Health estimate differences between six independent web surveys: different web surveys, different results?

Rainer Schnell[1*] and Jonas Klingwort[2]

Abstract

Most general population web surveys are based on online panels maintained by commercial survey agencies. Many of these panels are based on non-probability samples. However, survey agencies differ in their panel selection and management strategies. Little is known if these different strategies cause differences in survey estimates. This paper presents the results of a systematic study designed to analyze the differences in web survey results between agencies. Six different survey agencies were commissioned with the same web survey using an identical standardized questionnaire covering factual health items. Five surveys were fielded at the same time. A calibration approach was used to control the effect of demographics on the outcome. Overall, the results show differences between probability and non-probability surveys in health estimates, which were reduced but not eliminated by weighting. Furthermore, the differences between non-probability surveys before and after weighting are larger than expected between random samples from the same population.

Keywords Nonresponse, Undercoverage, Calibration, Non-probability sample, Probability sampling, Weighting techniques, Survey agencies, Recruitment strategy, Access panel

Introduction

Web surveys are increasingly used for public health research [1–3], official statistics [4], social and marketing research [5], and election and opinion polls [6, 7]. Due to the low costs, short fieldwork periods, the ease of elaborate filtering, and the option to use multimedia elements in questionnaires [8], the popularity of web surveys is plausible. Furthermore, the absence of interviewers often seems to be an argument for less socially desirable response behaviour [9]. Finally, during the Covid-19 pandemic, web surveys had the additional advantage of preventing any physical contacts.

However, the main methodological problem of web surveys is sampling. For general population surveys, with very few exceptions, no sampling frame for direct contacts to respondents (e.g., email addresses) is available. Population covering sampling frames containing email addresses are commonly restricted to special populations, such as specific professions. Therefore, most commercial web surveys are based on non-probability samples, usually recruited online. In general, online recruitment of new panel members uses, for example, social networks, online communities/social media, affiliate networks, or website banners [10].

Since most web surveys are based on non-probability samples, differences between differently recruited surveys can be expected [11]. Because probability samples are based on easily reproducible procedures with error bounds given by sampling error, differences between probability samples should be small. In contrast, the

*Correspondence:
Rainer Schnell
rainer.schnell@uni-due.de
[1] University of Duisburg-Essen, Duisburg, Germany
[2] Centraal Bureau voor de Statistiek, Heerlen, The Netherlands

procedures for non-probability samples are hardly described in detail, making reproduction difficult. Therefore, we expect larger differences between non-probability samples and probability samples, as could be expected between different random samples from the same population.

There is a lack of empirical studies evaluating whether different agencies will produce similar results for the same questionnaire on factual items using web surveys. A recent study by [12], for example, compares demographics and voting outcomes, but no other factual items, between two random samples samples and three non-probability samples.

This article presents the results of the first systematic empirical study of potential differences between health-related web surveys. Therefore, six independent web surveys were commissioned. Four agencies used non-probability samples, two agencies probability-based samples.

Research background

Undercoverage and nonresponse in web surveys

Currently, due to the absence of a sampling frame of the general population, random sampling for single-mode web surveys is impossible [13] under almost all jurisdictions. Although other sampling procedures are possible, in practice, web surveys are often based on online-recruited access panels, or participants are recruited on websites visited. Therefore, elements are missing in the sample due to coverage errors, and inclusion probabilities are unknown for those responding. Accordingly, design-based estimates of web surveys are not legitimated by mathematical statistics [14]. Some survey agencies recruit offline by drawing random samples, for example from a population register and inviting sampled persons to participate in a web survey to remedy this problem. Furthermore, if persons without internet access are provided with online access, recruiting offline may reduce undercoverage bias [15–17].

Although the level of internet access at the European Union's household level has increased steadily, differences still exist: from 85% in Greece to 99% in Norway (in 2022) [18]. For Germany in 2022, official statistics reported internet access at the household level at 91% and individual internet usage at about 93% [18, 19]. Based on the American Community Survey 2021, the internet penetration rate in the USA is estimated at 90% [20]. Depending on the excluded proportion of people without internet access and the difference in the target variable between people with internet and without internet [21], excluding subpopulations might bias survey estimates.

While undercoverage addresses the internet access requirement, nonresponse refers to the ability and motivation to participate in the survey. Both selection mechanisms, undercoverage and nonresponse, may cause bias in the survey estimates. Accordingly, a survey data set is the result of both selection mechanisms. Disentangling between coverage and nonresponse errors is impossible due to the absence of a sampling frame [22]. The larger the proportion of non-respondents and the stronger the correlation between the target variable and the missing data mechanism, the larger the nonresponse bias.

An equation by [21] allows the estimation of the difference between a population mean $\bar{Y}$ and a mean of a sample with nonreponse $\bar{Y}_{nr}$:

$$\bar{Y} - \bar{Y}_{nr} \approx \frac{R_{\rho Y} S_{\rho} S_Y}{\bar{\rho}}. \tag{1}$$

The equation (1) assumes that every person has a response propensity ρ, with an overall mean $\bar{\rho}$, and standard deviation S_{ρ} in the population. $R_{\rho Y}$ is the correlation between Y and ρ, and S_Y is the standard deviation of Y. The bias ($\bar{Y} - \bar{Y}_{nr}$) depends on three quantities: the correlation between the response propensity and the variable to be estimated, the variance of the response propensity, and the variance of the variable of interest. Accordingly, the bias will be small if either the participation rate in the non-probability sample is high or at least one of the other factors ($R_{\rho Y}$, $S\rho$, S_Y) is small.

Regardless of response mode or mandatory participation, surveys show a downward trend in response rates [23–26]. Web surveys yield even lower response rates than other response modes [27]. In general, with increasing proportions of nonresponse and increasing correlations between the response variable and the mechanisms causing nonresponse, the risk for biased estimates increases [28]. However, although probability-based surveys suffer from decreasing response rates, empirically, their estimates seem to be still more accurate than estimates obtained by non-probability samples [11]. Given equation (1), this empirical result is mathematically plausible only if the correlation between response propensity and variables of interest is low or the differences in response propensities are small in probability-based surveys.

