)[b]	1.49%	(0.570) [a]	(0.394) [b]
With Numeric Code and Notice	1.46%	(0.077)[a]		2.00%	(0.268) [a]	

[a] P-values for testing against the group without numeric code.

[b] P-value for testing both groups with numeric code together against the group without numeric code.

The group without numeric code has an average GFE-item nonresponse rate of 0.91%, compared to 1.51% in the group with numeric code and 1.46% in the group with notice in the cover letter (Table 4). The item nonresponse is low in all three groups and the differences are small in absolute numbers, but relatively we see a 66.1% and 60.5% rise in item nonresponse as an effect of the numeric code.

In Figure 2 and Table 4 we compare these results on GFE-items to group differences in nonresponse to non-sensitive items. Surprisingly, the non-sensitive items in general have a higher item nonresponse than the GFE-items. However, the items show no clear rise in nonresponse due to the numeric code. The group differences in nonresponse to GFE-items are close to significance ($p_{\text{without code vs. code without notice}} = 0.059$, $p_{\text{without code vs. code with notice}} = 0.077$), whereas in case of the non-sensitive items they are clearly nonsignificant ($p_{\text{without code vs. code without notice}} = 0.570$, $p_{\text{without code vs. code with notice}} = 0.268$). If we pool both groups with numeric code and test the difference in item nonresponse between this pooled group and the group without numeric code, the difference turns out to be significant in GFE-items ($p_{\text{without code vs. with code}} = 0.032$) while it is still insignificant in the non-sensitive items ($p_{\text{without code vs. with code}} = 0.394$).[9] A numeric code on the questionnaire does indeed increase item nonresponse but only in sensitive questions. The notice in the cover letter does not influence this effect.

4.3 Misreporting

To analyze the impact of numeric codes on given responses, we estimate group effects for all 38 standardized GFE-items applying a multivariate regression. In order to reduce the residual error all models include sex, age, age squared, religious affiliation, German citizenship, educational status, employment status, income per capita, and city district of residence. This regression on the standardized GFE-items results in parameters for respondents with numeric code but without notice and respondents who received a notice. The group without numeric code serves as reference.

Figure 3 shows the coefficients for the groups from the multivariate regression. Each dot represents the estimated difference in agreement with an item between the groups with numeric codes and the group without numeric code, which serves as reference. As items are standardized, coefficients represent the differences measured in standard deviations of the item. Using the example in section 4.2 regarding skepticism about customs and habits of Islam, respondents who received a questionnaire with a numeric code reported a 0.158 standard deviations lower agreement with this statement. Respondents, who were informed about the use of the numeric code, reported a 0.157 standard deviations lower agreement (see Arrows in Figure 3).

[9] To take into account that the answers within respondents are not independent, the test calculates the p-value by applying a Monte Carlo simulation. Values are drawn from a multivariate normal distribution using the variance covariance matrix from a multivariate regression. The multivariate regression contains only the experimental groups as explanatory variables and the nonresponse to each item as dependent variables. As the explanatory variables are all binary, common standard errors can be used. Sociodemographics do not contribute to the explanation of item nonresponse and therefore are not included in the model. As we expect that the numeric code increases nonresponse, we use one-sided tests.

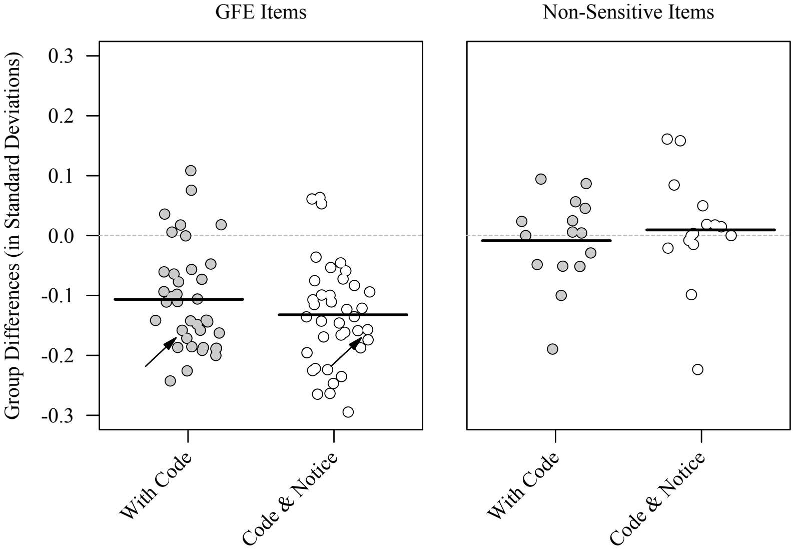


Figure 3 Regression coefficients for standardized GFE-items and standardized non-sensitive items (reference group: without code). Black lines represent the average group differences. Arrows point to the example item in section 4.2 (skepticism about customs and habits of Islam).

Table 5 Average regression coefficients for standardized GFE-items and non-sensitive items (p-values in parantheses)[a]

Treatment Group	GFE		Non-Sensitive Items	
With Numeric Code	-0.106	(0.022)	-0.009	(0.474)
With Numeric Code and Notice	-0.132	(0.011)	0.009	(0.612)

[a] Differences in standard deviations. P-values for testing against the group without numeric code.

All estimates of the multivariate GFE-regression for the groups with numeric code are between -0.294 and 0.108. The majority of coefficients are negative indicating a decrease in socially undesired responses resulting from the scanner code. Respondents with numeric code show lower GFE, i.e. their answers to sensitive questions tend more strongly towards desirability. The average difference over all items on GFE between the group without and the group with numeric code and without notice is -0.106 standard deviations and -0.132 standard deviations to the group with numeric code and notice (Table 5). In case of the non-sensitive items the differences are minimal (-0,009 and 0,009, Figure 3, right panel).

To compare the average effect of numeric codes, significance tests need to account for the dependencies between GFE-items within respondents. The test procedure is based on the variance-covariance matrix of the multivariate regression (similar to the one in the analysis of item nonresponse in 4.2, see footnote 9).[10] On average, the respondents with numeric code show significantly lower GFE ($p_{\text{without code vs. code without notice}} = 0.022$, $p_{\text{without code vs. code with notice}} = 0.011$). Respondents with numeric codes gave more socially desirable answers to sensitive questions. In case of the non-sensitive items the differences are insignificant ($p_{\text{without code vs. code without notice}} = 0.474$, $p_{\text{without code vs. code with notice}} = 0.612$).

We can conclude that the numeric code discourages respondents to give socially undesirable answers to sensitive questions but seems to have no impact concerning non-sensitive questions. While the effect of an explanation of the code's usage is not definite, it is safe to state that it has no positive effect.

4.4 Impact on GFE-regression results

So far, the results show higher item nonresponse and more socially desired answers in the treatment groups with numeric code. As most studies are interested in relationships between variables, we calculate a typical GFE-model for xenophobia and compare how the numeric code and the notice affect the regression coefficients.[11] All variables are interacted with the group dummies "code without notice" and "code with notice". If numeric codes do not influence respondents' behavior, the interactions should be insignificant. If numeric codes influence regression results, estimated interaction parameters should show significant effects.

The interaction coefficients and confidence intervals of the model are shown in Figure 4. With a critical value of 0.05, 26 explaining variables, and two interaction groups one would expect 2.6 parameters to be significant by chance. Three interactions terms with the numeric code group without notice became significant (gender, catholic religion and other religion). Interactions with notice did not exceed the critical value. All in all there are not more significant parameters than expected, which indicates that code and notice do not affect the regression results systematically.

[10] To test the standardized non-sensitive items we use the absolute coefficients, as the pooling of the items would otherwise influence the result. However, pooling is not possible, as there is no general expectation what kind of answer is more socially undesirable. For the standardized GFE-items the normal coefficients are used. P-values are calculated by drawing values from a multivariate normal distribution with the covariance matrix taken from the multivariate regressions for standardized GFE-items and standardized non-sensitive items. The values (GFE) or absolute values (non-sensitive items) from the drawn sample serve as error distribution. The multivariate normal distribution accounts for the dependencies between items.

[11] The variables used in this regression are guided by models in the report on the survey Steinbeißer et al. (2013).

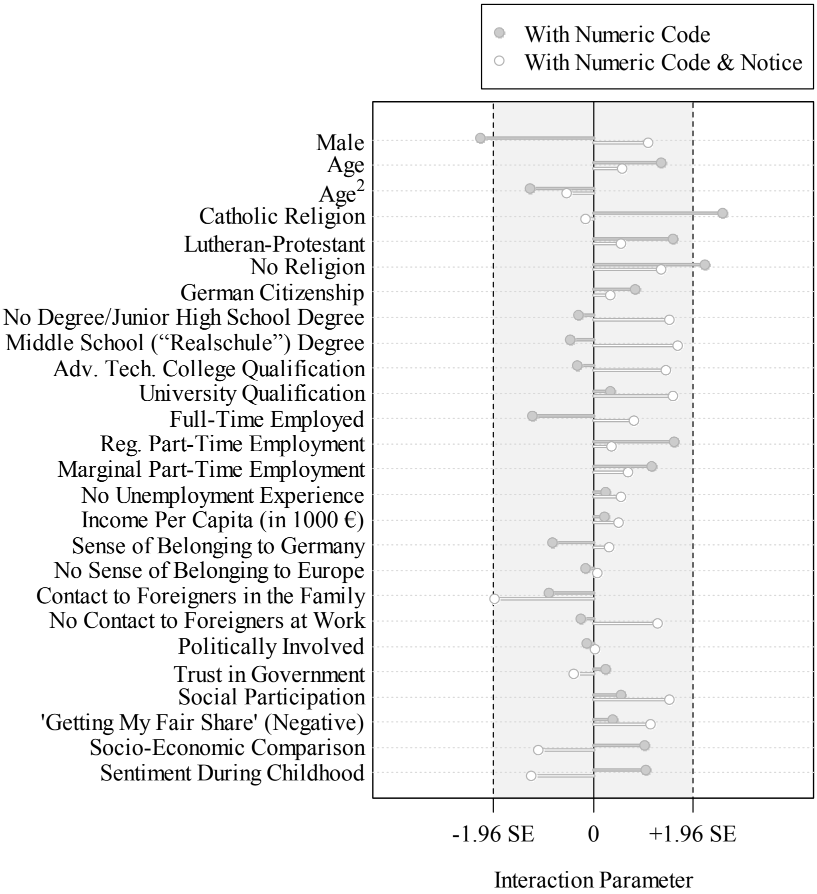


Figure 4 Interaction parameters of the treatment group dummy variables (ref: no numeric code) with several typical independent variables and 95%-confidence interval from a regression for xenophobia (n = 706).

Several other regression models support this result. We estimate additional models with the same independent variables for islamophobia, anti-Semitism and attitudes towards unemployed, homosexuality and National Socialism. When looking at the distribution of the interaction parameters, there are never more than four significant interaction parameters per model. This is consistent with the distribution of t-values for all interactions. If there is no relationship and the interaction parameters descend from a random distribution, t-values should converge to a standard normal distribution. The interaction terms are close to a standard normal distribution. The Kolmogorov-Smirnov test, Shapiro-Wilk test and Anderson-Darling test show no significant deviation for the group without notice ($p_{KS} = 0.676$, $p_{SW} = 0.614$,

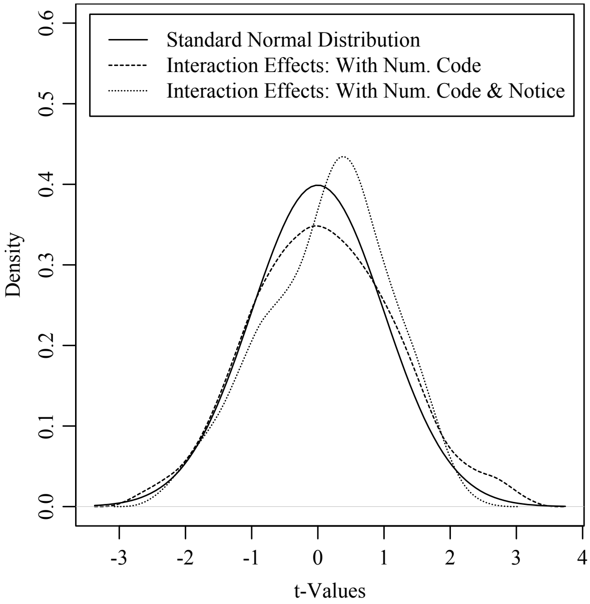


Figure 5 Distribution of 156 interaction parameters for treatment group dummy variables (ref: no numeric code) with independent variables from six regression models for GFE-variables (kernel density estimation).

$p_{AD} = 0.556$). In the group with notice we find a significant deviation with the Kolmogorov-Smirnov test ($p_{KS} = 0.007$) at the 1% significance level and with the Shapiro-Wilk test and Anderson-Darling test, at the 10% level ($p_{SW} = 0.092$, $p_{AD} = 0.052$). This can also be seen in Figure 5. While the peak of the t-value distribution for the group without notice is very close to zero, the peak for the group with notice is not. The test supports the visual impression, that the distribution of the group with notice is not normal. Given these results, an effect on the regression result, caused by the combination using a numeric code and pointing respondents to the code, can be suspected.

5 Conclusion and implications

We hypothesized that numeric codes on a questionnaire might influence respondents' behavior. Such a code could induce privacy concerns in the respondents. Respondents can possibly react by refusing to answer the complete questionnaire (H1), by skipping sensitive questions (H2) or by giving biased answers to sensitive questions (H3). If the code did have such an impact on respondents' behavior, this might also bias typical regression results (H4). These effects on unit nonresponse, item nonresponse, misreporting, and consequently on regression results can be attenuated or raised by a statement on the code's usage in the cover letter (H5.1, H5.2, H5.3, H5.4). To test these hypotheses a survey experiment was conducted: We printed a numeric code on half of the questionnaires and half of the respondents with a numeric code were informed about it in the cover letter.

Although the response rate was slightly lower in the groups with numeric code, the differences were far from significant. In line with other studies (Campbell & Waters, 1990; King, 1970; Reuband, 1999, 2006, 2015; Wildman, 1977), H1 and H5.1 are not supported.

Respondents with a numeric code had a significantly higher item nonresponse rate in sensitive questions on group-focused enmity, but no higher item nonresponse in non-sensitive items. Explaining the numeric values on the questionnaire in the cover letter did not have an influence on these relationships. H2 is supported, but there is no support for H5.2.

A significant misreporting bias towards socially desired answers was found in both treatment groups. An explanation of the numeric code appears to have no positive effect, however, results give no clear indication whether the notice introduces bias or not. H3 can be confirmed, but there is no support for H5.3.

The influence of these differences on regression results was insignificant. H4 cannot be confirmed. However, if the numeric code is addressed in the cover letter, there is some indication that regression coefficients might be affected, as the bias in single variables seems to accumulate. While the result is ambiguous, there is some support for H5.4.

In contrast to other studies (King, 1970; Wildman, 1977), we find at least some systematic differences in reporting behavior regarding item nonresponse and misreporting. The zero results of older studies might be due to limitations in the sample size and in the statistical techniques. Our survey uses a bigger sample, we apply tests based on multiple variables and we are able to reduce the unexplained variance by adjusting for suitable sociodemographic variables. The results are in accordance with Yang and Yu (2011): respondents who had a questionnaire with a scanner code gave more socially desirable answers. Given that non-sensitive items were not affected, it seems reasonable to assume that respondents reacted to a perceived breach in anonymity. We suspect that there are two ways in which a smaller item mean emerges. Several respondents reacted to the numeric code and adjusted their response a little in direction of a social desirable answer. In addition, a small group, who would honestly give extreme answers, completely changed their response to the opposite socially desirable extreme. We think that both influences are possible; however, our data does not allow distinguishing between these effects.

Given that optical mark recognition software very often uses numeric codes, we also want to give recommendations for survey methodologists. First, we recommend to avoid an explicit statement on that code in the cover letter (if legally and ethically justifiable) as it does not ease biases. Given our results, a numeric code can be problematic if there is a specific interest in descriptive results on a sensitive topic (like unemployment experiences). It is also possible that some groups of respondents react strongly to a numeric code on the questionnaire. On the other hand, the effects are small compared to other sources of bias like selective nonresponse to surveys (Schnell, 1997) and mismatch of answers and behavior (see e.g. Diekmann & Preisendörfer, 1992).