Internet use and health

The mechanisms causing differences in internet use depending on health conditions are rarely discussed. However, the selection process from the target population to the population covered by web surveys can be characterized by six requirements. First, the technical requirements of a working internet connection by line, WiFi, digital cellular networks or satellites must be fulfilled. Second, sufficient financial resources by the

respondents are necessary if internet access is not provided for free. Third, using a smartphone or a computer requires the physical ability to see (or hear) and the ability to type or speak. Fourth, answering survey questions requires cognitive abilities such as understanding abstract concepts, word finding and judgment. Fifth, recruitment for a survey needs a mode to contact the respondent, which usually requires a sampling frame. Such a frame for web-based population surveys is rare. Therefore, offline sampling requires other frames, such as phone numbers or address lists, which can only be used indirectly for web surveys. For online sampling, river sampling or similar non-probability sampling techniques are necessary. Sixth, the designated respondent needs sufficient motivation to answer a survey request.

Physical or cognitive capabilities might be impaired depending on the symptoms of a medical condition. Due to hospital stays or caregivers, the probability of contact with the designated respondent may vary between contact modes and medical conditions. Finally, motivation for a response might decrease (or increase) depending on the type and severity of the medical condition.

The effects might not necessarily be linear or additive. For example, physical disabilities might impact survey participation only for severely disabled persons. Increasing physical or cognitive problems might reduce motivation as well. Therefore, neither a diagnosis (a reported code from the International Classification of Diseases, ICD) nor a specific symptom alone will be a sufficient or necessary condition for survey response. Hence, no simple pattern of health conditions and survey participation can be expected.

However, some studies are available if the potential bias could be reduced by weighting. [29] reported that weighting by age and gender did not eliminate differences between a web and a CAPI survey in BMI, eating habits, physical activity, alcohol consumption, and smoking. [2] reported that web responders currently smoked less, had fewer children, and less often had a chronic disease. Observed health differences between internet users and non-users (based on the Michigan Behavioral Risk Factor Surveillance System (BRFSS) survey 2003 and the Health and Retirement Study 2002) are described by [30]. Based on reported internet usage in European (European Social Survey) and US data (BRFSS) obtained by conventional surveys (F2F and CATI), [31] note that '(...) calibration on age, gender, ethnic background, urban residence, education and household income does not eliminate the observed health differences'. In a probability-based survey with the web as a response mode, [32] showed significant differences between web and face-to-face respondents after controlling for gender, age, region, marital status, household size, educational attainment, and country of

birth. Recently, [33] showed in a comprehensive analysis of American data persisting differences between internet users and non-users including age, employment, cultural activities, and education. Using British and Swedish data, [34] provided evidence for similar differences regarding, for example, age, low-level of education, and living alone. Both publications noticed health issues (such as disability) as predictors for internet use. [33] summarized their findings: 'Without some reasonable adjustment, a variable like health status has a high risk of being significantly biased in studies that do not cover the non-internet population'. Hence, there is growing evidence of correlations between health indicators and internet use, which cannot be corrected by weighting procedures.

Selection mechanisms and weighting

The data missing due to coverage and nonresponse errors can be described by three different data generating mechanisms: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR) [35, 36].

If the data generating mechanism is MAR, the generalized regression estimator (GREG) can be used to correct the effect of such missing data by calibration [37]. The calibration estimator for respondents (*r*) of a target variable *Y* is defined as

$$\hat{Y}_w = \sum_r w_i y_i, \tag{2}$$

with w_i the calibrated correction weights and y_i the response. The w_i are a product of the initial weights wi_i and the correction factor v_i, where $w_i = wi_i * v_i$ (for details on the calibration estimator, see [38]). When the MAR assumption does not hold, GREG estimates are still biased. In such cases, the probability of being included in the survey and response depends on y_i, and (unobserved) x_i cannot fully explain the selection probability. Therefore, the missing data generating mechanism is considered as MNAR.

Survey agencies and recruitment strategies

The research design of the study presented here intended to commission five different survey agencies to collect health data using their online panels. The agencies included are the largest commercial market research companies offering their panels for academic research in Germany. However, since none of these panels were based on probability samples, the only commercial probability-based web panel was also included. Due to an additional grant, a sixth web survey could be conducted five months after the last interview of the other five surveys. This panel is the only general purpose academic probability-based web survey in the country under study,

Table 1 Agency, sampling type, recruitment, fieldwork duration, and sample size of the six surveys

Agency	Sampling	Recruitment	Fieldwork period	Sample size
NPS-1	Quota sample	online	13.09. – 30.09.2016	5.002
NPS-2	Quota sample	online	13.09. – 28.09.2016	5.000
NPS-3	Quota sample	online	13.09. – 16.11.2016	5.501
NPS-4	Quota sample	online	13.09. – 19.09.2016	2.877
PS-1	Random sample	offline	13.09. – 07.10.2016	5.001
PS-2	Random sample	offline	19.04. – 13.06.2017	3.065

which conducts research on behalf of external research groups. The target population was defined as the residential population aged 18 years and older. Details of the fieldwork are shown in Table 1.

NPS-1, NPS-2, and NPS-3 are globally operating providers of online panels.[1] These agencies use opt-in access panels where participants have deliberately registered. The commissioned agencies also used quota sampling to approximate the population demographics with their samples. NPS-4 is an ongoing online opt-in panel managed by a university. This panel invited all members of the panel to obtain a sample for this study.

PS-1 and PS-2 used probability samples. PS-1 is a commercial research company, PS-2 a publicly financed research agency. While survey agency PS-1 used a random-digit dialing probability sample [40], PS-2 used F2F interviews of a register-based sample to recruit members for panel-participation [17]. At the time of this study, PS-1 provided internet access to previously offline respondents.

All agencies were contractually obliged to implement the same survey, with a questionnaire developed and tested in advance. However, all details of the fieldwork were left to the agencies to reflect their actual practice.[2]

To control for time-dependent influences, the initially planned five surveys started fieldwork at the same time. The sample size to be delivered was contractually set at

$n = 5.000$. Agencies with smaller panels (NPS-4 and PS-2) delivered smaller sample sizes. No further data pre-processing was applied. Hence the data delivered by the agencies were analyzed. To sum up: NPS-1, NPS-2, NPS-3, NPS-4 used undefined mixtures of non probability samples. PS-1 and PS-2 used probability samples.

Questionnaire and pretest

Health was chosen as the survey topic because undercoverage and nonresponse due to health issues seem likely given previous research (see the Internet use and health section). The questionnaire contained non-attitudinal questions on general health status, health-care utilization, injuries and accidents, disabilities and chronic diseases, and health-related behaviour. In total, 36 health items were used. The majority of items were taken from three general population health monitoring studies (DEGS-1 [41, 42], GEDA-12 [43], and GEDA-14/15 [44]).

Six items on demographics were asked to be used in the weighting model. The questions on age and municipality size were taken from a reference survey [45]. The questions on education were adopted from official statistics [46]. The questions on gender and federal state were developed in collaboration with the survey agency that conducted the pretest.

The questionnaire was pretested using an online recruited access-panel of a market research agency. In total, 550 respondents of a quota sample tested the questionnaire. Only seven aborted interviews, not clustered at specific questions, were identified. No question produced implausible responses or large proportions of item-non-response. The questionnaire required, on average, about 5 minutes to complete.

Dependent variables

In total, 36 health items will be used as dependent variables in the analysis. A dichotomized five-point ordered subjective health indicator (good/very good: 1; else: 0), and the Body-Mass-index (calculated from questions on weight and height) were used for general health status. Six questions asked for the use of health services:

1. the number of general practitioner visits in the last four weeks,
2. the number of general practitioner visits in the last 12 months,
3. the number of calendar days being ill and unable to perform usual duties in the last 12 months,
4. the number of working days being diagnosed by a physician as unable to work within the last 12 months,
5. the number of overnight hospital stays for inpatient treatment within the last 12 months and

---

[1] NPS-1 is operated by GMI (now part of Kantar), NPS-2 is operated by SSI (now named Dynata), NPS-3 is operated by Ipsos, NPS-4 is the WiSo-panel [39], PS-1 is operated by Forsa, and PS-2 is the GESIS panel.

[2] Therefore, many different panel management strategies could impact differences between agencies, for example, recruitment, payment, control and web interface. Furthermore, providers may have different panel attrition problems or suffer from different panel conditioning effects. Separating these effects could form a research program on its own.

6 if an artificial hip joint replacement was implemented in the last 12 months.

For the latter question as for all dichotomous questions, the code 0 indicates 'no', code 1 indicates 'yes'.

Four dichotomous indicators were asked on traffic accidents, accidents at home, accidents in leisure time, and injuries at work in the last 12 months. 15 items related to ever diagnosed diseases (high blood pressure, allergy, chronic back pain, sleep disorder, joint diseases (arthrosis/rheumatism), depressive disorder, migraine, heart diseases (heart failure/cardiac insufficiency), chronic bronchitis, diabetes, osteoporosis, liver diseases (fatty liver/liver infection/hepatitis/liver shrinkage/cirrhosis), asthma, stroke, and cancer.

Further, the respondents were asked to indicate their use of glasses/contact lenses, their ability to read a newspaper without difficulty and their use of hearings aids. Finally, they should report if a disability condition was officially certified, and if so, which degree of disability (20-100) was certified.

Finally, four items covered health-related behaviour. Smoking was assessed by asking for the number of smoked cigarettes per day and week. Sports activity was dichotomized (less than one hour/at least one hour per week). Drinking alcoholic beverages was also dichotomized (less than four days or at least four days per week). However, due to a restrictive interpretation of German data protection law by PS-2, only 14 out of 36 variables are provided for the PS-2 data set.[3]

Methods
Methods for comparisons
Data was analyzed in four steps: Beginning with unweighted survey estimates, ANOVA was used for testing mean differences in unweighted estimates between surveys. In the second step, Generalized Regression Estimation (GREG) was used for weighting the surveys. In the third step, multiple pairwise comparisons between the weighted means of surveys using Tukey's Honest Significant Difference Test (Tukey's HSD) were computed. Finally, we pooled the non-probability samples into a single NPS group to reduce the number of comparisons and use t-tests for comparison.[4]

Since the comparisons for 14 variables are based on the comparisons between all six web surveys, we have $(6 \times 5)/2 = 15$ pairwise comparisons, in total $15 \times 14 = 210$ comparisons for six surveys. In addition, we have 22 variables for five surveys, giving $(5 \times 4)/2 = 10$ pairwise comparisons of 22 variables giving $22 \times 10 = 220$. Overall, we have $210 + 220 = 430$ group comparisons.

To simplify the results by reducing the number of comparisons we pooled the non-probability samples into a single NPS group in a separate analysis. For these comparisons between PS and NPS samples, t-tests were used. For PS-1 versus NPS, 36 items are available for comparison and for PS-2 versus NPS, 14 items. We first compare unweighted estimates and then weighted estimates. Therefore, in total, there are $k = 2 \times 36 + 2 \times 14 = 100$ comparisons.[5] To account for potential problems due to multiple testing, we apply a Bonferroni correction ($\alpha_{adj} = \frac{\alpha}{k}$), which is widely regarded as conservative [49].

Weighting
Weighting adjustments for web surveys are commonly based on demographics, such as age and gender [50, 51]. The weighting model in the study reported here is based on demographics as well. The surveys are weighted by region, size of the municipality, age, gender, and education. Population totals of the Census were used for calibration.[6] However, due to item-nonresponse, the information required for weighting was unavailable for some respondents.[7] These observations were removed from the analysis.[8]

Figure 1 shows the distribution of the computed GREG weights for each survey. The upper row shows the initial weights. Especially the large weights indicate selection

---

[3] The data would have been available in a closely supervised research data center, but initially, PS-2 was not able to grant access within six months to the research data center. Later, Covid-19 restrictions delayed access to the research data center.

[4] Since two heterogeneous kinds of samples have to be compared, we have no meta-analysis problem, which excludes standard measures of heterogeneity. Therefore, we use multiple pairwise comparisons (Tukey's HSD) between the weighted means of surveys.