Second, researchers need to trade off the potential biases brought about by a numeric code and the importance of the code on the questionnaire for the research project. In case of automated capture of questionnaires instead of manual data entry, numeric codes will save money that can be used to reduce selective unit nonresponse.

While this study focuses on scanner codes, there are other potential purposes of such codes. Numbers on questionnaires can be utilized as panel identifiers, treatment group identifiers in survey experiments, and for region mapping. As far as the conditions for such codes are similar, our results should largely be transferable. If numeric codes are applied thoughtfully, their benefit for the usage of statistical procedures to reduce errors could well outweigh the code's negative effect.

References

Angrist, J. D., & Pischke, J.-S. (2009). *Mostly harmless econometrics: An empiricist's companion*. Princeton, NJ: Princeton Univ. Press.

Babbie, E. (2013). *The Practice of Social Research: Thirteenth Edition*. Belmont: Wadsworth.

Barnett, J. (1998). Sensitive Questions and Response Effects: An Evaluation. *Journal of Managerial Psychology, 13*(1/2), 63-76.

Barton, A. H. (1958). Asking the Embarrassing Question. *Public Opinion Quarterly, 22*(1), 67-68.

Bauer, J. J. (2014). *Verzerrungen in Random-Route Stichproben und Lösungsansätze: Vortrag ETH Zürich*. Retrieved from http://www.ls4.soziologie.uni-muenchen.de/forschung/zusatzinfos/sens_quest/index.html.

Benson, L. E. (1941). Studies in Secret-Ballot Technique. *Public Opinion Quarterly, 5,* 79-82.

Bernstein, R., Chadha, A., & Montjoy, R. (2001). Overreporting Voting. Why it Happens and Why it Matters. *Public Opinion Quarterly, 65*(1), 22-44.

Beyer, H., & Krumpal, I. (2010). „Aber es gibt keine Antisemiten mehr“: Eine experimentelle Studie zur Kommunikationslatenz antisemitischer Einstellungen. *Kölner Zeitschrift für Soziologie und Sozialpsychologie, 62*(4), 681-705.

Bradburn, N. M., Sudman, S., & Wansink, B. (2004). *Asking Questions.* San Francisco: Wiley.

Campbell, M. J., & Waters, W. E. (1990). Does Anonymity Increase Response Rate in Postal Questionnaire Surveys about Sensitive Subjects? A Randomised Trial. *Journal of Epidemiology and Community Health, 44*(1), 75-76.

Church, A. H. (1993). Estimating the Effect of Incentives on Mail Survey Response Rates: A Meta-Analysis. *Public Opinion Quarterly, 57*(1), 62-79.

Couper, M. P., Singer, E., Conrad, F. G., & Groves, R. M. (2008). Risk of Disclosure, Perceptions of Risk, and Concerns about Privacy and Confidentiality as Factors in Survey Participation. *Journal of Official Statistics, 24*(2), 255-275.

de Leeuw, E. D., & Hox, J. (2008). Self-administred Questionnaires: Mail Surveys and Other Applications. In E. D. de Leeuw, J. Hox, & D. A. Dillman (Eds.), *International Handbook of Survey Methodology* (pp. 239-263). New York, London: Lawrence Erlbaum Associates.

de Leeuw, E. D., Hox, J., & Huisman, M. (2003). Prevention and Treatment of Item Nonresponse. *Journal of Official Statistics, 19*(2), 153-176.

de Rada, V. D. (2005). Influence of Questionnaire Design on Response to Mail Surveys. *International Journal of Social Research Methodology, 8*(1), 61-78. doi:10.1080/1364557021000025991

Diekmann, A., & Preisendörfer, P. (1992). Persönliches Umweltverhalten: Diskrepanzen zwischen Anspruch und Wirklichkeit. *Kölner Zeitschrift für Soziologie und Sozialpsychologie, 44*(2), 226-251.

Dillman, D. A. (1991). The Design and Administration of Mail Surveys. *Annual Review of Sociology, 17,* 225-249.

Dillman, D. A. (2007). *Mail and Internet Surveys: The Taylored Design Method* (2nd Edition). New York: John Wiley & Sons.

Dillman, D. A. (2008). The Logic and Psychology of Constructing Questionnaires. In E. D. de Leeuw, J. Hox, & D. A. Dillman (Eds.), *International Handbook of Survey Methodology* (pp. 161-175). New York, London: Lawrence Erlbaum Associates.

Dillman, D. A., Singer, E., Clark, J. R., & Treat, J. B. (1996). Effects of Benefits Appeals, Mandatory Appeals, and Variations in Statements ofConfidentiality on Completion Rates for Census Questionnaires. *Public Opinion Quarterly, 60*(3), 376-389.

Edwards, P., Roberts, I., Clarke, M., Di Giusseppi, C., Pratap, S., Wentz, R., & Kwan, I. (2002). Increasing Response Rates to Postal Questionnaires: Systematic Review. *British Medical Journal, 324,* 1183-1192. doi:10.1136/bmj.324.7347.1183

Fick, P., & Diehl, C. (2013). Incentivierungsstrategien bei Minderheitsangehörigen. Ergebnisse eines Methodenexperiments. *methoden, daten, analysen, 7*(1), 59-88.

Ganster, D. C., Hennessey, H. W., & Luthans, F. (1983). Social Desirability Response Effects: Three Alternative Models. *Academy of Management Journal, 26*(2), 221-331.

Heitmeyer, W. (Ed.). (2002a). *Deutsche Zustände: Folge 1.* Frankfurt: Suhrkamp.

Heitmeyer, W. (2002b). Gruppenbezogene Menschenfeindlichkeit. Die theoretische Konzeption und erste empirische Ergebnisse [Group-focused enmity. Theoretical Conception and First Empirical Results]. In W. Heitmeyer (Ed.), *Deutsche Zustände. Folge 1* (pp. 15-36). Frankfurt: Suhrkamp.

Hyman, H. (1944). Do They Tell the Truth? *Public Opinion Quarterly, 8,* 557-559.

King, F. W. (1970). Anonymous versus Identifiable Questionnaires in Drug Usage Surveys. *American Psychologist, 25*(10), 982-985.

Kreuter, F., Presser, S., & Tourangeau, R. (2008). Social Desirability Bias in CATI, IVR, and Web Surveys. The Effects of Mode and Question Sensitivity. *Public Opinion Quarterly, 72*(5), 847-865.

Krumpal, I. (2011). Determinants of Social Desirability Bias in Sensitive Surveys: A Literature Review. *Quality & Quantity, 47,* 2025-2047.

Lee, R. M. (1993). *Doing Research on Sensitive Topics*. Thousand Oaks: SAGE.

Lensvelt-Mulders, G. (2008). Surveying Sensitive Topics. In E. D. de Leeuw, J. Hox, & D. A. Dillman (Eds.), *International Handbook of Survey Methodology* (pp. 461-478). New York, London: Lawrence Erlbaum Associates.

Lessler, J. T., & Kalsbeek, W. D. (1992). *Nonsampling Error in Surveys*. New York: John Wiley & Sons.

Ong, A. D., & Weiss, D. J. (2000). The Impact of Anonymity on Responses to Sensitive Questions. *Journal of Applied Social Psychology, 30*(8), 1691-1708.

Preisendörfer, P., & Wolter, F. (2014). Who Is Telling the Truth? A Validation Study on Determinants of Response Behavior in Surveys. *Public Opinion Quarterly, 78*(1), 126-146.

Reuband, K.-H. (1999). Anonyme und nicht-anonyme postalische Bevölkerungbefragungen. *Planung & Analyse, 26*(1), 56-58.

Reuband, K.-H. (2006). Postalische Befragungen alter Menschen: Kooperationsverhalten, Beantwortungsstrategien und Qualität der Antworten. *ZA-Information / Zentralarchiv für Empirische Sozialforschung, 59,* 100-127.

Reuband, K.-H. (2015). Ausschöpfung und Nonresponse Bias in postalischen Befragungen. In J. Schupp & C. Wolf (Eds.), *Schriftenreihe der ASI - Arbeitsgemeinschaft Sozialwissenschaftlicher Institute. Nonresponse Bias. Qualitätssicherung sozialwissenschaftlicher Umfragen* (pp. 209-251). Wiesbaden: Springer VS.

Richman, W. L., Kiesler, S., Weisband, S., & Drasgow, F. (1999). A Meta-Analytic Study of Social Desirability Distortion in Computer-Administered Questionnaires, Traditional Questionnaiers, and Interviews. *Journal of Applied Psychology, 84*(5), 754-775.

Schnell, R. (1997). *Nonresponse in Bevölkerungsumfragen: Ausmass, Entwicklung und Ursachen*. Opladen: Leske + Budrich.

Schwarz, N. (2008). The Psychology of Survey Response. In W. Donsbach & M. W. Traugott (Eds.), *The SAGE Handbook of Public Opinion Research* (pp. 374-387). Los Angeles, London: SAGE.

Singer, E. (1978). Informed Consent: Consequences for Response Rate and Response Quality in Social Surveys. *American Sociological Review, 43*(2), 144-162.

Singer, E., von Thurn, D. R., & Miller, E. R. (1995). Confidentiality Assurances and Response: A Quantitative Review of the Experimental Literature. *Public Opinion Quarterly, 59*(1), 66. doi:10.1086/269458

Steinbeißer, D., Bader, F., Ganser, C., & Schmitt, L. (2013). *Gruppenbezogene Menschenfeindlichkeit in München: Forschungsbericht des Instituts für Soziologie der Ludwig-Maximilians-Universität München*. Retrieved from http://www.ls4.soziologie.uni-muenchen.de/forschung/aeltere_projekte/gmf/bericht_gmf_18_10_2013.pdf

Stocké, V. (2004). Entstehungsbedingungen von Antwortverzerrungen durch soziale Erwünschtheit. Ein Vergleich der Rational-Choice Theorie und des Modells der Frame-Selektion. *Zeitschrift für Soziologie, 33*(4), 303-320.

Stocké, V. (2007). The Interdependence of Determinants for the Strength and Direction of Social Desirability Bias in Racial Attitude Surveys. *Journal of Official Statistics, 23*(4), 493-514.

Tourangeau, R., Rips, L. J., & Rasinski, K. A. (2000). *The Psychology of Survey Response*. Cambridge: Cambridge university press.

Tourangeau, R., & Yan, T. (2007). Sensitive Questions in Surveys. *Psychological Bulletin, 133*(5), 859-883.

Warwick, D. P., & Lininger, C. A. (1975). *The Sample Survey: Theory and Practice*. New York: McGraw-Hill.

Wildman, R. C. (1977). Effects of Anonymity and Social Setting on Survey Responses. *Public Opinion Quarterly, 41*(4), 74-79.

Wilson, E. B. (1927). Probable Inference, the Law of Succession, and Statistical Inference. *Journal of the American Statistical Association, 22*(158), 209-212. doi:10.1080/01621459.1927.10502953

Wolter, F. (2012). *Heikle Fragen in Interviews. Eine Validierung der Randomized Response-Technik*. Wiesbaden: Springer-VS.

Yang, M.-L., & Yu, R.-R. (2011). Effects of Identifiers in Mail Surveys. *Field Methods, 23*(3), 243-265. doi:10.1177/1525822X11399401

Zick, A., Wolf, C., Küpper, B., Davidov, E., Schmidt, P., & Heitmeyer, W. (2008). The Syndrome of Group-Focused Enmity: The Interrelation of Prejudices Tested with Multiple Cross-Sectional and Panel Data. *Journal of Social Issues, 64*(2), 363-383. doi:10.1111/j.1540-4560.2008.00566.x

Appendix

A Items on GFE

In the items on GFE the survey respondents were asked to gradually agree or disagree with these statements on a scale of 1 to 5 (“Totally disagree”, “Rather disagree”, “Neither agree nor disagree”, “Rather agree”, “Totally agree”). The items we used are:[12]

GFE-Items [1-low approval, 5-high approval]	Mean
To what extent do you agree with the following statements concerning disabled persons?	
In Germany, more should be done for disabled persons.[a]	3.845
I find many demands by disabled persons excessive.	2.120
Disabled persons receive too many privileges.	1.695
The following is about opinions on unemployed persons. In your opinion, to what extent do the following statements apply?	
Most unemployed persons make an effort to find a job.[a]	3.006
Unemployed persons who don’t find a job after longer search are themselves to blame.	2.459
I find it outrageous how permanently unemployed persons live a comfortable life at the society’s expense.	2.635
How some people systematically dodge work makes me angry.	3.857
Permanently unemployed persons should receive more support so they can find back to a working life.[a]	3.614
To what extent do you agree with the following statements concerning homeless persons?	
Most homeless persons have gotten into this situation through no fault of their own.[a]	3.178
Begging homeless persons should be removed from pedestrian areas.	2.666
Most homeless persons are disinclined to work.	2.452
To what extent do you agree with the following statements concerning homosexuality?	
Homosexuality is immoral.	1.453
Marriages between two women or, two men, respectively, should be allowed.[a]	3.994
Adopting children should stay forbidden to same-gender couples.	2.263

[a] Item scale is reversed in all calculations.

[12] The complete questionnaire (in German) can be found in the online appendix http://www.ls4.soziologie.uni-muenchen.de/forschung/zusatzinfos/sens_quest/.

GFE-Items [1-low approval, 5-high approval]	Mean
Now we would like to know to what extent you agree with the following statements.	
Jewish culture is an important part of Germany.[a]	3.610
What our country needs today is a tough and forceful assertion of German interests towards other countries.	2.581
A National Socialist dictatorship must never be allowed to happen again.[a, b]	4.911
National Socialism also had its good sides.	1.481
Like in nature, the strongest should always prevail in society.	1.683
Actually, Germans are, by nature, superior compared to other peoples.	1.335
Even today the influence of Jews is too great.	2.014
We should have a leader who governs Germany with a strong hand for the good of all.	1.272
Jews simply have something special and peculiar about them and don’t quite fit with us.	1.526
In the past months there has been a lot of debate in the public about immigration and integration. Therefore we are interested in the extent to which you agree with the following statements.	
Muslim culture fits with Germany well.[a]	2.549
Foreigners only come here to exploit our welfare state.	2.529
Naturalization of immigrated foreigners should be facilitated.[a]	3.010
The building of mosques enriches cultural life in Munich.[a]	2.888
When jobs become scarce, foreigners should be sent back home.	2.008
There are too many foreigners in our neighborhood.	2.025
An employer should be allowed to hire only Germans.	1.608
Having a variety of different religions is good for a country.[a]	3.804
I would only reluctantly register my child in a kindergarden/a school with many foreign children.	2.749
The customs and habits of Islam feel creepy to me.	2.847
Foreigners should leave Germany as fast as possible.	1.422
Foreigners living here threaten my own financial situation.	1.365
Munich is superalienated by foreigners to a dangerous degree.	1.779
In our society, too little regard is taken for minorities.	2.884
We have to protect our culture against the influence of other cultures.	2.499
There are too many Muslims in Germany.	2.425

[a] Item scale is reversed in all calculations.

[b] Due to the extremely high approval rate, there was almost no variation within the item. It is therefore not used for analyses which focus on single items (sections 4.2 and 4.3). It is, however, part of the National Socialism scale (Table 2 and section 4.4).

B Non-Sensitive Items

These items are used as test group in order to notice biased answers to less sensitive items.

Non-Sensitive Items	Mean
Please state how comfortable you are in your neighborhood.	
[very uncomfortable-1, very comfortable-5]	4.188
In your personal opinion, to what extent do the following statements apply to your nearer living environment?	
[low approval-1, high approval-5]	
The people here help each other.	3.380
The people here know each other well.	2.997
The people here get along well with each other.	3.719
All things considered, how satisfied are you with your life as a whole these days?	
[very unsatisfied-1, very satisfied-5]	3.975
How do you rate your current financial situation?	
[very bad-1, very good-5]	3.628
How much of what you want can you afford?	
[nearly nothing/nothing-1, nearly everything/everything-5]	3.448
Are you worried about your job?	
[very much-1, not at all-5, I don’t work (anymore)-Missing]	4.606
How much trust do you have in…	
[very little-1, very much-5, I don’t know-Missing]	
… the Bundestag?	2.919
… the German economy?	3.444
… churches?	2.283
… courts/the legal system?	3.329
… schools/the educational system?	3.110
… the current federal government?	2.860
… the police?	3.492

methods, data, analyses | Vol. 10(1), 2016, pp. 73-96 DOI: 10.12758/mda.2016.004

Retrospective Measurement of Students' Extracurricular Activities with a Self-administered Calendar Instrument

Peter Furthmüller
German Youth Institute (DJI)

Abstract

With the expansion of all-day schooling in Germany, students' extracurricular activities are being brought into greater focus in educational and social sciences. However, the diverse range of activities and individual biographies makes it difficult to gather data on the variety and periods of extracurricular activities in classroom-based surveys. This paper introduces a tailored calendar instrument that was applied by the Study on the Development of All-Day Schools (StEG) to retrospectively survey the activities of senior students since the fifth grade. Unlike other calendar applications, the calendar was not filled in by trained staff but self-administered by the students in a group setting. We discuss methodological issues regarding this procedure by examining the current state of research and by sharing experiences of tests of the instrument prior to the survey. By further analysing the survey data, we find no indication that the calendar task induces higher non-response as a result of overburdening the respondents. Calendar elements with an open-ended format resulted in heterogeneous reports, which were nonetheless mostly suitable for further analysis. According to our findings, the number of reported activities does not vary for students with longer intervals of retrospection. From our results, we conclude that a calendar instrument can be successfully applied in classroom-based surveys but should be implemented with a step-by-step procedure under a supervisor's guidance.

Keywords: calendar, retrospective, questionnaire, extracurricular activities, StEG

1 Introduction

Since the introduction and expansion of all-day schooling in Germany, new approaches to informal and non-formal learning have found their way into standard educational institutions. This expansion of educational content has been accompanied by an increase in autonomy and freedom at both an institutional and an individual level: schools are almost unconstrained by curricular, organisational and pedagogical prescripts when conceptualising their all-day programmes. Students, on the other hand, can now choose from a wider range of extracurricular activities not only outside the institutional context but also in school. Overall, these developments have led to more heterogeneous and individualised combinations of educational activities among adolescents. Students can pursue different activities at different ages and for different lengths of time. Investigating the complexity of extracurricular activity participation in large samples has been identified as an important but rare practice (see e.g. Feldman & Matjasko, 2005). The scarcity of appropriate studies might also be due to a lack of suitable research tools. It is a methodological challenge to survey extracurricular activities by theme in a precise and differentiated manner while at the same time recording their biographical sequencing including interruptions and overlaps.

In the *Study on the Development of All-Day Schools (StEG)*, researchers from the German Youth Institute (DJI) are investigating how participation in extracurricular activities affects the transition from school to vocational training at the "first threshold". Extracurricular activities are viewed from a biographical perspective and surveyed *retrospectively* among students in the ninth and tenth grade at secondary schools.[1] Taking into account the heterogeneity of educational biographies and with a view to improving the quality of retrospective reports, StEG applied a tailored calendar instrument[2] to study extracurricular activities since the fifth grade. Unlike most calendar applications, the calendar is not filled in by trained staff during face-to-face-interviews; instead, whole classes are sampled and the

[1] The DJI's project only conducts interviews at secondary schools that are not "Gymnasiums".

[2] Different terminology is used for calendar instruments in the relevant literature, e.g. event history calendar, illustrated life history, life events calendar, life history calendar, life history matrix, month-by-month calendar, time axes or timeline (see Glasner & van der Vaart, 2009).

Direct correspondence to
Peter Furthmüller, German Youth Institute (DJI), Study on the Development of All-Day Schools (StEG), Nockherstr. 2, 81541 Munich, Germany
E-Mail: furthmueller@dji.de

Acknowledgements: The author would like to thank Bettina Arnoldt and Christine Steiner for their support and assistance with this paper.

questionnaire is self-administered by the students. Thus, relatively large amounts of data can be collected in a short time, but it is not possible to check every retrospective report individually during the collection stage.

This paper seeks to introduce the calendar instrument for retrospective measurement of extracurricular activities and carries out analyses that are relevant to assessing data quality. In the first section, current methodological views on calendar instruments are summarised. The second and third sections contain background information on the study design and explain the construction and pretesting of the calendar instrument. The extent to which the calendar instrument is appropriate for collecting data about extracurricular activities is discussed in the concluding sections.

2 Current State of Research

2.1 The Methodological Challenges of Recalling and Dating

Compared to panel studies, retrospective surveys enable less expensive collection of life history information and faster availability of biographical data, and do not suffer from panel attrition. On the other hand, retrospective studies are considered risky in terms of memory bias (see Dex, 1995; Solga, 2001). Interdisciplinary research since the 1980s has led to a better understanding of how cognitive processes (e.g. for storing and retrieving memories), characteristics of events (e.g. passed time, salience), individual traits (e.g. gender, age) and contextual factors (e.g. interviewing situation, social desirability of topics) have a positive or negative effect on recall (see e.g. Dex, 1995; Sudman, Bradburn & Schwarz, 1996; Schwarz, 2007). It is essential, in this regard, that memories generally have to be reconstructed by mentally linking together multiple notions and that retrospection can be guided by cues (see Pohl, 2007, p. 34). Dating of events is an especially difficult task since only a few events are memorised with a "time stamp" (see Glasner & van der Vaart, 2008, p. 3) and respondents are usually required to recall an event as well as its context before they can date it (see Reimer, 2001, p. 16). Recall and dating can also be distinguished in terms of their error dimensions: while recall of events is mainly associated with memory *gaps*, the dating of events is also prone to timing *errors* (see ibid.; Auriat, 1993; Glasner, 2011).

How can a survey instrument be designed to minimise these risks? Balán et al. (1969) reported that the quality of retrospective data was improved through the use of a chronologically structured *schedule*. Freedman et al. (1988) referred to this idea when they developed the *life history calendar* after thorough pretesting. Like a coordinate system, the life history calendar provides a grid for dating biographical details by specifying axes for times and themes. The theme axis displays all domains, issues and events on which respondents are asked to provide information. In this way, the time axis enables each of the topics to be dated with predefined time units. Events can be related to each other and the recording of associated dates is simplified and standardised by the grid (see ibid., p. 41; for recent examples see e.g. Das, Martens & Wijnant, 2011; Rudin & Müller, 2013).

Based on findings from the cognitive sciences, R. F. Belli (1998) showed how calendar instruments support the process of memory reconstruction. Memory is organised like a network where recalled events can serve as cues to stimulate further memories. A calendric presentation of life events encourages respondents to retrieve information via different pathways, namely thematic "top-down retrieval", cross-thematic "parallel retrieval" and temporal "sequencing"[3] (see ibid., p. 394; Matthes, Reimer & Künster, 2007, p. 72). In these ways, calendar instruments promote a contextualisation of recall (Reimer, 2001, p. 100) since respondents can not only use temporal bounding strategies, but also relate events to each other.

2.2 Findings on the Quality of Calendar Data

The extent to which the calendar method leads to better data quality has been evaluated with non-experimental and quasi-experimental designs (for a synopsis see Glasner & van der Vaart, 2009). A comparison of results is difficult, however, because instruments and procedures for data collection differ considerably from each other in these studies. The calendar methods vary in terms of research themes, retrospective periods (from a few weeks to several years), the time scales that are used (e.g. years, months, days), graph designs (e.g. grids or timelines), survey techniques (e.g. face-to-face or CAT interviewing) and samples (e.g. adolescents or adults). Although mostly small and in some cases ambivalent effects on data quality are reported, the authors in general draw positive conclusions: memory performance seems to be improved by the use of calendar tools, which were successfully applied to the recollection of various events over the course of respondents' lifetimes in different studies (see e.g. Freedman et al., 1988; Caspi et al., 1996; Martyn, 2009). Specifically with regard to *educational* activities, Dürnberger, Drasch & Matthes conclude that a contextualised approach supports recall for retrospective periods of five years (see ibid., 2011). Calendar instruments produce more complete reports, particularly if events are of the distant past or difficult to remember (see Goldman, Moreno & Westhoff, 1989; Becker & Sosa, 1992; van der Zouwen, Dijkstra & van der Vaart, 1993; Engel, Keifer & Zahm, 2001; Belli et al. 2004; van der

[3] For example: a top-down retrieval strategy could result in a (fictional) statement like "When I attended all-day school, I always attended the computer courses". A cross-thematic parallel retrieval could result in "In addition to all-day school I attended a private music school once per week". A temporal sequencing strategy could result in "After I stopped taking guitar lessons, I joined our school band".

Vaart, 2004; Yoshihama et al., 2005). Depending on their topic, calendric retrospections show medium to high consistency with data that was collected earlier from the same respondents (see Freedman et al., 1988; Caspi et al., 1996; Lin, Ensel & Wan-Foon, 1997; Belli, Shay & Stafford, 2001). It was also found that calendar instruments lead to more accurate dating, although the advantage over conventional question lists is sometimes small (see Becker & Sosa, 1992; van der Zouwen et al., 1993; Belli et al., 2007; Sayles, Belli & Serrano, 2010). Calendar instruments are also supposed to reduce heaping, i.e. respondents rounding time periods to typical values such as 6, 12 or 24 months rather than reporting the precise date (see Goldman et al., 1989; Becker & Diop-Sidibé, 2003). However, it was not always possible to confirm the reduction of heaping (see van der Vaart, 2004). The calendar method is particularly recommended if complex biographies with overlapping events need to be reconstructed (see Engel et al., 2001; Belli et al., 2007). On the other hand, complicated histories have also proven to be one cause of low reliability in calendar data (see Callahan & Becker, 2012). Collection, coding and clearing of calendar data is more demanding and advantages over conventional question lists depend on the survey topic. But Glasner and van der Vaart (2009) maintain that calendar instruments have never led to poorer quality of retrospective data than question lists (see ibid., p. 343). More recent findings suggest, however, that this may not be the case if events can be recalled very easily by the respondents (see Belli, Bilgen & Baghal, 2013).

It is repeatedly pointed out in the relevant literature that interviewers and respondents described calendar instruments as helpful tools for recall and conversation. Furthermore, respondents were often motivated to record their biographies as accurately and completely as possible with such a template (see Freedman et al., 1988; Hoppin et al., 1998; Engel et al., 2001; Belli et al., 2004; Martyn, Reifsnider & Murray, 2006; Belli et al., 2007). Some authors reason, however, that calendar instruments may increase non-response because they look particularly demanding at first sight and could discourage respondents (see Glasner & van der Vaart, 2009). Prior to their telephone interviews, van der Vaart and Glasner (2005) sent a calendar instrument to some of the survey participants as a memory aid during the conversation. While the response rate was only 39 percent in the group which had received a calendar, it reached 67 percent amongst those who had no access to such a supplement (quoted in Glasner & van der Vaart, 2009, p. 344). Belli et al. (2001), on the other hand, did not find an essential difference between the response rates of traditional question list surveys and calendar interviews (see ibid. p. 52). Taking additional results into consideration, the findings on non-response in calendar instruments do not allow a definite conclusion (see e.g. Mortimer & Johnson, 1999; Yoshihama et al., 2005; Martyn et al., 2006; Cotugno, 2009).

In most of the present studies, the calendars were filled in by trained interviewers and not respondents themselves. Likewise, CATI designs contain calendric grid views primarily for the interviewers to reveal gaps and inconsistencies in biographical reports and to clarify them.[4] There only appear to be a few studies in which calendar instruments were administered by the respondents themselves. In a panel study, Mortimer and Johnson (1999) sent out a *life history calendar* annually to collect data about important events and activities in the respondents' lives. Cotugno (2009) successfully applied a self-administered calendar in a question list paper-and-pencil-interview with more than 200 participants. Martyn and Martin (2003) interviewed adolescents about sensitive topics like drug abuse and sexual behaviour using an *event history calendar (EHC)*. To fill out the form, the participants could choose the support of a trained interviewer or administer the EHC themselves. About 86 percent (n=43) decided to fill in the EHC autonomously. A self-administered calendar instrument seems to comply with the need for confidentiality and therefore reduces social desirability bias when intimate and private domains are being surveyed. "In addition, when one-on-one interviews are not required, the EHC can be administered to groups of participants like large-scale surveys are administered, obtaining comprehensive data while saving time and money" (see Martyn, 2009, p. 73). Based on comparative data from prior surveys, personal interviews and plausibility analyses, Martyn views the self-administered calendar method as a suitable instrument for youth research (see ibid. 2009).

As an interim summary, we can note that calendar instruments are not only recommended because they improve recall and stimulate more complete reports, but also because they are an effective survey technique for recording a multitude of events in a compact format. Therefore, calendar instruments seem particularly suited for retrospective collection of data about complex biographies of extracurricular activities. Self-administered calendars have already been employed successfully in the past and their application in a classroom setting seems promising.

3 Data Base

StEG is a research programme being carried out by a consortium of several institutions that conducts surveys on all-day schooling in Germany on a regular basis. The participating institutions are the German Institute for International Educational Research (DIPF), the Institute for School Development Research (IFS), Justus Liebig University Giessen (JLU) and the German Youth Institute (DJI). StEG is sponsored by the German Federal Ministry of Education and Research (BMBF).

Based on a representative survey of head teachers from all over Germany in 2012, the participating institutions of StEG carry out in-depth studies on the effects

[4] No significant differences could be found in face-to-face and CATI calendar interviews (see Freedman et al., 1988, p. 65).

of all-day schooling. The DJI's in-depth study is conducted with a subsample of secondary schools ("Schulen der Sekundarstufe I") that are not grammar schools ("Gymnasiums") and participated in the survey of head teachers.[5] 65 all-day schools from 12 federal states agreed to in-depth studies as part of the StEG programme.[6] This article contains the data of 1,901 students in graduating classes who were interviewed with a self-administered paper-and-pencil questionnaire in spring 2013. Of these adolescents, 608 were in the ninth grade and 1,292 were in the tenth grade.[7] The questionnaire consisted of a tailored calendar instrument and a classic question list on different topics, e.g. school careers and family background.

4 Pretesting and Tailoring the Calendar

The calendar instrument was intended to enable analysis and profiling of extracurricular activities, i.e. to collect data on the *kinds* of activities carried out by students as well as the biographical *periods* of activity. Like the question list section of the questionnaire, the calendar was supposed to be filled in by the respondents as autonomously as possible. To identify problems, the instrument was pretested with students in three different classes prior to the survey.[8] The test gave no indication that the respondents were overwhelmed by the task of dating events with a calendar instrument in principle. It became clear, however, that respondents often had to deal with subtasks that were implicitly imposed or assumed by the instrument, which made it more difficult to fill in the grid as a whole (e.g. converting recalled dates to the calendar scale). Pretesting provided valuable feedback and suggestions for how to improve the calendar instrument. The most important steps for creating the final instrument (see Appendix A-1 and A-2) are explained in the following sections.

Theme axis: open question about activities

While other calendar instruments typically cover many domains on the theme axis, the calendar in StEG focusses on extracurricular activities to keep the task as simple as possible. The theme axis of the pretested calendar included 12 categories

[5] The DJI's survey was carried out in cooperation with the IEA Data Processing and Research Center (DPC) in Hamburg.

[6] The participating schools are from Bavaria (n=15), Baden-Württemberg (n=7), Brandenburg (n=4), Bremen (n=4), Hesse (n=4), Mecklenburg-West Pomerania (n=4), Lower Saxony (n=7), North Rhine-Westphalia (n=4), Rhineland-Palatinate (n=7), Saxony (n=3), Saxony-Anhalt (n=4) and Thuringia (n=2).

[7] Depending on their individual school career, students in the ninth and tenth grades are typically 15 to 16 years old.

[8] One ninth-grade class at a lower secondary school and two preparation classes at a vocational school ("Berufsgrundbildungsjahr" and "Berufsvorbereitungsjahr" respectively).

which brought together a wide range of activities from different domains. To match each recalled activity to a date, the respondents first had to assign them to a specific category by reading and interpreting the labels provided. If the adolescents were simultaneously engaged in various activities within a category and at a certain date, they were supposed to fill in the number of parallel activities for that point in time. *Matching* and at the same time *counting* activities turned out to be a very difficult task and the calendar was substantially revised in this regard for the main survey: instead of asking about the type of activities with a matching task, predefined categories were abandoned in favour of an open-ended format (see Appendix A-2). Since the respondents could list every single activity separately with the new format and did not need to summarise them in any way, counting also became unnecessary in the final instrument.