[5] Comparing *p*-values with a fixed threshold is rarely advised [47]. We use t-tests here as rough indicators for differences larger than expected, not to make decisions about a hypothesis. However, the effect measure Cohen *d* is related to *t*: ($|t| = \sqrt{(n_1 n_2)/(n_1 + n_2)}d$) [48]. The factor for multiplying *d* to yield *t* is about 38.7 and 50 for all comparisons. Due to this monotonic transformation, an analysis based on *d* would, therefore, yield comparable results. To help interpreting the results, we additionally report effect sizes using Cohen's *d*.

[6] Age was used with six categories (18–24, 25–29, 30–39, 40–49, 50–64, and 65+), gender with two categories, education with five categories, size of the municipality with three categories (10.000–20.000, 20.001–100.000, 100.000 and more inhabitants) and region with 16 categories (the German federal states). The GREG weighting model can be written as age ∗ gender ∗ education ∗ size of municipality ∗ federal state.

[7] Between 0.4% (NPS-2) and 8.3% (PS-1) respondents did not answer at least one question on demography.

[8] During the weight computations, empty cells in the weighting model were replaced with one pseudo-observation for each missing cell. The number of created pseudo-observations per survey were 1.566 for NPS-1, 1.628 for NPS-2, 1.505 for NPS-3, 1.965 for NPS-4, 1.714 for PS-1, and 1.839 for PS-2. After calculating the weights, the pseudo-observations were removed from the data set.
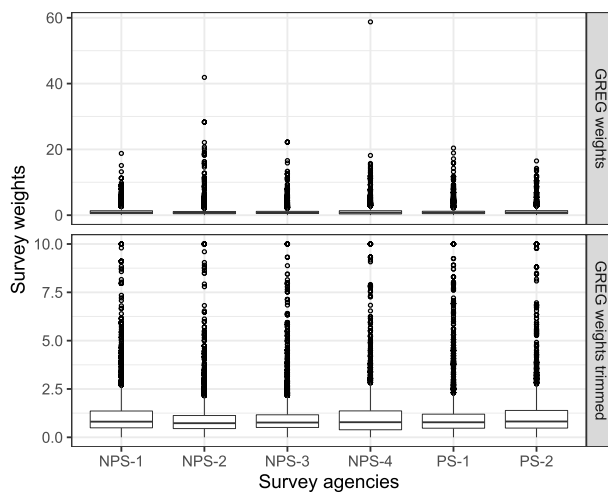
Fig. 1 Boxplots of weight distributions for each survey. The upper row shows initial survey weights, the lower row shows trimmed survey weights

problems. The use of such weights would substantially increase the sampling error. In accordance to survey practice, the weights were trimmed [52, 53]. Trimming weights will introduce a bias in the weighted estimates [54]. Therefore a moderate amount of trimming was used by setting the maximum value of weights to 10. The lower row shows the distribution of the trimmed weights used for the following analysis below.

Results
Differences between unweighted estimates
Figure 2 shows the unweighted survey estimates and 95%-confidence intervals for each survey. Overall, the unweighted differences between the web surveys are larger than expected by sampling alone. Regarding general health status variables, PS-1 has the healthiest respondents in terms of health self-assessment, and the mean BMI is also one of the lowest but still indicates a pre-obese state. The mean of respondents of NPS-3 and NPS-4 corresponds to obesity grade I.

For most variables, estimates of PS-1 are among the lowest for the use of health services. Respondents of NPS-1, NPS-2, and NPS-4 show the largest numbers of consultations.

Concerning the 15 items on ever-diagnosed diseases, the survey PS-1 shows the highest proportions in five variables (high blood pressure, allergy, joint diseases, heart disease, and cancer). These are mostly diseases with high incidence in the general population. Furthermore, larger proportions of PS-1 and PS-2 respondents wear glasses/contact lenses and read the print of a newspaper without problems. PS-1 has the largest proportion of respondents wearing hearing aids.

Furthermore, PS-1 and PS-2 show the lowest proportions of respondents with an officially certified disability. In addition, respondents of PS-1 and PS-2 are more active in sports and smoke less per day. However, on a weekly basis, respondents of PS-2 smoke the most. Finally, PS-1 and PS-2 show the largest proportions of respondents drinking alcoholic beverages ≥ 4 times per week.

Separate ANOVAs for each variable showed significant group differences between survey agencies for 34 out of 36 reported variables (94%). Non-significant ($p > 0.05$) differences are found for 'the number of overnight hospital stays for inpatient treatment within the last 12 months' and 'ever diagnosed with osteoporosis'.

Table 2 shows the count of Bonferroni adjusted p-values of t-tests according to conventional significance levels. For the unweighted estimates, 19 variables (56%) were found to differ significantly between PS-1 and NPS. For PS-2 versus NPS, 8 variables (57%) showed significant differences.

The mean effect sizes d [55] in the unweighted comparisons are about $d = 0.155$ (PS-1 vs NPS) and $d = 0.201$ (PS-2 vs NPS). The maxima are $d = 0.703$ and $d = 0.602$.

To sum up, after accounting for multiple testing, more than half of the variables show significant differences. In addition, the average effect sizes indicate weak effects, but these are larger than, for example, mode effects reported in the literature.[9] However, the strong effect sizes of the maximum values indicate considerable heterogeneity of samples which should reflect the same population.

Effects of weighting
After weighting, Tukey's HSD confirmed remaining significant group differences in 30 out of 36 analyzed health variables (83%). Due to weighting, significant group differences vanished for four variables (injury at work in the last 12 months, asthma, cancer, and wearing a hearing aid).