Time axis: grades as a personalised time scale

In the pretest, students preferred the calendar's time axis to be scaled according to school grades. Grades seem especially suitable for dating extracurricular activities at schools since they correspond to the institutionalised schedule there. They structure every individual's school career and provide further cues for recall. It cannot be assumed, however, that every respondent passes through school grades in the same sequence: while some of the students have had a regular career, other students may have repeated or skipped specific grades. To take individual school careers into account, the calendar's time axis was not completely labelled with preset values in its final implementation. Instead, the grades had to be filled in by the students according to their school career after the fifth grade. They were specifically instructed to write down a grade multiple times if it had been repeated.

Narrowing the reference frame of relevant activities

Asking about activity types using an open-ended question involves the risk of respondents ignoring activities they consider inappropriate and reporting trivial ones instead (for further discussion see also Sudman et al., 1996, p. 55). To convey the range of relevant activities more clearly to the respondents, an additional task preceded the calendar instrument in the main study: the respondents were simply asked to recall what they did on a regular basis apart from attending school lessons and to record their thoughts on a separate sheet of the questionnaire ("memo", "memorandum" or "Merkzettel", see Appendix A-1). The memo contained one column for recalled activities from the in-school domain, and another one for activities from the non-school domain. In order to establish boundaries for the scope of activities that were of interest to StEG, a *reference frame* was outlined in simple terms: the respondents were instructed to only write down activities that were not lessons and that they had attended regularly (at least once per week) and steadily (for the duration of at least one term) at any point since fifth grade. Moreover, the range of relevant activities was specified by the supervisors through explanations and examples. It turned out to be very helpful to add an index to the columns' rows, since it encouraged respondents to fill in their activities more consistently.

Detailed instructions and step-by-step progress

Written guidance for the calendar was often not read by the pretest participants or was perceived as being too difficult to understand. If students tried to clarify uncertainty at all, they did so by asking their seatmates or the supervisor. This created a disturbance, irritated other students and distracted the whole class from filling in the questionnaire. Thus, a different approach was chosen for the main survey: written instructions were almost completely replaced by guidelines for verbal directives given by the supervisors. The supervisors instructed the respondents step by step and presented an example of each task on a poster-sized calendar. Care was taken to ensure that the students began each subtask together and got to ask questions if necessary. The whole procedure was organised as follows:

1. Filling in the memo
2. Labelling the calendar's time axis with completed grades
3. Transcribing[9] relevant activities from the memo to the calendar's theme axis
4. Dating each activity and completing the calendar

Analogous to the memo's separate columns for in-school and non-school activities, two calendar grids were included on different pages of the final questionnaire. Such a step-by-step process may seem contrary to the rationale of a typical calendar, which integrates all recall tasks into one instrument. However, the memo was not only supposed to convey an idea about the relevant activities, but also to provide the basis for the dating task, since the respondents have to recall some activities first before they can date them (see also Reimer, 2001, p. 16) and are thereby provided with cues for further activities.

[9] The designation "Merkzettel" was supposed to connote that the memos' content is only temporary and auxiliary for the respondents. In the main survey, the instructors merely asked the students to copy their activities from the memo to the calendar ("Bitte übertragt alle schulischen Angebote aus dem Merkblatt auf S.4 in die leeren Zeilen auf der linken Seite unter der Anleitung 'Trage hier deine Aktivitäten ein'"). The respondents were not asked to keep the entries on the memo and calendar in sync nor were they told not to add further entries to the calendar. For future applications of the procedure, however, it might be beneficial to invite respondents more explicitly to add further activities while filling out the calendar.

5 Assessment of Data Quality

In the following, an attempt will be made to assess the quality of the calendar data and to evaluate the validity of the school careers and extracurricular activities recorded with the instrument. With calendar instruments the *gold standard* of validity assessment includes a comparison of retrospective data and similar information gathered from the same individuals at an earlier date (see Alwin, 2009, p. 283). An example of this rare type of evaluation can be found in Belli et al. (2001). However, an alternative strategy has to be pursued if retrospective interviewing with a calendar instrument is applied due to the fact that no data was gathered in the past, as is usually the case. “Establishing the validity of survey measurement is difficult because within a given survey instrument there is typically little available information that would establish a criterion for validation” (Alwin, 2009, p. 282).

In StEG too, no longitudinal data is available that would allow a *gold standard* analysis. But verifiable quality criteria can be derived from the current state of research and experiences in the pretest: firstly, it can be assumed that overburdening students induces frustration or uncooperativeness with regard to the calendar task that, in the end, will lead to higher non-response. Therefore, high non-response would indicate that the calendar is too complicated and not suited for classroom-based self-administered surveying. Secondly, the quality of the calendar must be judged by the degree of activities entered in the calendar that are relevant, plausible and applicable for the research topic. Thirdly, the calendar would be insufficiently qualified for ascertaining extracurricular activities if biographical patterns can be ascribed to shortcomings of memory and recall. These issues will now be examined in greater detail.

5.1 Willingness to Fill in the Calendar

Despite the adjustments that were made after the pretest, the calendar instrument confronts respondents with an unfamiliar and relatively complicated task that demands a high level of cooperation. Due to the classroom-based nature of the survey, there is no individual supervision and few ways to deal with frustration and refusal to cooperate. With a self-administered instrument, data quality primarily depends on a sufficient number of completed calendars, but the data does not allow direct identification of uncooperative respondents: if students did not fill in the calendar, the reason may be deliberate refusal or the simple fact that no activities were carried out or could be recalled. A clearer picture emerges if the step-by-step progress is taken into account when assessing non-response (see Appendix B-1).

The differences between the filled-in segments on the memo page show that more respondents (n=1,901) reported activities for the non-school domain (92.7%) than for the school domain (88.5%). Only a few of the adolescents who filled in the memo omitted to list activities on the corresponding calendar page (in-school: 0.4%; non-school: 1.3%) and almost all of the students also dated the activities they entered on the theme axis of the calendar. All students were asked by the supervisors to fill in the time axis of the calendar with their completed grades, regardless of whether they could recall activities or not. This instruction was followed by 97.5 percent of all 1,901 students for the in-school calendar and 95 percent for the non-school calendar respectively.

If lack of cooperation was the main cause for missing data, the proportion of respondents who filled in *neither* the calendar page for activities in school *nor* the page for non-school activities (no table) should be high: 2.5 percent of the respondents did not supply dated activities for any domain, but only 0.9 percent repeatedly neglected the instruction to fill in the time axes with grades. This implies that there are almost no students who refused to cooperate at every step of the procedure. For comparison: the average proportion of non-response in the question list part of the questionnaire is 1.9 percent (n_{max}=1,901 with 358 items). Hence, willingness to fill in the calendar instrument can be regarded as positive. Missing data seems instead to be caused by a lack of activities being carried out or recalled within the specified boundaries.

5.2 Thematic Classification of Recalled Activities

The calendar was expected to yield heterogeneous textual data on a broad range of different activities that needed to be standardised in some way. Prior to the survey, a coding scheme was developed based on the scope of relevant activities, earlier findings and additional Internet research. The scheme was intended to enable individual activities to be recorded distinctively whilst mapping them to thematic categories. It includes 372 detailed activities in 15 categories. In a sense, the coding scheme also represents an *ex ante* explication of the thematic variety assumed in students' activities.

Appendix A-3 shows the proportions of adolescents (n=1,901) who reported activities from the 15 thematic categories. The distribution parallels known results in some respects (on the dominant role of sports, for example, see Züchner, Arnoldt & Vossler, 2008; Grgic & Züchner, 2013), and differences between the in-school and non-school domains suggest that the calendar data is able to portray social realities. As stated above, only a rudimentary assessment of the "correctness" of reports is possible since no precedent data is available from these respondents.[10] But the StEG questionnaire also included some extra questions on activities that were supposed to be more difficult to survey with a calendar. A comparison of both survey methods demonstrates the limitations of the instrument: with the calendar, only three percent of the adolescents autonomously recorded that they had once attended "homework support", a more formal programme that is typical at all-day schools. However, when they were asked directly, 23 percent of the students confirmed that they had attended "homework support". A third of the students affirmed the direct question "Do you earn money with a side part-time job?", but less than four percent recorded such a job in the calendar. If activities in school are strongly associated with formal lessons or if non-school activities are weakly associated with leisure, they may have been cued less often by the calendar. As expected prior to the survey a specific query about activities that may not be fully covered by the boundaries of relevant activities presented in the memo or the calendar seems a useful precaution to prevent missing data.

Furthermore, Appendix A-3 shows the proportion of students' statements that were difficult or could not be assigned to one of the 15 thematic categories. The *Miscellaneous* category includes all activities that could be identified as some sort of informal or non-formal practice, but were not more precisely attributable (e.g. statements like "project group" or "all-day programme"). Around one in ten students recorded in-school activities that were completely unusable and were therefore classified as *invalid* (mostly because they could not be read or were crossed out). Non-school activities were treated as invalid mostly due to deviations from the reference frame (e.g. statements like "meeting friends", "chilling out", "parties" or "shopping"). Approximately 16 percent of all respondents made an entry for either the in-school or non-school domain which was not usable for further analysis. The invalid and vague activities can also be related to the number of total records: in total, 5,181 entries were made for in-school activities of which 345 (6.7%) were considered invalid and 149 (2.9%) could not be assigned to the scheme. The respondents entered a total of 4,598 records in the calendar for non-school activities. Of these, 200 (4.3%) were invalid and 96 (2%) were too inaccurate for further classification. Hence, the vast majority of the information provided could be assigned to either a specifically defined activity or to at least one thematic category.

[10] It seems worth noting that, inter alia, the comparison of activity distributions from different studies depends heavily on the specified boundaries and the reference frame that is applied: StEG was gathering data on extracurricular activities that were regularly practised from the fifth grade for at least one term. In contrast, other studies sometimes concentrate on activities that are relevant at the time of the survey, and hence also include shorter periods of activity (for example).

5.3 Number of Recalled Activities

The calendric survey is supposed to reduce the risk of activities being reported incompletely, particularly for grades that are dated far in the past. The calendar would perform unsatisfactorily in this regard if fewer activities were reported for larger retrospective intervals.

The students reported significantly more activities for the non-school domain than the in-school one (in-school: M=0.81 (SD=0.97); non-school: M=1.38 (SD=1.25); n=1,901; p<.001)[11] and grades without activities are more frequent within the in-school domain (see Appendix A-4). Based on the calendar data, a distinct trend can be noted with regard to the extracurricular activities in both domains. While respondents entered fewer in-school activities for higher grades (r=–.11; n=1,901; p<.001), the sample differs with regard to non-school activities: on the one hand, the proportion of adolescents who did not report any activity in higher grades at all also increases slightly; on the other hand, the number of students who carried out multiple activities at the same time is higher. The non-school domain tends to feature more numerous activities in higher grades (r=.07; n=1,901; p<.001). The calendar data seems consistent with earlier findings on extracurricular activities by adolescents in this respect. Using a previous data base, StEG has already demonstrated an age-dependent decline of extracurricular activities at all-day schools by analysing longitudinal data of students from fifth to ninth grade (see Arnoldt, Furthmüller & Steiner, 2013; Züchner & Arnoldt, 2011). The developing activity pattern in the non-school domain corresponds to the findings of the study "Medien, Kultur und Sport bei jungen Menschen (MediKuS)" which revealed a common shift in activities, meaning that growing up does not necessarily imply a withdrawal from activities, but rather involves a change of contexts (see Grgic & Züchner, 2013, p. 258).

Neither the in-school nor the non-school data show a linear growth of inactive students for grades further back in the past. However, remarkably few activities were recorded for the fifth grade as the lower boundary of the retrospective interval. The sampling design can be utilised to examine whether this deviation should be ascribed to a lack of activity or rather to the passing of time, difficulties of recall and deficiencies of the instrument: the sample is composed of students for whom different periods of time have elapsed since the fifth grade. Adolescents who were in the ninth grade at the time of the survey and followed a regular career in school were fifth-graders four years prior to the interview. But respondents in the tenth grade or students who needed to repeat a grade must look back at least five years in order to recall the events of the fifth grade. If the calendar instrument could not provide sufficient support for recall, respondents with larger retrospective intervals should have recorded fewer activities for the fifth grade. This assumed relation

[11] In the following, the number of activities in repeated grades was not accounted for.

between activity count and retrospective interval size is examined with a Poisson regression model (see Appendix B-2 and B-3). To control other influences on extracurricular practice, the model includes independent variables that have been shown to be relevant for the attendance of all-day programmes (see e.g. Steiner, 2011; Steiner & Fischer, 2011; Züchner et al., 2008).

In school, adolescents from former East Germany carried out more extracurricular activities in the fifth grade, as did students with a migration background and those of high socioeconomic status. School context seems to be relevant, since students recorded fewer activities if they had attended fifth grade at a different school. The higher activity count of students who participated in all-day programmes in the fifth grade is unsurprising, but a promising sign of the retrospections' validity. Respondents who attended the fifth grade six years prior to the survey reported fewer activities than the reference group with a retrospective interval of four years, but not to a significant level. Thus, despite some other reasonable findings, no statistically significant influence of retrospective interval lengths was observed with regard to the recalled number of in-school activities. For the non-school domain, adolescents from the eastern states of Germany, with a migration background or of low socioeconomic status recorded significantly *fewer* activities than the reference group. Students who had attended all-day programmes in the fifth grade were also more active outside of school. While these controlled variables exhibit different patterns in relation to the number of in-school and non-school activities in the fifth grade, the retrospective intervals do not: the number of recalled non-school activities does not differ significantly depending on how far the students had to look back.

6 Summary

Self-administering a calendar instrument not only involves recalling and dating, but a wide range of secondary tasks that are sometimes carried out simultaneously by the respondents. A step-by-step approach and verbal guidance by supervisors proved to be the most important measure to enable the application of a self-administered calendar in a classroom-based survey. By gradually labelling the calendar's axes, the students created a personalised scheme which relieved them of most of the work of reading and interpretation. However, while this decreased difficulty for the respondents, it increased complexity for the researchers. With its open-ended components, the calendar requires greater effort to code and process the data, since variability usually needs to be standardised in some way prior to analysis. By far the biggest challenge in surveying extracurricular experiences with an open-ended format is to convey an idea of the relevant activities the respondents are supposed to enter. The preceding memorandum turned out to be a practical solution that further facilitated the calendar procedure for the students. Although most of the recorded activities could be interpreted and categorised after the survey, the accuracy of listed activities and dates remains undetermined since no data is available for comparison. However, the non-response rates give no indication of major problems due to frustration or a lack of cooperation. Furthermore, no differences in the number of recalled activities were found for respondents with longer retrospective intervals up to seven years. These results also seem to be an affirmation that respondents were supported in their main task of recalling and dating activities, and that, in contrast to the pretest, they were no longer overwhelmed by the instrument or diverted by complicated secondary tasks. Against this background, the calendar instrument has proven successful in gathering retrospective data about the complex biographies of extracurricular activities.

References

Alwin, D. F. (2009). Assessing the validity and reliability of timeline and event history data. In R. F. Belli, F. Stafford & D. F. Alwin (Eds.), *Calendar and time diary methods in life course research* (pp. 277-301). Thousand Oaks: Sage.

Arnoldt, B., Furthmüller, P. & Steiner, C. (2013). *Ganztagsangebote für Jugendliche – Eine Expertise zum Stellenwert von Ganztagsangeboten für Schüler/innen ab der 9. Klasse*. München: Deutsches Jugendinstitut (DJI).

Auriat, N. (1993). My wife knows best – a comparison of event dating accuracy between the wife, the husband, the couple, and the Belgium population register. *Public Opinion Quarterly*, *57*(2), 165-190.

Balán, J., Browning, H. L., Jelin, E. & Litzler, L. (1969). A computerized approach to the processing and analysis of life histories obtained in sample surveys. *Behavioral Science*, *14*(2), 105-120.

Becker, S. & Diop-Sidibé, N. (2003). Does use of the calendar in surveys reduce heaping? *Studies in Family Planning*, *34*(2), 127-132.

Becker, S. & Sosa, D. (1992). An experiment using a month-by-month calendar in a family planning survey in Costa Rica. *Studies in Family Planning*, *23*(6), 386-391.

Belli, R. F. (1998). The structure of autobiographical memory and the event history calendar: Potential improvements in the quality of retrospective reports in surveys. *Memory*, *6*(4), 383-406.

Belli, R. F., Shay, W. L. & Stafford, F. P. (2001). Event history calendars and question list surveys: A direct comparison of interviewing methods. *Public opinion quarterly*, *65*(1), 45-74.

Belli, R. F., Lee, E., Stafford, F. P. & Chou, C.-H. (2004). Calendar and question-list survey methods: Association between interviewer behaviors and data quality. *Journal of Official Statistics*, *20*(4), 185-218.

Belli, R. F., Smith, L. M., Andreski, P. M. & Agrawal, S. (2007). Methodological comparisons between CATI event history calendar and standardized conventional questionnaire instruments. *Public Opinion Quarterly*, *71*(4), 603-622.

Belli, R. F., Bilgen, I. & Baghal, T. A. (2013). Memory, communication, and data quality in calendar interviews. *Public opinion quarterly*, *77*(S1), 194-219.

Callahan, R. & Becker, S. (2012). The reliability of calendar data for reporting contraceptive use: evidence from rural Bangladesh. *Studies in Family Planning*, *43*(3), 213-222.

Caspi, A., Moffitt, T. E., Thornton, A., Freedman, D., Amell, J. W., Harrington, H., Smeijers, J. et al. (1996). The life history calendar: A research and clinical assessment method for collecting retrospective event-history data. *International Journal of Methods in Psychiatric Research*, *6*, 101-114.

Cotugno, M. Z. (2009). Creating a self-administered event history calendar. *Survey Practice*, *2*(9). Retrieved from www.surveypractice.org

Das, M., Martens, M. & Wijnant (2011). Survey Instruments in SHARELIFE. In M. Schröder (Ed.), *Retrospective data collection in the survey of health, ageing and retirement in Europe. SHARELIFE Methodology* (pp. 20-28). Mannheim: Mannheim Research Institute for the Economics of Aging (MEA).

Dex, S. (1995). The reliability of recall data: A literature review. *Bulletin de Methodologie Sociologique*, *49*(1), 58-89.

Dürnberger, A., Drasch, K. & Matthes, B. (2011). Kontextgestützte Abfrage in Retrospektiverhebungen. *Methoden - Daten - Analysen*, *5*(1), 3-35.

Engel, L. S., Keifer, M. C. & Zahm, S. (2001). Comparison of a traditional questionnaire with an icon/calendar-based questionnaire to assess occupational history. *American Journal of Industrial Medicine*, *40*, 502-511.

Feldman, A. F. & Matjasko, J. L. (2005). The role of school-based extracurricular activities in adolescent development: A comprehensive review and future directions. *Review of Educational Research*, *75*(2), 159-210.

Freedman, D., Thornton, A., Camburn, D., Alwin, D. & Young-DeMarco, L. (1988). The life history calendar: A technique for collecting retrospective data. *Sociological Methodology*, *18*, 37-68.

Glasner, T. (2011). *Reconstructing event histories in standardized survey research: cognitive mechanisms and aided recall techniques*. University of Amsterdam, Amsterdam. Retrieved from www.researchgate.net

Glasner, T. & Vaart, W. van der. (2008). *Cognitive processes in event history calendar interviews: A verbal report analysis*. In S. Balbi, G. Scepi, G. Russolillo & A. Stawinoga (Eds.), Proceedings of the Seventh International Conference on Survey Methodology. Retrieved from library.wur.nl

Glasner, T. & van der Vaart, W. (2009). Applications of calendar instruments in social surveys: A review. *Quality and Quantity*, *43*(3), 333-349.

Goldman, N., Moreno, L. & Westoff, C. F. (1989). Collection of survey data on contraception: An evaluation of an experiment in Peru. *Studies in Family Planning*, *20*(3), 147-157.

Grgic, M. & Züchner, I. (Eds.). (2013). *Medien, Kultur und Sport. Was Kinder und Jugendliche machen und ihnen wichtig ist. Die MediKuS-Studie*. Weinheim/Basel: Beltz Juventa.

Hoppin, J. A., Tolbert, P. E., Flagg, E. W., Blair, A. & Zahm, S. H. (1998). Use of a life events calendar approach to elicit occupational history from farmers. *American journal of industrial medicine*, *34*(5), 470-476.

Lin, N., Ensel, W. M. & Wan-Foon, G. (1997). Construction and use of the life history calendar: Reliability and validity of recall data. In I. H. Gotlib & B. Wheaton (Eds.), *Stress and adversity over the life course: Trajectories and turning points* (pp. 249-272). Cambridge: Cambridge University Press.

Martyn, K. K. (2009). Adolescent health research and clinical assessment using self-administered event history calendars. In R. F. Belli, F. Stafford & D. F. Alwin (Eds.), *Calendar and time diary methods in life course research* (pp. 69-86). Thousand Oaks: Sage Publications.

Martyn, K. K. & Martin, R. (2003). Adolescent sexual risk assessment. *Journal of Midwifery & Women's Health*, *48*(3), 213-219.

Martyn, K. K., Reifsnider, E. & Murray, A. (2006). Improving adolescent sexual risk assessment with event history calendars: A feasibility study. *Journal of Pediatric Health Care*, *20*(1), 19-26.

Matthes, B., Reimer, M. & Künster, R. (2007). Techniken und Werkzeuge zur Unterstützung der Erinnerungsarbeit bei der computergestützten Erhebung retrospektiver Längsschnittdaten. *Methoden - Daten - Analysen*, *1*(1), 69-92.

Mortimer, J. T. & Johnson, M. K. (1999). Adolescent part-time work and postsecondary transition pathways in the united states. In W. R. Heinz (Ed.), *From education to work. cross-national perspectives* (pp. 111-148). New York: Cambridge University Press.

Pohl, R. (2007). *Das autobiographische Gedächtnis: Die Psychologie unserer Lebensgeschichte*. Stuttgart: Kohlhammer Verlag.

Reimer, M. (2001). *Die Zuverlässigkeit des autobiographischen Gedächtnisses und die Validität retrospektiv erhobener Lebensverlaufsdaten: Kognitive und erhebungspragmatische Aspekte*. Materialien aus der Bildungsforschung (Vol. 71). Berlin: Max-Planck-Institut für Bildungsforschung.

Rudin, M. & Müller, C. (2013). Kann es sein, dass ich das Beginndatum falsch erfasst habe? *Methoden - Daten - Analysen*, *7*(3), 433-463.

Sayles, H., Belli, R. F. & Serrano, E. (2010). Interviewer variance between event history calendar and conventional questionnaire interviews. *Public Opinion Quarterly*, *74*(1), 140-153.

Schwarz, N. (2007). Cognitive aspects of survey methodology. *Applied Cognitive Psychology*, *21*(2), 277-287.

Solga, H. (2001). Longitudinal surveys and the study of occupational mobility: Panel and retrospective design in comparison. *Quality & Quantity*, *35*(3), 291-309.

Steiner, C. (2011). Teilnahme am Ganztagsbetrieb. In N. Fischer, H. Holtappels, E. Klieme, T. Rauschenbach, L. Stecher & I. Züchner (Eds.), *Ganztagsschule: Entwicklung, Qualität, Wirkungen. L*ängsschnittliche *Befunde der Studie zur Entwicklung von Ganztagsschulen (StEG)* (pp. 57-75). Weinheim/Basel: Juventa.

Steiner, C. & Fischer, N. (2011). Wer nutzt Ganztagsangebote und warum? *Zeitschrift für Erziehungswissenschaft*, *14*(3), 185–203.

Sudman, S., Bradburn, N. M. & Schwarz, N. (1996). *Thinking about answers: The application of cognitive processes to survey methodology*. San Fransisco: Jossey-Bass.

Vaart, W. van der. (2004). The time-line as a device to enhance recall in standardized research interviews: A split ballot study. *Journal of Official Statistics*, *20*(2), 301-317.

Vaart, W. van der & Glasner, T. (2005). Enhancing recall accuracy in surveys by cognitively tailored timeline methods: A record check study. Paper presented at the First EASR Conference, Barcelona.

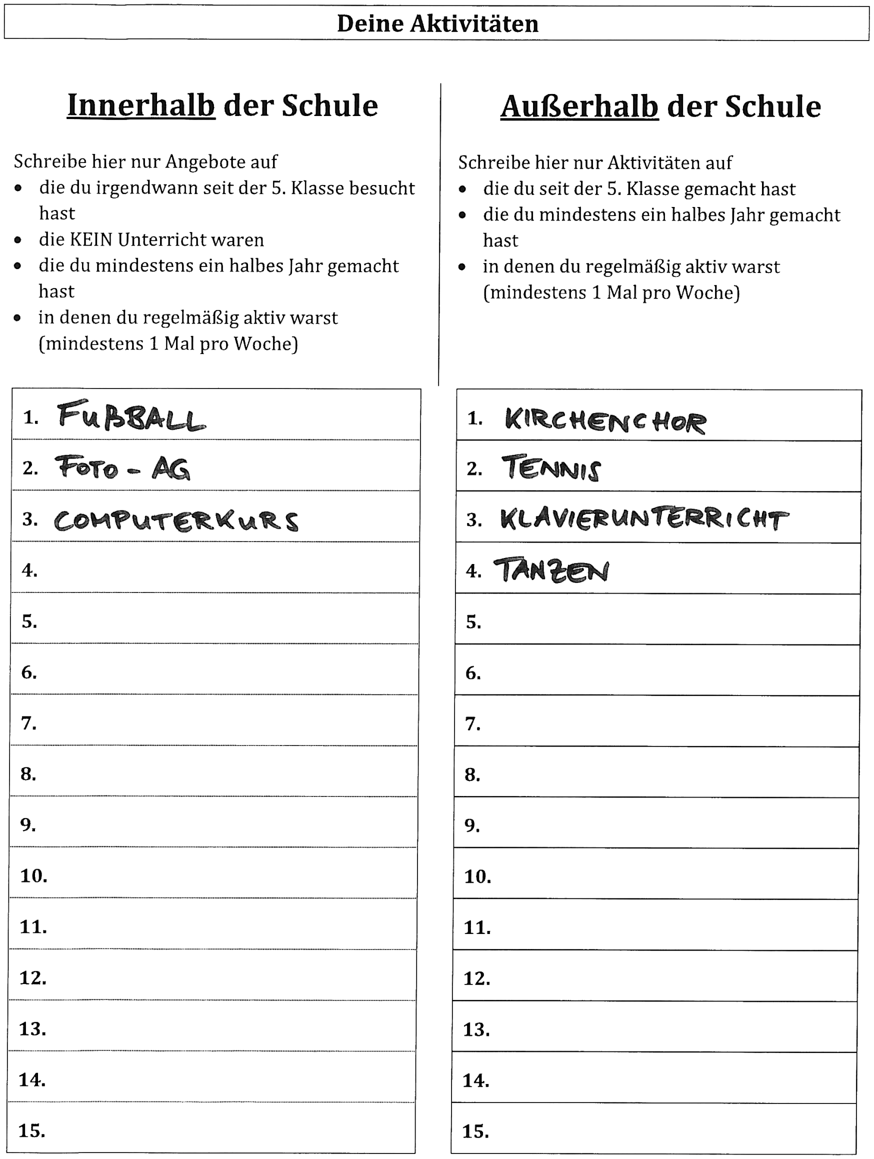
Yoshihama, M., Gillespie, B., Hammock, A. C., Belli, R. & Tolman, R. M. (2005). Does the life history calendar method facilitate the recall of intimate partner violence? Comparison of two methods of data collection. *Social Work Research*, *29*(3), 151-163.

Zouwen, J. van der, Dijkstra, W. & Vaart, W. van der. (1993). Effects of measures aimed at increasing the quality of recall data. *Bulletin de Methodologie Sociologique*, *39*(1), 3-19.

Züchner, I. & Arnoldt, B. (2011). Schulische und außerschulische Freizeit- und Bildungsaktivitäten. In N. Fischer, H. Holtappels, E. Klieme, T. Rauschenbach, L. Stecher & I. Züchner (Eds.), *Ganztagsschule: Entwicklung, Qualität, Wirkungen. L*ängsschnittliche *Befunde der Studie zur Entwicklung von Ganztagsschulen (StEG)* (pp. 267-290). Weinheim/Basel: Juventa.

Züchner, I., Arnoldt, B. & Vossler, A. (2008). Kinder und Jugendliche in Ganztagsangeboten. In H. Holtappels, E. Klieme, T. Rauschenbach & L. Stecher (Eds.), *Ganztagsschule in Deutschland. Ergebnisse der Ausgangserhebung der Studie zur Entwicklung von Ganztagsschulen (stEG)* (pp. 106-122). Weinheim/München: Juventa.

Appendix A



Deine Aktivitäten

Innerhalb der Schule

Schreibe hier nur Angebote auf

- die du irgendwann seit der 5. Klasse besucht hast
- die KEIN Unterricht waren
- die du mindestens ein halbes Jahr gemacht hast
- in denen du regelmäßig aktiv warst (mindestens 1 Mal pro Woche)

1. FUßBALL
2. FOTO - AG
3. COMPUTERKURS
4.
5.
6.
7.
8.
9.
10.
11.
12.
13.
14.
15.

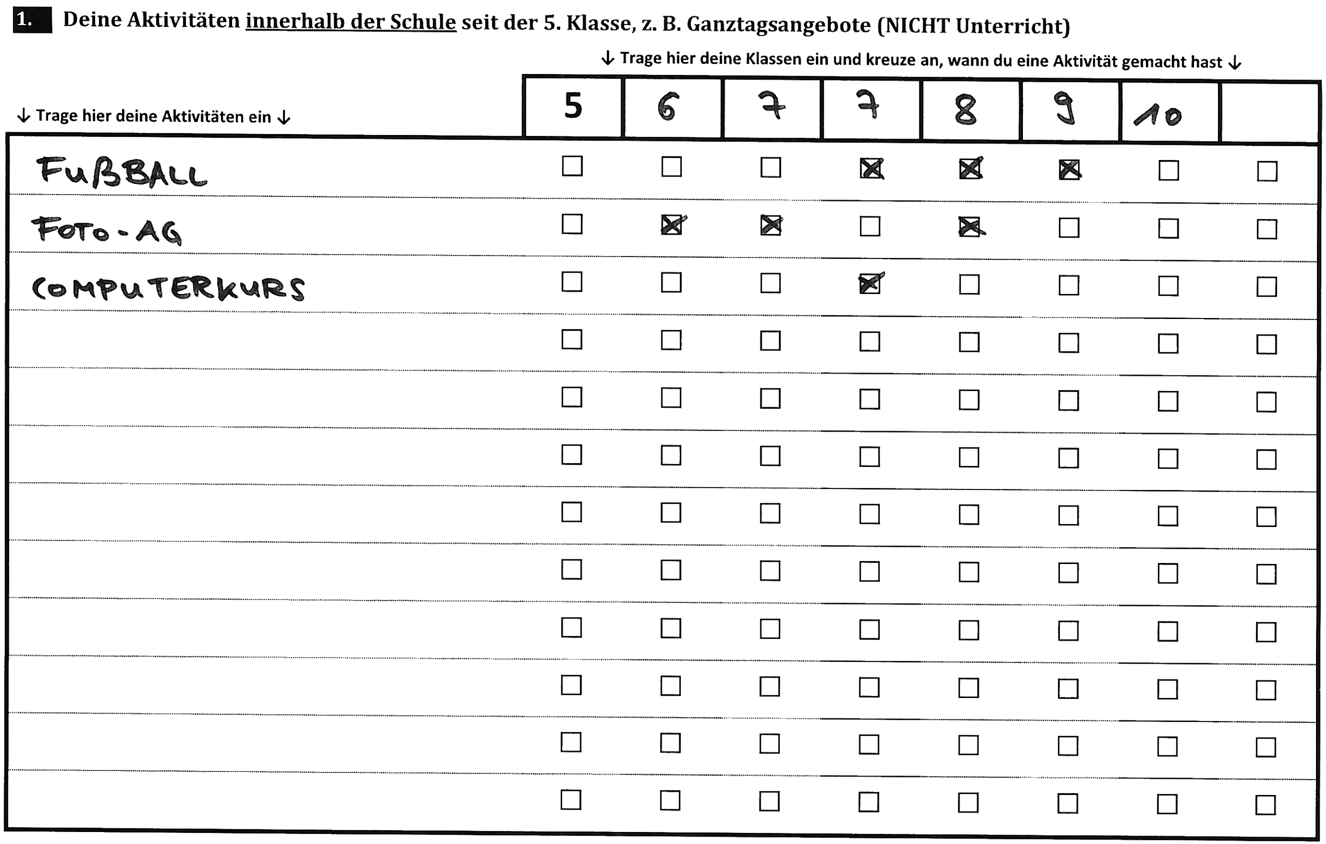
Außerhalb der Schule

Schreibe hier nur Aktivitäten auf

- die du seit der 5. Klasse gemacht hast
- die du mindestens ein halbes Jahr gemacht hast
- in denen du regelmäßig aktiv warst (mindestens 1 Mal pro Woche)