Of all 430 pairwise differences, 156 are significant (36%). Of those, 77 significant group differences were found between the online recruited surveys (49%), 76 between online and offline recruited surveys (49%), and three between the two offline recruited surveys (2%). These three were reading ability, sports activity and the number of smoked cigarettes per week. Denoting the number of differences as k, most differences between the online recruited surveys were found between NPS-4 and

---

[9] Effect sizes of mode differences are rarely published in survey methodology. However, [56] reports 0.04 as the mean of Cohen's d for 138 items compared between a face-to-face survey and a mixed-mode survey. Compared to these values, the mean effects of NPS vs PS are larger.
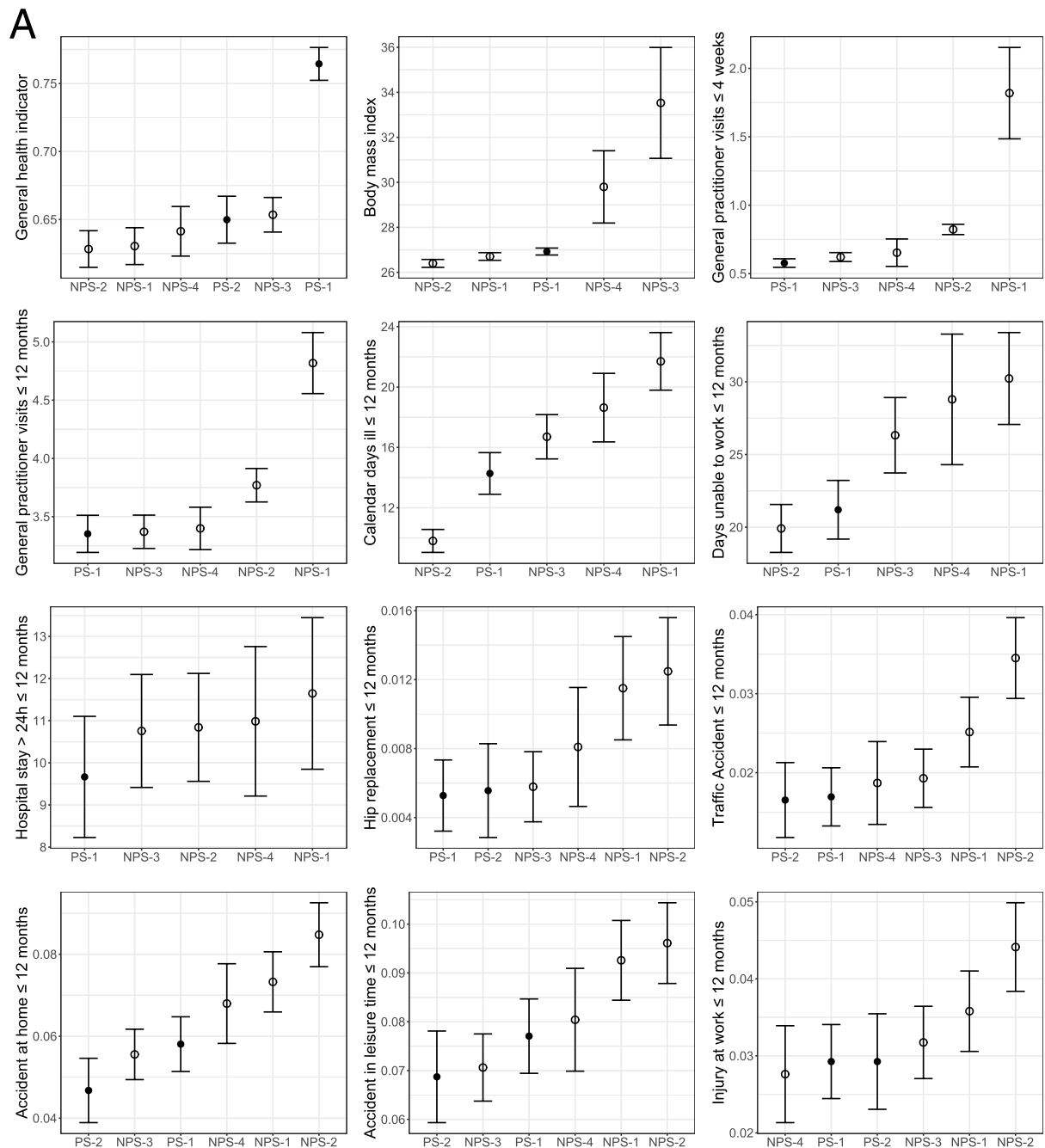
Fig. 2 Unweighted survey estimates and 95%-confidence intervals for each survey. The order of surveys on the x-axis of each subgraph corresponds to the size of the estimate. Empty dots indicate non-probability samples; filled dots indicate offline recruited panels
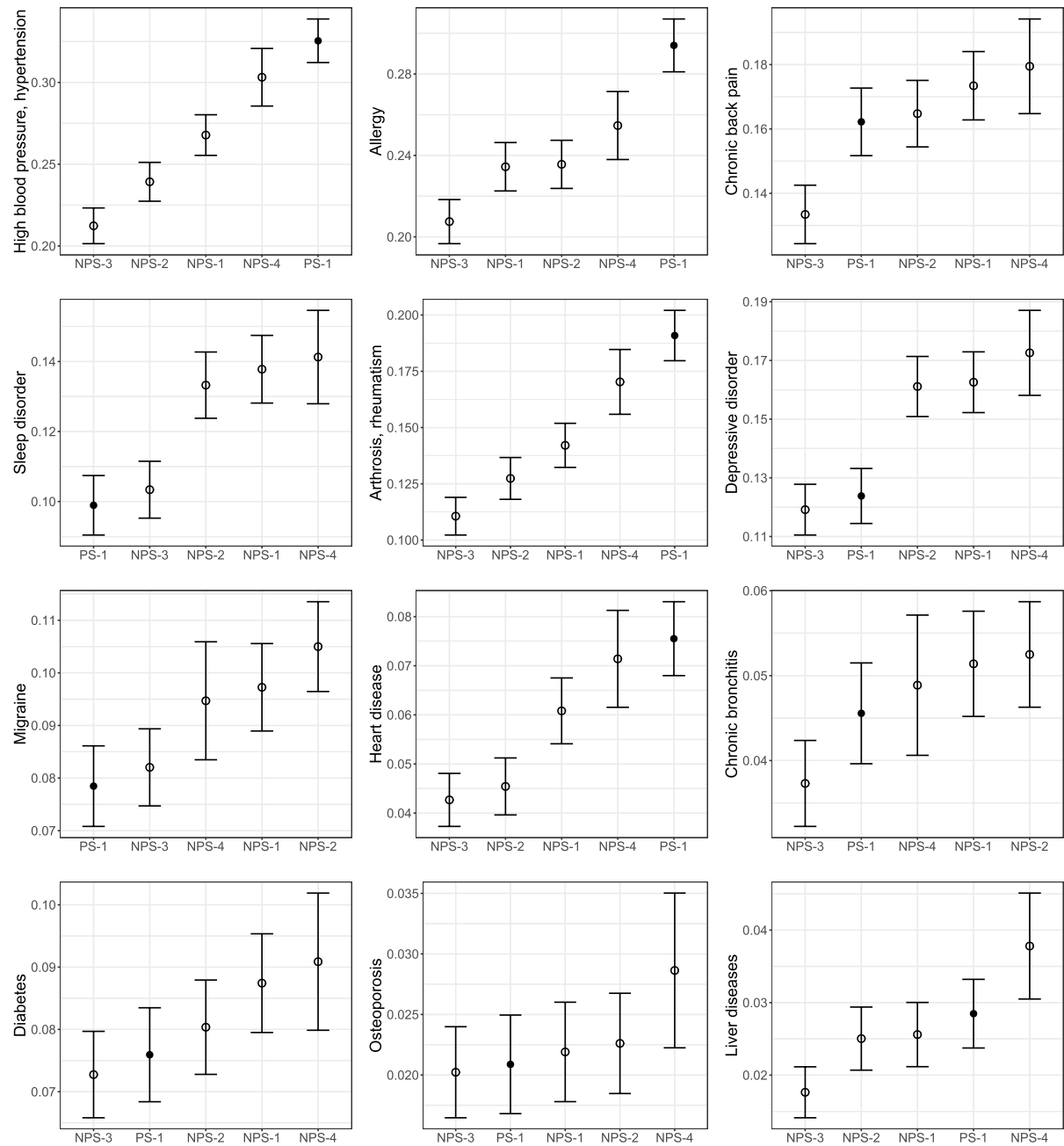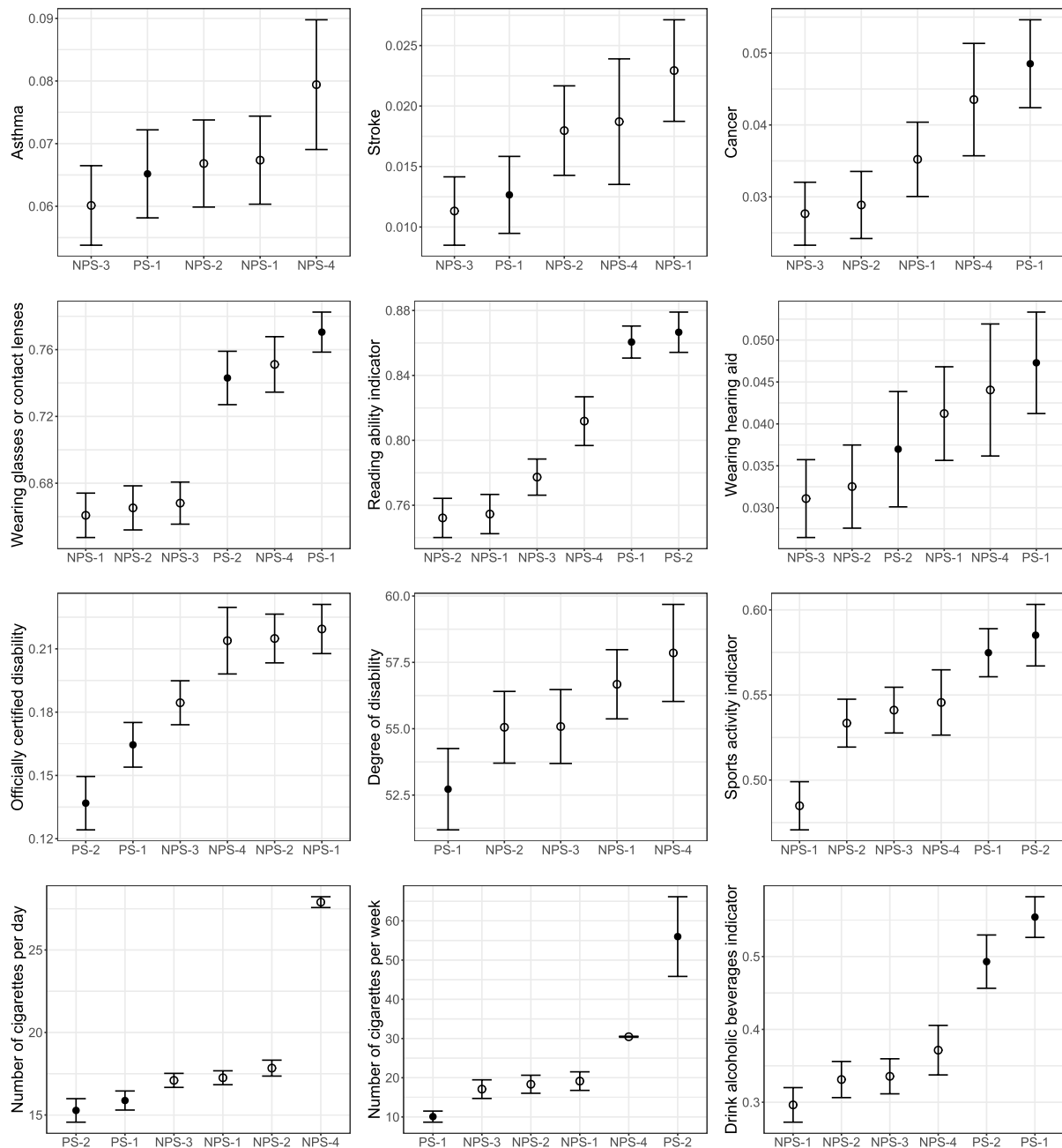
Fig. 2 continued