1. KIRCHENCHOR
2. TENNIS
3. KLAVIERUNTERRICHT
4. TANZEN
5.
6.
7.
8.
9.
10.
11.
12.
13.
14.
15.

Figure A-1. Example of a filled-in memorandum for in-school and non-school activities



1. **Deine Aktivitäten <u>innerhalb der Schule</u> seit der 5. Klasse, z. B. Ganztagsangebote (NICHT Unterricht)**

↓ Trage hier deine Klassen ein und kreuze an, wann du eine Aktivität gemacht hast ↓

↓ Trage hier deine Aktivitäten ein ↓	5	6	7	7	8	9	10	
FUßBALL	☐	☐	☐	☒	☒	☒	☐	☐
FOTO-AG	☐	☒	☒	☐	☒	☐	☐	☐
COMPUTERKURS	☐	☐	☐	☒	☐	☐	☐	☐
	☐	☐	☐	☐	☐	☐	☐	☐
	☐	☐	☐	☐	☐	☐	☐	☐
	☐	☐	☐	☐	☐	☐	☐	☐
	☐	☐	☐	☐	☐	☐	☐	☐
	☐	☐	☐	☐	☐	☐	☐	☐
	☐	☐	☐	☐	☐	☐	☐	☐
	☐	☐	☐	☐	☐	☐	☐	☐
	☐	☐	☐	☐	☐	☐	☐	☐
	☐	☐	☐	☐	☐	☐	☐	☐

Figure A-2. Example of a filled-in calendar of extracurricular activities in school

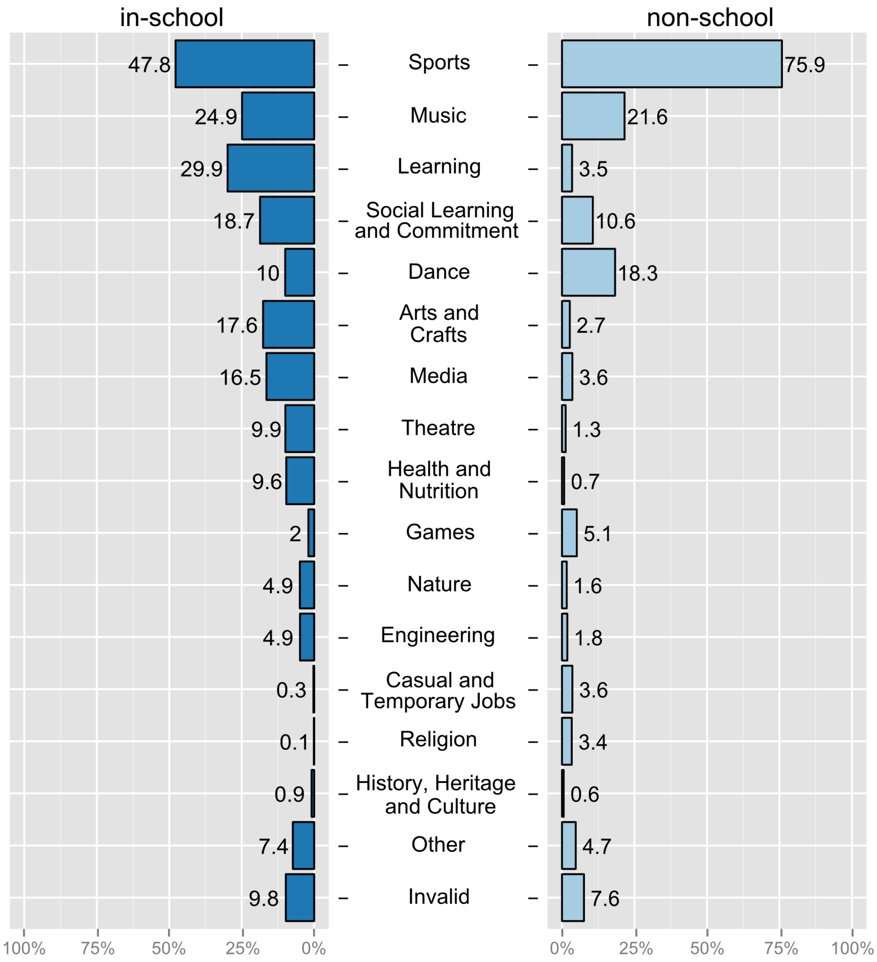


Figure A-3. Percentage of respondents with entered activities by categories and domain (n=1,901)

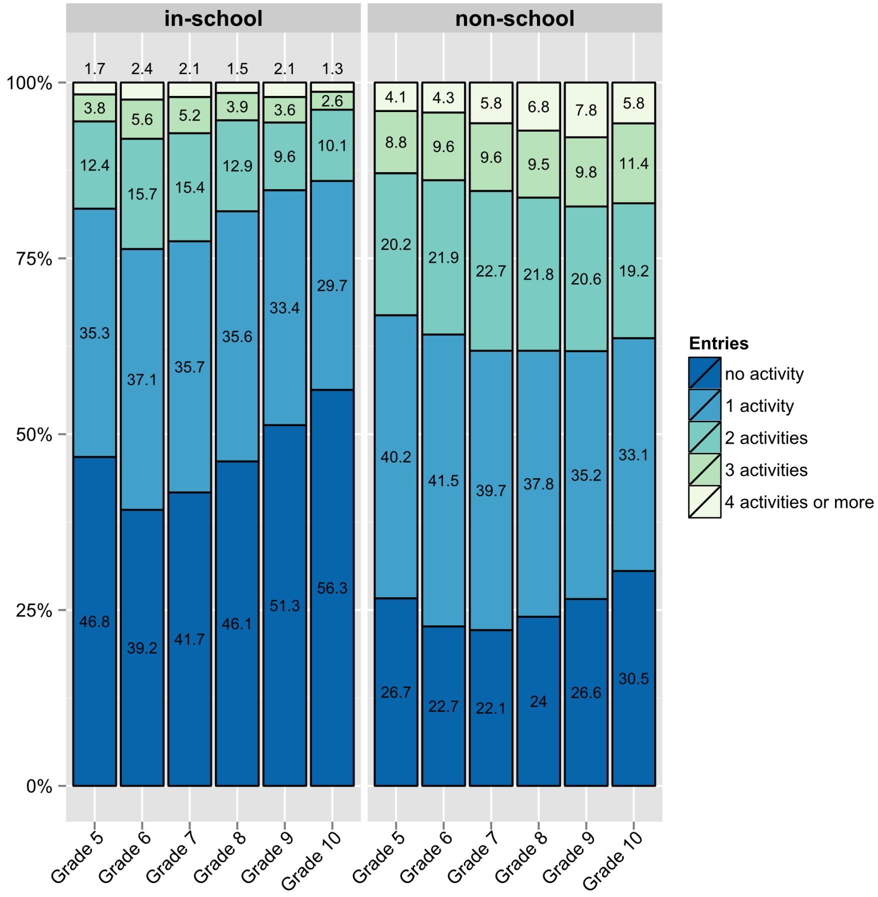


Figure A-4. Count of activities by grades (grade 5 to 9: n=1,901; grade 10: n=1,293; repeated grades are excluded)

Appendix B

Table B-1

Completion of the calendar instrument by separate steps of the guided procedure

Step of procedure	segment/ domain: in-school	segment/ domain: non-school
1. filled in memo	88.5	92.7
2. filled in grades	97.7	95.0
3. filled in activities	88.1	91.4
4. filled in dates for activities	88.0	91.3

Note: figures in %. n=1,901

Table B-2

Poisson regression and Incidence Rate Ratios (IRR) on fifth grade activity counts of the in-school calendar

		95% Confidence interval		
	β (SE[1])	CI_{min}	IRR	CI_{max}
Constant	-0.59*** (0.13)	0.43	0.55	0.71
Retrospective intervals (Reference: 4 years)				
5 years	0.01 (0.11)	0.81	1.01	1.25
6 years	-0.22 (0.14)	0.60	0.80	1.06
7 years	0.13 (0.18)	0.80	1.14	1.62
Controlled variables				
Former East Germany	0.27* (0.12)	1.02	1.31	1.67
Female	0.09 (0.07)	0.95	1.10	1.27
Intermediate school change	-0.26† (0.14)	0.59	0.77	1.01
Migration background	0.21** (0.08)	1.07	1.24	1.43
Lower HISEI[2] quartile	0.03 (0.08)	0.87	1.03	1.21
Upper HISEI[2] quartile	0.23* (0.10)	1.03	1.26	1.53
All-day-participant	0.62*** (0.10)	1.51	1.85	2.27

Note: n = 1,617. McFadden‘s Pseudo R² = .05. Wald $\chi^2(10) = 145.27^{***}$. AIC = 3,762.9.

[1] Adjusted standard errors for 65 school clusters. [2] Highest International Socio-Economic Index of Occupational Status (HISEI) of parents

† p < .10. * p < .05. ** p < .01. *** p < .001.

Table B-3

Poisson regression and Incidence Rate Ratios (IRR) on fifth grade activity counts of the non-school calendar

		95% Confidence interval		
	β (SE[1])	CI_{min}	IRR	CI_{max}
Constant	0.37*** (0.06)	1.30	1.45	1.62
Retrospective intervals (Reference: 4 years)				
5 years	0.00 (0.06)	0.88	1.00	1.13
6 years	0.10 (0.10)	0.91	1.11	1.35
7 years	0.12 (0.18)	0.79	1.12	1.60
Controlled variables				
Former East Germany	-0.19*** (0.05)	0.74	0.83	0.92
Female	0.02 (0.05)	0.93	1.02	1.12
Migration background	-0.12* (0.05)	0.80	0.89	0.98
Lower HISEI quartile[2]	-0.23*** (0.05)	0.72	0.80	0.87
Upper HISEI quartile[2]	0.03 (0.04)	0.95	1.03	1.11
All-day-participant	$0.12^{\dagger}$ (0.07)	0.98	1.13	1.30

Note: n = 1,617; Wald $\chi^2(9) = 61.18$*** ; McFadden's Pseudo R^2 =.01; AIC = 4,654.3

[1] Adjusted standard errors for 65 school clusters. [2] Highest International Socio-Economic Index of Occupational Status (HISEI) of parents

† $p < .10$. * $p < .05$. ** $p < .01$. *** $p < .001$.

Cluster Size and Aggregated Level 2 Variables in Multilevel Models. A Cautionary Note

Reinhard Schunck

GESIS – Leibniz Institute for the Social Sciences

Abstract

This paper explores the consequences of small cluster size for parameter estimation in multilevel models. In particular, the interest lies in parameter estimates (regression weights) in linear multilevel models of level 2 variables that are functions of level 1 variables, as for instance the cluster-mean of a certain property, e.g. the average income or the proportion of certain people in a neighborhood. To this end, a simulation study is used to determine the effect of varying cluster sizes and number of clusters. The results show that small cluster sizes can cause severe downward bias in estimated regression weights of aggregated level 2 variables. Bias does not decrease if the number of clusters (i.e. the level 2 units) increases.

Keywords: multilevel modeling, hierarchical linear model, sample size, survey research, cluster sampling

1 Introduction

Multilevel models (also known as hierarchical linear models and mixed models) are a common statistical tool for the analysis of clustered data (De Leeuw, Meijer, & Goldstein, 2008; Langer, 2010; Rabe-Hesketh & Skrondal, 2012; Raudenbush & Bryk, 2002; Snijders & Bosker, 2012). Their advantages are obvious: instead of treating observations incorrectly as unrelated, they explicitly take the clustering of observations into account and allow for modeling how characteristics of the higher level impact units at the lower level – for example, how neighborhood characteristics affect residents or how school characteristics affect students.

It is common in multilevel modeling to aggregate level 1 information to generate level 2 information, i.e. to characterize the clusters in which the lower level units are nested. For instance, the proportion of immigrant children in schools, the proportion of unemployed respondents in neighborhoods, the average income in neighborhoods and similar measures are frequently used in multilevel analysis (Fauth, Roth, & Brooks-Gunn, 2007; Gross & Kriwy, 2013; Pong & Hao, 2007; Schunck & Windzio, 2009; Windzio, 2004; Windzio & Teltemann, 2013).

In multilevel analysis cluster means are frequently assumed to have a meaningful interpretation, which is substantively different from the level 1 variables from which they are calculated. For instance, the mean household income in a neighborhood may be seen as a measure of neighborhood quality.[1] This paper investigates how level 1 sparseness, that is having few observations per cluster, affects the estimation of the regression weights of such aggregated level 2 variables in linear multilevel models.

Level 1 sparseness is not uncommon in empirical research. Research is often confronted with data that is of a hierarchical nature but contains only few observations per cluster. This is common in surveys that follow stratified sampling designs, where only few respondents are clustered in geographical units (Clarke & Wheaton, 2007; Schunck & Windzio, 2009).

Questions regarding adequate sample sizes at each level in multilevel analysis have been discussed before (Bell, Ferron, & Kromrey, 2008; Clarke, 2008; Clarke

[1] This sets multilevel modeling apart from longitudinal modeling in which such between-effects are often considered of having no meaningful interpretation (Allison, 2009; Schunck, 2013).

Direct correspondence to
Reinhard Schunck, GESIS – Leibniz Institute for the Social Sciences,
Unter Sachsenhausen 6-8, 50667 Köln, Germany
E-Mail: reinhard.schunck@gesis.org

Acknowledgements: I would like to thank Carsten Sauer and three anonymous reviewers for the helpful and critical comments.

& Wheaton, 2007; Hox, 1998; Kreft, 1996; Maas & Hox, 1999, 2005). Prior research suggests that level 1 sparseness does not lead to serious bias in parameter estimates (Bell et al., 2008; Clarke, 2008; Clarke & Wheaton, 2007; Maas & Hox, 2005). The number of clusters (level 2 units) seems to be more important than the number of observations per cluster. However, previous research has not systematically investigated how small sample sizes at level 1 impacts the estimates in multilevel models if these models include aggregated level 2 variables that are a function of the level 1 variables. In this case, small cluster size may cause noisy and unreliable aggregations. This becomes obvious if we consider the reliability of aggregated variables in multilevel models. For an aggregated indicator the reliability of the group mean can be expressed by

$$\lambda_j = \frac{\sigma_B^2}{\sigma_B^2 + \sigma_W^2 / n_i} \tag{1}$$

where σ_B^2 is the between group-variance of the indicator, σ_W^2 is the within-group variance, and n_i is the common cluster size (Snijders & Bosker, 2004, pp. 25-26). Reliability increases if the number of level 1 units per cluster increases and reliability decreases when the number of observations per cluster decreases.[2] In linear models, low reliability will create an error-in-variables problem and will cause an attenuation bias (Wooldridge, 2010, p. 81). This study therefore considers the effects of very small cluster sizes in linear two-level multilevel models on parameter estimates of regression weights of level 2 variables that are a function of level 1 variables.

2 Methods

To this end, this study uses Monte Carlo simulations, varying a) the cluster size, i.e. the number of level 1 units per cluster (n_i = 5, 10, 20, 40, 80) and b) the number of level 2 units (n_j = 20, 40, 100, 1000). The number and size of clusters is chosen to include the range of cluster sizes and numbers of clusters typically encountered in multilevel modeling – ranging from data with few clusters and relatively large cluster sizes to data with a large number of clusters but very few observations within clusters. Very large clusters as in country data are not considered, since the interest lies on level 1 sparseness. Data were generated based on a two-level multilevel model specified as

[2] Obviously, reliability also depends on the amount of variance between and within clusters. Reliability is also high when there are large differences between clusters.

$$y_{ij} = \alpha + \beta_1 x_{ij} + \beta_2 c_j + \beta_3 \bar{x}_j + u_j + \varepsilon_{ij} \tag{2}$$

with i indicating level 1 and j indicating level 2. x_{ij} was generated as continuous level 1 covariate from a normal distribution with a mean of 0 and a variance of 1 ($x_{ij} \sim N(0,1)$), $\bar{x}_j$ is the level 2 covariate that is a function (the cluster mean) of the level 1 covariate x_{ij}, and c_j was generated as continuous level 2 covariate from a normal distribution with a mean of 0 and a variance of 1 ($c_j \sim N(0,1)$)[3]. The level 1 error was generated from a normal distribution with a mean of 0 and a variance of 1 ($\varepsilon_{ij} \sim N(0,1)$) and the level 2 error similarly as $u_j \sim N(0,1)$. The constant was specified as $\alpha = 1$ and the regression weights as $\beta_1 = 1$, $\beta_2 = 1$, and $\beta_3 = 1$.

To simulate the data generating process more realistically, the data were generated by assuming that the cluster size (n_i) is 100 in the population. The different cluster sizes (n_i = 5, 10, 20, 40, 80) were realized by drawing random samples out of the population clusters. This corresponds for instance to drawing random samples of residents out of larger neighborhoods or students out of schools. This has important and intended consequences of the cluster mean. While the true cluster mean $\bar{x}_j$ is used to generate the data (2), the multilevel model used to analyze the data relies on the estimate $\bar{x}_j^{'}$ from the cluster samples.

For each of the 20 conditions (5 cluster sizes * 4 different numbers of level 2 units), 1,000 data sets were simulated using Stata 13.1 (StataCorp, 2013). After data generation, the simulated samples were analyzed using a linear two-level multilevel model. The examined outcomes were the estimated fixed effects, that is the regression coefficients $\hat{\beta}_1$, $\hat{\beta}_2$, and $\hat{\beta}_3$ under the specified conditions. Bias in parameter estimates is indicated by the percentage relative bias, which is assessed as ($\left(\hat{\beta} - \beta\right) / \beta$ * 100) (Maas & Hox, 1999). For instance, if the true parameter is $\beta = 1$ and the estimated parameter is $\hat{\beta}$ = 1.5, this leads to (1.5 – 1)/1 * 100 = 50, indicating the estimated parameter is upward biased by 50%. If $\hat{\beta}$ = 0.5, this leads to (0.5 – 1)/1 * 100 = -50, indicating that the estimated parameter is biased downward by 50%.

3 Results

The results of the simulation for the linear two-level multilevel model are presented in Table 1 and in Figures 1, 2, and 3.

The results show that there are very low levels of bias in the estimates of $\hat{\beta}_1$, the regression weight associated with the level 1 variable x_{ij} (Table 1). Even under

[3] Note that since a proportion is a special case of a mean, the results extend to dichotomous level 1 variables, which for instance classify observations according to a binary characteristic.

extreme conditions (n_i = 5 and n_j = 20), the estimated regression weights were very close to the true value. This is also apparent from Figure 1, which displays the mean percentage relative bias in $\hat{\beta}_1$. In all conditions, the percentage relative bias is below +/- 1%. Bias decreases on average if the cluster size or if the number of clusters increases, as can be seen from Figure 1. As regards the estimate $\hat{\beta}_2$ – the regression weight associated with the level 2 variable c_j – the results similarly show only insubstantial bias in the estimates (Table 1). Again, the percentage relative bias does not exceed +/- 1% in any condition (Figure 2). Bias decreases further when the number of level 2 units increases (Figure 2). Accordingly, for both $\hat{\beta}_1$ and $\hat{\beta}_2$ bias caused by level 1 sparseness appears negligible.

However, the results show a strikingly different picture when it comes to the estimate of $\hat{\beta}_3$, the regression weight associated with the cluster mean $\overline{x}_j^{'}$. Again, the true value for the parameter was set to equal 1. If the cluster size is very small (n_i = 5), the estimated regression weights show an extreme downward bias being close to zero (Table 1). Bias decreases when the size of the clusters increases – from an average percentage relative bias of -95.25% in the condition of extreme level 1 sparseness (n_i = 5) to -21.20% if the clusters comprise 80 level 1 observations (n_i = 80) (Figure 3). Even with moderate cluster sizes, i.e. n_i = 40, the average percentage relative bias is still -59.94. Importantly, bias does not decrease if the number of clusters increases. The number of level 2 units (n_j = 20, 40, 100, 1000) is not statistically significantly related to the size of the bias (n_i = 5: F (3, 3996) = 0.15, p<0.932; n_i = 10: F (3, 3996) = 0.65, p<0.582; n_i = 20: F (3, 3996) = 0.39, p<0.759; n_i = 40: F (3, 3996) = 0.84, p<0.474; n_i = 80: F (3, 3996) = 0.13, p<0.9446).

Table 1 Estimated regression weights (means and standard deviations)

		Estimate $\hat{\beta}_1$ (true value = 1)					Estimate $\hat{\beta}_2$ (true value = 1)					Estimate $\hat{\beta}_3$ (true value = 1)				
		n_i (cluster size)					n_i (cluster size)					n_i (cluster size)				
		5	10	20	40	80	5	10	20	40	80	5	10	20	40	80
n_j (number of clusters)	20	1.004	1.003	1.001	1.001	0.999	1.004	1.005	1.010	1.005	1.004	0.041	0.076	0.183	0.346	0.769
		(0.112)	(0.077)	(0.052)	(0.038)	(0.026)	(0.278)	(0.259)	(0.254)	(0.253)	(0.251)	(0.626)	(0.819)	(1.156)	(1.573)	(2.238)
	40	0.997	0.999	0.998	1.000	1.000	0.994	0.993	0.994	0.992	0.993	0.049	0.094	0.170	0.416	0.796
		(0.081)	(0.053)	(0.036)	(0.025)	(0.018)	(0.181)	(0.174)	(0.172)	(0.170)	(0.169)	(0.429)	(0.555)	(0.777)	(1.079)	(1.540)
	100	1.000	1.000	0.999	0.998	0.999	1.001	1.000	1.000	1.000	1.001	0.052	0.109	0.204	0.387	0.781
		(0.048)	(0.033)	(0.022)	(0.015)	(0.011)	(0.110)	(0.106)	(0.103)	(0.102)	(0.101)	(0.258)	(0.343)	(0.492)	(0.652)	(0.900)
	1,000	1.000	1.000	1.000	1.000	1.000	1.001	1.001	1.001	1.001	1.001	0.048	0.098	0.195	0.392	0.806
		(0.016)	(0.010)	(0.007)	(0.005)	(0.004)	(0.034)	(0.032)	(0.031)	(0.031)	(0.030)	(0.080)	(0.105)	(0.144)	(0.208)	(0.274)

Note: standard deviations in parentheses. Each value is averaged across 1,000 simulations.

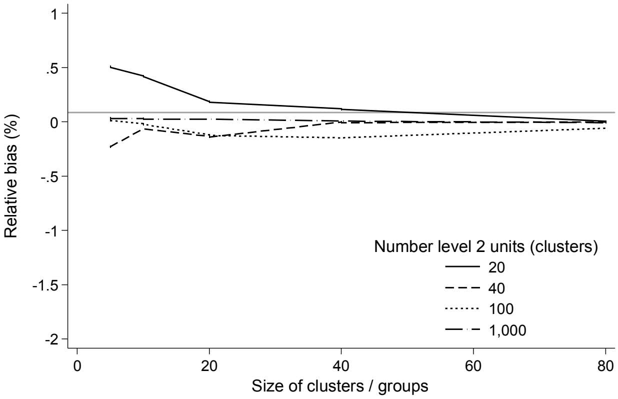


Figure 1 Percentage relative bias in $\hat{\beta}_1$

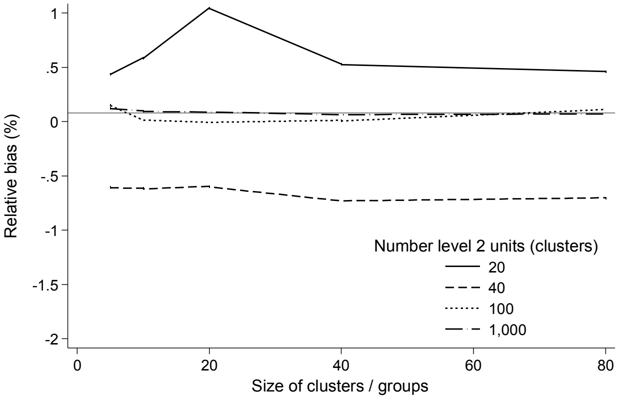


Figure 2 Percentage relative bias in $\hat{\beta}_2$

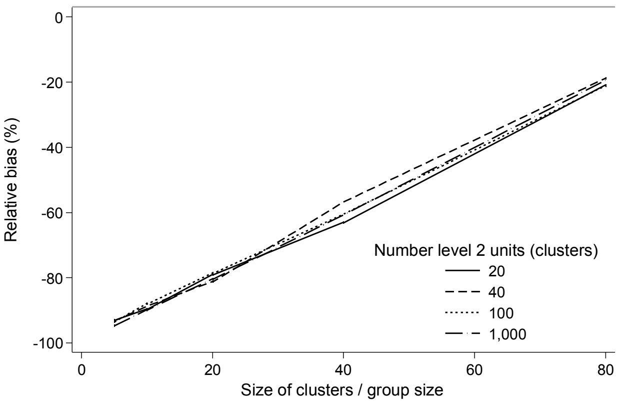


Figure 3 Percentage relative bias in $\hat{\beta}_3$

4 Conclusions

The results of this study show that level 1 sparseness (i.e. small cluster size) in multilevel models can cause large bias in estimated regression weights of level 2 variables that are aggregated from level 1 variables.

To assess the effect of level 1 sparseness, this study simulated multilevel data varying the number and the size of clusters and analyzed the data to evaluate the impact of level 1 sparseness on the estimated regression weights. The number and size of clusters had relatively little impact on the estimated effect of regression weights of normal level 1 and level 2 variables. In this respect, this study links up with previous research (Bell et al., 2008; Clarke, 2008; Clarke & Wheaton, 2007; Maas & Hox, 1999, 2005).

However, if multilevel models include level 2 variables that are a function of the level 1 variables, e.g. the average income or the proportion of unemployed people in a neighborhood, the study found severe downward bias in estimated regression weights. In situation of extreme level 1 sparseness, that is if the clusters comprise only 5 or 10 observations, the average percentage relative bias was more than 93%. Importantly, bias does not decrease if the number of level 2 units increases. Bias reduces if the number of observations within each cluster increases. However, even with moderate cluster sizes (20 or 40 observations per cluster), bias is still substantial.

What is the reason for such bias? Reliability of aggregated variables depend on cluster size (Snijders & Bosker, 2004, pp. 25-26). If very few level 1 units are used to generate the level 2 characteristic, we are dealing with measurement error: The (aggregated) level 2 characteristic is a noisy estimate of the true level 2 characteristic. It is a well-known fact that error-in-variables causes attenuation (i.e. downward) bias in estimated regression weights in linear models (Wooldridge, 2010, p. 81). The problem we are therefore facing is a measurement error or error-in-variables problem, respectively.

We have to assume that this is a prevalent problem. Most multilevel data comprise samples of level 1 units drawn out of a population of level 2 units, e.g. respondents living in larger neighborhoods, students attending different schools, or employees working in different establishments. In all these data, estimated effects of aggregated level 2 variables will be biased downward.

Obviously, the problem only applies if the clusters are samples. If the multilevel data comprises the full clusters, i.e. if all observations within a cluster are included, such as all students nested in a class, the problem will not apply – even if the clusters are small.

What can be done about this? The first and most obvious remedy is to increase the (relative) size of the clusters. The larger the number of level 1 units per cluster, the lower is the bias. A second remedy is to use external data sources to generate the aggregated level 2 characteristics. For instance, administrative data may be used to complement survey data with the level 2 variables of interest. A third remedy lies in methods that adjust for measurement error. Measurement error can, for instance, be accommodated by using a latent variable approach (Bollen, 1989; Reinecke & Pöge, 2010; Skrondal & Rabe-Hesketh, 2003). This would require using multiple level 1 indicators to model the (latent) level 2 characteristic. For instance, neighborhood characteristics could be assessed by relying on several measures, e.g. (mean) income, (mean) education, (proportion of) unemployment, etc. While these three potential remedies appear promising, one may still encounter situations in which none is applicable and should therefore treat aggregated variables in multilevel models with caution.

References

Allison, P. D. (2009). *Fixed effects regression models*. Los Angeles: SAGE.

Bell, B. A., Ferron, J. M., & Kromrey, J. D. (2008). Cluster size in multilevel models: the impact of sparse data structures on point and interval estimates in two-level models. *JSM Proceedings, Section on Survey Research Methods*, 1122-1129.

Bollen, K. A. (1989). *Structural equations with latent variables*. New York: Wiley.

Clarke, P. (2008). When can group level clustering be ignored? Multilevel models versus single-level models with sparse data. *Journal of Epidemiology and Community Health, 62*(8), 752-758.

Clarke, P., & Wheaton, B. (2007). Addressing data sparseness in contextual population research using cluster analysis to create synthetic neighborhoods. *Sociological Methods & Research, 35*(3), 311-351.

De Leeuw, J., Meijer, E., & Goldstein, H. (2008). *Handbook of multilevel analysis*: Springer.

Fauth, R. C., Roth, J. L., & Brooks-Gunn, J. (2007). Does the neighborhood context alter the link between youth's after-school time activities and developmental outcomes? A multilevel analysis. *Developmental Psychology, 43*(3), 760.

Gross, C., & Kriwy, P. (2013). Einfluss regionaler sozialer Ungleichheits- und Arbeitsmarktmerkmale auf die Gesundheit.

Hox, J. (1998). Multilevel modeling: When and why *Classification, data analysis, and data highways* (pp. 147-154): Springer.

Kreft, I. G. (1996). Are multilevel techniques necessary? An overview, including simulation studies: California State University Press, Los Angeles.

Langer, W. (2010). Mehrebenenanalyse mit Querschnittsdaten. In C. Wolf & H. Best (Eds.), *Handbuch der sozialwissenschaftlichen Datenanalyse* (pp. 741-774). Wiesbaden: Springer.

Maas, C. J., & Hox, J. J. (1999). Sample sizes for multilevel modeling. *American Journal of Public Health, 89*, 1181-1186.

Maas, C. J., & Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology, 1*(3), 86-92.

Pong, S. l., & Hao, L. (2007). Neighborhood and School Factors in the School Performance of Immigrants' Children1. *International Migration Review, 41*(1), 206-241.

Rabe-Hesketh, S., & Skrondal, A. (2012). *Multilevel and longitudinal modeling using Stata* (3rd ed.). College Station, Tex.: Stata Press Publication.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical Linear Models. Applications and Data Analysis Methods.* Thousand Oaks: Sage.

Reinecke, J., & Pöge, A. (2010). Strukturgleichungsmodelle. In H. Best & C. Wolf (Eds.), *Handbuch der sozialwissenschaftlichen Datenanalyse* (pp. 775-804). Wiesbaden: VS.

Schunck, R. (2013). Within- and Between-Estimates in Random Effects Models. Advantages and Drawbacks of Correlated Random Effects and Hybrid Models. *Stata Journal, 13*(1), 65-76.

Schunck, R., & Windzio, M. (2009). Ökonomische Selbstständigkeit von Migranten in Deutschland: Effekte der sozialen Einbettung in Nachbarschaft und Haushalt. *Zeitschrift für Soziologie, 38*(2), 111-128.

Skrondal, A., & Rabe-Hesketh, S. (2003). Some applications of generalized linear latent and mixed models in epidemiology: repeated measures, measurement error and multilevel modeling. *Norsk epidemiologi, 13*(2).