C

Fig. 2 continued

NPS-2 ($k = 20$), between NPS-4 and NPS-3 ($k = 18$), and least between NPS-2 and NPS-3 ($k = 8$).

In sum, after weighting Tukey's HSD still showed significant group differences between web surveys in 30 out of 36 analyzed health variables. These remaining differences were not specific to a topic but persisted across question groups (general health status, use of health services, accidents/injuries, disabilities/chronic diseases, and health-related behavior).

As mentioned above in the Methods for comparisons section, to reduce the number of comparisons, in addition to the HSD results presented above, we pooled the NPS estimates for the next analysis. Using weighted estimates, comparing PS-1 and NPS yielded 12 variables (36%) with significant differences. Comparing PS-2 and NPS, 7 variables (50%) showed significant differences (see Table 2).

Table 2 Number of Bonferroni adjusted *p*-values of t-tests below conventional thresholds. Reading example for the first two lines: (17+2) t-tests corresponds to *p*-values less or equal 0.05 after correcting for multiple tests. After weighting, (8+4) t-tests still show *p*-values less or equal 0.05

Comparison	$p \leq .001$	$p \leq .01$	$p \leq .05$	$p > .05$
PS-1 vs NPS (unweighted)	17	0	2	17
PS-1 vs NPS (weighted)	8	4	0	24
PS-2 vs NPS (unweighted)	8	0	0	6
PS-2 vs NPS (weighted)	6	1	0	7

Moreover, weighting reduced the mean effect sizes to $d = 0.063$ and $d = 0.127$. The maximum effect sizes after weighting are about $d = 0.255$ (PS-1 vs NPS) and $d = 0.429$ (PS-2 vs NPS), still indicating medium effects, given the traditional classification of effect sizes [55].

To analyse the effect of weighting separately for each item, the relative difference *RD* between the unweighted $\widehat{Y}_u$ and weighted estimates $\widehat{Y}_w$ is computed as

$$RD = \frac{\widehat{Y}_u - \widehat{Y}_w}{\widehat{Y}_w}. \tag{3}$$

Hence, a negative *RD* indicates an underestimation of the unweighted estimate. Figure 3 shows the results. Most comparisons show a negative *RD* (63%). Negative differences up to -45% and positive differences up to 25% are found. In most cases calibration increased the estimates. Hence unweighted estimates would result in underestimations. Therefore, unweighted surveys would overestimate the health of the population. As discussed in the Research background section, it is likely that the weighted estimates are still negatively biased.

However, as Fig. 3 shows, some items such as traffic accidents, liver diseases, or asthma result in large positive differences. Why these specific items produce considerable large overestimations in some NPS and considerable large underestimations in other NPS is unclear. A few outliers are always possible, but the unsystematic pattern, even after calibration, shows that the weighting mechanism does not entirely capture the generating process for the deviations.

Summary

In general, respondents of non-probability samples were less healthy, used more health services, had more accidents and injuries, and showed more unhealthy behaviour than respondents of the probability samples. Weighting using standard demographics was not capable of removing all significant group differences. About 36% of the differences between PS-1 and NPS, and 50% between PS-2 and NPS remained significant after controlling for demographics. Overall, the results show differences between probability and non-probability surveys in health estimates, which were reduced but not eliminated by weighting. Furthermore, the differences between non-probability surveys before and after weighting are larger than expected between random samples from the same population.

Discussion and conclusion

This study was specifically designed to test for differences between web survey agencies conducting the same study. Weighting reduced some differences in estimates between the surveys, but not all. Therefore, weighting seems insufficient to make estimates of different agencies comparable. Health estimates of non-probability based web surveys rely more on a valid weighting model.

Given that the missing generating mechanism might be MNAR and the still large variation of estimates between different surveys after weighting, collecting health data for population parameter estimates with non-probability based surveys seems not advised.

Limitations

Although specifically designed, the study presented has certain limitations. Since we wanted to study differences in actual survey practice, the separate effects of non-probability sampling and offline versus online recruitment or different recruitment techniques can not be estimated separately as both factor groups are confounded. However, separating these factors would require the combination of a non-probability sampling with offline recruitment, which we consider unlikely in practice for general populations.

Due to data protection regulations, 22 variables of PS-2 could not be used. Therefore, we could not compare all estimates of all surveys. Furthermore, PS-2 was conducted later than the other surveys, which could potentially affect comparisons of this survey with the other surveys. However, neither of these two problems are likely to change the main conclusions.

Since we used trimmed weights, it could be argued that untrimmed weights might reduce bias. Trimmed weights are used in practice to avoid a large impact of few observations on results, which would increase the sampling variance [57]. Therefore, trimming weights is common and using a fixed threshold is widespread among commercial survey agencies.

Since only specific auxiliary variables have been used for calibration, other variables might have reduced bias. However, these auxiliary variables must have been measured, and relevant reference data is needed. For particular topics, such information may be available. For example, [58] used 'webographic' variables, and [59] used early adopter
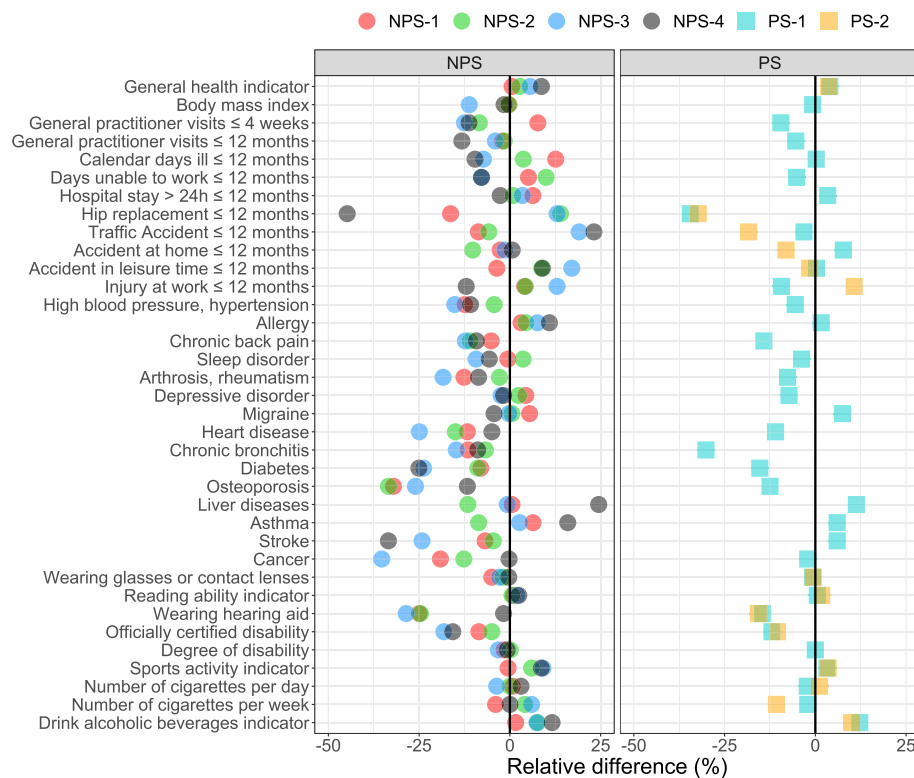
Fig. 3 Relative difference between unweighted and weighted survey estimates per survey agency. The black vertical line indicates a zero difference. Y-axis corresponds to the sequence in the questionnaire