Snijders, T. A., & Bosker, R. J. (2004). *Multilevel Analysis. An Introduction to Basic and Advanced Multilevel Modeling.* London: Sage.

Snijders, T. A., & Bosker, R. J. (2012). *Multilevel analysis. An introduction to basic and advanced multilevel modeling* (2nd ed.). Los Angeles: Sage.

StataCorp. (2013). *Stata: Release 13. Statistical Software.* College Station, TX: StataCorp LP.

Windzio, M. (2004). Kann der regionale Kontext zur „Arbeitslosenfalle “werden? *KZfSS Kölner Zeitschrift für Soziologie und Sozialpsychologie, 56*(2), 257-278.

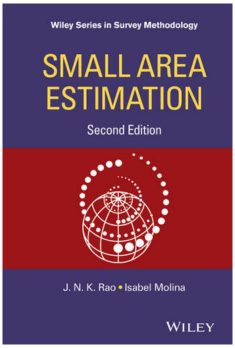
Windzio, M., & Teltemann, J. (2013). Empirische Methoden zur Analyse kontextueller Faktoren in der Bildungsforschung *Bildungskontexte* (pp. 31-60): Springer.

Wooldridge, Jeffrey M. (2010). *Econometric analysis of cross section and panel data.* (2nd edition). Bosten, MA: MIT press.

Appendix

```
// Stata code

clear all
version 13.1

global data "..."      // define file path here

//   #1
//   define program

capture program drop l2linear
program define l2linear
     clear
     drop _ all
     args i j
     set obs 'j'
     gen j = _ n
     gen c _ j = rnormal(0,1)
     gen u _ j = rnormal(0,1)
     expand 100
     bysort j: gen i = _ n
     gen x _ ij = rnormal(0,1)
     bysort j: egen x _ j = mean(x _ ij)
     gen e _ ij = rnormal(0,1)
     gen y _ ij = 1 + 1*x _ ij + 1*c _ j + 1*x _ j + u _ j + e _ ij
     bysort j: sample 'i', count
     bysort j: egen x _ j _ noise = mean(x _ ij)
     xtreg y _ ij x _ ij x _ j _ noise c _ j, i(j) re
end

//   #2
//   simulate

foreach j of numlist 20 40 100 1000 {
     foreach i of numlist 5 10 20 40 80 {

           simulate _ b, seed(12345) reps(1000): l2linear 'i' 'j'
           gen n _ j = 'j'
           gen n _ i = 'i'
           sum

           if ('j'==20 & 'i'==5) save "${data}\sim _ linear.dta", replace
           else  {
                 append using "${data}\sim _ linear.dta"
                 save "${data}\sim _ linear.dta", replace
                 }
     }
}
```

Book Review


Wiley Series in Survey Methodology
SMALL AREA ESTIMATION
Second Edition
J. N. K. Rao • Isabel Molina
WILEY

Rao, J. N. K. and Molina, I. (2015) (2nd edition).
Small Area Estimation.
Wiley Series in Survey Methodology. John Wiley & Sons, Inc., Hoboken, New Jersey.
ISBN: 978-1-118-73578-7 (cloth)
480 pages
€ 99.30 (Hardcover)

In most of the references in articles after 2003 concerning small area estimation (SAE) the first edition of the pioneering book of Rao has been cited. Google tells us that Rao's book has been cited by 1553 related articles. This shows the importance of the first edition of Rao's book which was the standard textbook on SAE until today. The demand for reliable small area estimates in various applications can be stated worldwide. As SAE methods have further evolved since the years after the first edition was published, a second edition of the book was highly anticipated by the research community. The size increased from 313 pages to 441 pages. The layout has changed, now a page contains significantly more text. This time, Rao co-authored the book with Isabel Molina, PhD, Associate Professor at the Departament of Statistics of Universidad Carlos III de Madrid, Spain, since 2009. The second edition provides new and additional developments in the field of SAE. On the back of the book some of the following innovations are mentioned:

- Additional sections describe an R package for SAE and applications with R data sets that readers can replicate
- Numerous examples of SAE applications throughout the book, including recent applications in U.S. Federal programs
- New topical coverage on extended design issues, synthetic estimation, further refinements and solutions to the Fay-Herriot area level model, basic unit level models, and spatial and time series models
- A discussion of the advantages and limitations of various SAE methods for model selection from data as well as comparisons of estimates derived from models to reliable values obtained from external sources, such as previous census or administrative data

Chapter	First Edition	Second Edition
1	Introduction	Introduction
2	Direct Domain Estimation	Direct Domain Estimation
3	Traditional Demographic Methods	Indirect Domain Estimation
4	Indirect Domain Estimation	Small Area Models
5	Small Area Models	Empirical Best Linear Unbiased Prediction (EBLUP): Theory
6	Empirical Best Linear Unbiased Prediction: Theory	Empirical Best Linear Unbiased Prediction: Basic Area Level Models
7	EBLUP: Basic Models	Basic Unit Level Models
8	EBLUP: Extensions	EBLUP: Extensions
9	Empirical Bayes (EB) Method	Empirical Bayes (EB) Method
10	Hierarchical Bayes (HB) Method	Hierarchical Bayes (HB) Method

A comparison between the first and second edition shows that the third chapter: *Traditional Demographic Methods* is no longer included in the second edition. A listing of the chapter titles is given below.

The list of figures increased from 4 to 13, the list of Tables from 20 to 23. Also, the number of examples is more numerous, e.g. subsection 1.6.6 poverty mapping. Some of the headings changed, e.g. Modified Direct Estimators to Modified GREG Estimator. New sections and old sections with significant changes are indicated by an asterisk in the book. This applies to 1 Introduction, 2.7 Optimal Sample Allocation for Planned Domains, numerous parts of 3.2 Synthetic Estimation, part of 4.4 Extensions: Area Level Models, part of 4.6 Generalized Linear Mixed Models, 5.4 Model Identification and Checking, 5.5 Software, part of 6.1 EBLUP Estimation, numerous parts of 6.2 MSE Estimation, 6.3 Robust Estimation in the Presence of Outliers, 6.4 Practical Issues, 6.5 Software, parts of 7 Basic Unit Level Model, especially 7.3 Applications, 7.4 Outlier Robust EBLUP Estimation, 7.5 M-Quantile Regression, 7.6 Practical Issues, 7.7 Software, 7.8 Proofs, most of sections in 8 EBLUP: Extensions, 9.4 EB Estimation of General Finite Population Parameters, 9.7 Design-Weighted EB Estimation: Exponential Family Models, 9.11 Software, parts of 10.3 Basic Area Model, 10.4 Unmatched Sampling Variances ${}_i$, 10.7 HB Estimation of General Finite Population Parameters, 10.12 Two-Part Nested Error Model, 10.14 Missing Binary Data and 10.17 Approximate HB Inference and Data Cloning.

331 references were cited in the first edition on 20 pages, about 500 on 26 pages in the second edition. Amongst the references in the second edition are 23 from the

year 2015, 22 from the year 2014 and more than 140 from years 2003-2013. Also older references which were not in the first edition have been added.

The increase of items in the Author Index and Subject Index is not surprising.

Just like the first edition, the second edition is also intended primarily as a research monograph, but is also suited as a fundamental textbook for graduate-level courses in SAE and reliable small area statistics, as is cited in the preface of the book.

Summarized, the second edition of *Small Area Estimation* is a must read for all survey methodologists as well as for practitioners interested in SAE methods. Because of the immense growth in research and applications to SAE methods, it can be expected that a third edition or a new book maybe asked for in near future.

Siegfried Gabler

DOI: 10.12758/mda.2016.006

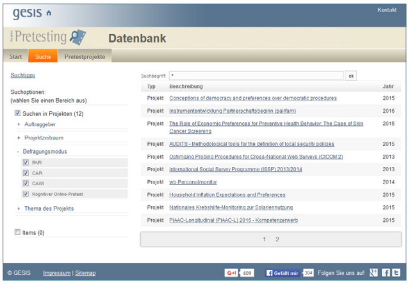
gesis Leibniz Institute for the Social Sciences

GESIS Pretest Database – pretest.gesis.org

The GESIS pretest database provides researchers access to findings of cognitive pretests conducted by the GESIS pretest lab. You can find, for example, the following information on a tested question:

- How do respondents understand the question or specific terms?
- Is the question understood in the way intended by the researcher?
- How easy or difficult is it to answer the question?

You can either perform a keyword search and filter the results by question topic, survey mode, multiitem scale (yes/no) and several other search options or browse the documented pretest projects.



Suchtipps

Suchoptionen:
(wählen Sie einen Bereich aus)

☑ Suchen in Projekten (12)
+ Auftraggeber
+ Projektzeitraum
- Befragungsmodus
☑ PAPI
☑ CAPI
☑ CAWI
☑ Kognitiver Online-Pretest
+ Thema des Projekts

☐ Items (0)

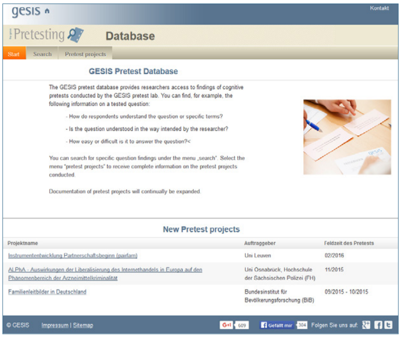
Suchbegriff: * ok

Typ	Beschreibung	Jahr
Projekt	Conceptions of democracy and preferences over democratic procedures	2015
Projekt	Instrumententwicklung Partnerschaftsbeginn (pairfam)	2016
Projekt	The Role of Economic Preferences for Preventive Health Behavior: The Case of Skin Cancer Screening	2015
Projekt	AUDITS - Methodological tools for the definition of local security policies	2015
Projekt	Optimizing Probing Procedures for Cross-National Web Surveys (CICOM 2)	2013
Projekt	International Social Survey Programme (ISSP) 2013/2014	2013
Projekt	wb-Personalmonitor	2014
Projekt	Household Inflation Expectations and Preferences	2015
Projekt	Nationales Krebshilfe-Monitoring zur Solarnutzung	2015
Projekt	PIAAC-Longitudinal (PIAAC-L) 2016 - Kompetenzerwerb	2015

1 2

For more informationen about our services and the costs and duration of questionnaire pretests please visit:

http://www.gesis.org/en/services/study-planning/pretest-lab/



GESIS Pretest Database

The GESIS pretest database provides researchers access to findings of cognitive pretests conducted by the GESIS pretest lab. You can find, for example, the following information on a tested question:

- How do respondents understand the question or specific terms?
- Is the question understood in the way intended by the researcher?
- How easy or difficult is it to answer the question?<

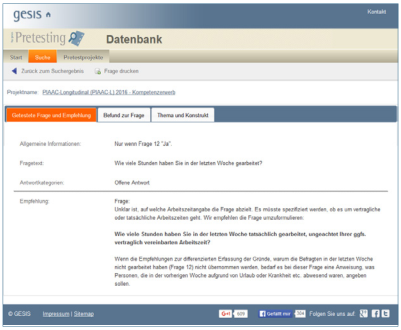
You can search for specific question findings under the menu „search". Select the menu "pretest projects" to receive complete information on the pretest projects conducted.

Documentation of pretest projects will continually be expanded.

New Pretest projects

Projektname	Auftraggeber	Feldzeit des Pretests
Instrumententwicklung Partnerschaftsbeginn (pairfam)	Uni Leuven	02/2016
ALPhA - Auswirkungen der Liberalisierung des Internethandels in Europa auf den Phänomenbereich der Arzneimittelkriminalität	Uni Osnabrück, Hochschule der Sächsischen Polizei (FH)	11/2015
Familienleitbilder in Deutschland	Bundesinstitut für Bevölkerungsforschung (BiB)	09/2015 - 10/2015

For every question tested, you can find the detailed test results and, if applicable, recommendations for improvement.



Projektname: PIAAC-Longitudinal (PIAAC-L) 2016 - Kompetenzerwerb

Allgemeine Informationen: Nur wenn Frage 12 "Ja".

Fragetext: Wie viele Stunden haben Sie in der letzten Woche gearbeitet?

Antwortkategorien: Offene Antwort

Empfehlung: Frage:
Unklar ist, auf welche Arbeitszeitangabe die Frage abzielt. Es müsste spezifiziert werden, ob es um vertragliche oder tatsächliche Arbeitszeiten geht. Wir empfehlen die Frage umzuformulieren:

Wie viele Stunden haben Sie in der letzten Woche tatsächlich gearbeitet, ungeachtet Ihrer ggfs. vertraglich vereinbarten Arbeitszeit?

Wenn die Empfehlungen zur differenzierten Erfassung der Gründe, warum die Befragten in der letzten Woche nicht gearbeitet haben (Frage 12) nicht übernommen werden, bedarf es bei dieser Frage eine Anweisung, was Personen, die in der vorherigen Woche aufgrund von Urlaub oder Krankheit etc. abwesend waren, angeben sollen.

Kontakt: (pretest.gesis.org)

Member of the Leibniz Association

Information for Authors

Methods, data, analyses (mda) publishes research on all questions important to quantitative methods, with a special emphasis on survey methodology. In spite of this focus we welcome contributions on other methodological aspects.

Manuscripts that have already been published elsewhere or are simultaneously submitted to other journals will not be considered. As a rule we do not restrict authors' rights. All rights remain with the author, and articles in mda are published under the CC-BY open-access license.

Mda aims for a quick peer-review process. All papers submitted to mda will first be screened by the editors for general suitability and then double-blindly reviewed by at least two reviewers. The decision on publication is made by the editors based on the reviews. The editorial team will contact the authors by email with the result at the latest eight weeks after submission; if the reviews have not been received by then, we provide a status update with a new target date.

When preparing a paper for submission, please consider the following guidelines:

- Please submit your manuscript by e-mail to mda(at)GESIS(dot)org.
- The total length of the manuscript shall not exceed 10.000 words.
- Manuscripts should…
  - be written in English, using American English spelling. Please use correct grammar and punctuation. Non-native English speakers should consider a professional language editing prior to publication.
  - be typed in a 12 pt Roman font, double-spaced throughout.
  - start with a cover page containing the title of the paper and contact details / affiliations of the authors, but be anonymized for review otherwise.
- Please also send us an abstract of your paper (approx. 300 words), a brief biographical note (no longer than 250 words), and a list of 5-7 keywords for your paper.
- Acceptable formats for Graphics are
  - Tiff
  - Jpeg (uncompressed, high quality)
  - pdf
- Please ensure a resolution of at least 300 dpi and take care to send hiqh-quality graphics. Line art images should have a resolution of 500-1000 dpi. Please note that we cannot print color images.
- The type area of our journal is 11.5 cm (width) x 18.5 cm (height). Please consider this when producing tables or graphics.
- Footnotes should be used sparingly.
- By submitting a paper to mda the authors agree to make data and program routines available for purposes of replication.

Please follow the APA guidelines when preparing in-text references and the list of references.

Entire Book:

Groves, R. M., & Couper, M. P. (1998). *Nonresponse in household interview surveys*. New York: John Wiley & Sons.

Journal Article (with DOI):

Klimoski, R., & Palmer, S. (1993). The ADA and the hiring process in organizations. *Consulting Psychology Journal: Practice and Research*, 45(2), 10-36. doi:10.1037/1061-4087.45.2.10

Journal Article (without DOI):

Abraham, K. G., Helms, S., & Presser, S. (2009). How social processes distort measurement: The impact of survey nonresponse on estimates of volunteer work in the United States. *American Journal of Sociology*, 114(4), 1129-1165.

Chapter in an Edited Book:

Dixon, J., & Tucker, C. (2010). Survey nonresponse. In P. V. Marsden & J. D. Wright (Eds.), *Handbook of Survey Research*. Second Edition (pp. 593-630). Bingley: Emerald.

Internet Source (without DOI):

Lewis, O., & Redish, L. (2011). *Native American tribes of Wisconsin*. Retrieved April 19, 2012, from the Native Languages of the Americas website: www.native-languages.org/wisconsin.htm

For more information, please consult the Publication Manual of the American Psychological Association (Sixth ed.).

gesis
Leibniz Institute for the Social Sciences

ISSN 1864-6956 (Print)
ISSN 2190-4936 (Online)