characteristics for weighting. However, since the missing data-generating mechanism for each potential dependent variable could be different, it is unlikely that a universal set of auxiliary variables will be suitable for all variables of interest in a multi-purpose survey. Furthermore, there is no comparative study of other calibration variables than demographics for correcting health bias in web surveys.

Many different weighting procedures have been suggested in the literature; we considered only the most widely used model in official statistics (calibration). Of course, other models could be applied, for example, multilevel regression with poststratification (MRP) [60] or various versions of propensity score adjustment [61, 62]. However, these methods will fail if no relevant information on the missing data generating mechanism is contained in the auxiliary information used in the model [63–66]. Testing this proposition will be the topic of a follow-up paper.

The selection steps for a web-only health survey were described in the Internet use and health section. It is unclear if the effect of health-related issues on participation in health surveys can be explained by ICD codes alone. If symptoms and their severeness and not diagnosis are relevant for survey participation, much more detailed questions than usual within general population surveys are required. Therefore, detailed studies of specific diagnoses and symptoms as factors in answering web surveys seem to be advised.

Only survey estimates were compared, and no external data was used for validation in the study reported here. The details of a validation study comparing weighted and unweighted estimates of the six surveys with external data are subject of ongoing work of the authors.

Availability of data and materials
The programs and survey datasets used in the current study are available from the corresponding author upon reasonable request. The administrative datasets

supporting this study's findings are available from German Official Statistics under access restrictions, requiring an individual licence, which was obtained for this study.

Declarations

Ethics approval and consent to participate

Data was collected online after information on data processing was provided. Informed consent was inferred by the completion of the web survey. There were no experiments on humans, and no human tissue samples were used in this study. An ethical statement for study was not required, since criteria for the need of an ethical statement were not given (risk for the respondents, lack of information about the aims of the study, examination of patients). This is in accordance with the German Research Foundation-guidelines (Deutsche Forschungsgemeinschaft, DFG) available at: http://dfg.de/foerderung/faq/geistes_sozialwissenschaften/ (only available in German language). Data collection was in accordance with the 1964 Helsinki declaration and its later amendments. The Ethics Committee of the Faculty of Social Sciences, University of Duisburg-Essen approved this study (4/23).

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

References

1. Liu H, Cella D, Gershon R, Shen J, Morales LS, Riley W, et al. Representativeness of the Patient-reported Outcomes Measurement Information System Internet Panel. J Clin Epidemiol. 2010;63(11):1169–78. https://doi.org/10.1016/j.jclinepi.2009.11.021.
2. Russell CW, Boggs DA, Palmer JR, Rosenberg L. Use of a Web-based Questionnaire in the Black Women's Health Study. Am J Epidemiol. 2010;172(11):1286–91. https://doi.org/10.1093/aje/kwq310.
3. Haddad C, Sacre H, Zeenny RM, Hajj A, Akel M, Iskandar K, et al. Should samples be weighted to decrease selection bias in online surveys during the COVID-19 pandemic? Data from seven datasets. BMC Med Res Methodol. 2022;22(1):1–11. https://doi.org/10.1186/s12874-022-01547-3.
4. Klingwort J, Buelens B, Schnell R. Early Versus Late Respondents in Web Surveys: Evidence from a National Health Survey. Stat J IAOS. 2018;34(3):461–71. https://doi.org/10.3233/SJI-170421.
5. ADM. Jahresbericht 2019 [annual report 2019, in German]. 2020. https://www.adm-ev.de/wp-content/uploads/2020/09/ADM_Jahresbericht_2019_020920_WEB.pdf. Accessed 01 Sept 2021.
6. Sohlberg J, Gilljam M, Martinsson J. Determinants of Polling Accuracy: The Effect of Opt-in Internet Surveys. J Elections Public Opin Parties. 2017;27(4):433–47. https://doi.org/10.1080/17457289.2017.1300588.
7. Sturgis P, Kuha J, Baker N, Callegaro M, Fisher S, Green J, et al. An Assessment of the Causes of the Errors in the 2015 UK General Election Opinion Polls. J R Stat Soc A (Stat Soc). 2018;181(3):757–81. https://doi.org/10.1111/rssa.12329.
8. Blair J, Czaja R, Blair EA. Designing Surveys: A Guide to Decisions and Procedures. 3rd ed. Thousand Oaks: Sage; 2014.
9. Kreuter F, Presser S, Tourangeau R. Social Desirability Bias in CATI, IVR, and Web Surveys: The Effect of Mode and Question Sensitivity. Public Opin Q. 2008;72(5):847–65. https://doi.org/10.1093/poq/nfn063.
10. McPhee C, Barlas F, Brigham N, Darling J, Dutwin D, Jackson C, et al. Data Quality Metrics for Online Samples: Considerations for Study Design and Analysis. 2023. https://aapor.org/wp-content/uploads/2023/02/Task-Force-Report-FINAL.pdf. Accessed 18 Mar 2023.
11. Cornesse C, Blom AG, Dutwin D, Krosnick JA, De Leeuw ED, Legleye S, et al. A Review of Conceptual Approaches and Empirical Evidence on Probability and Nonprobability Sample Survey Research. J Surv Stat Methodol. 2020;8(1):4–36. https://doi.org/10.1093/jssam/smz041.
12. Pekari N, Lipps O, Roberts C, Lutz G. Conditional distributions of frame variables and voting behaviour in probability-based surveys and opt-in panels. Swiss Political Sci Rev. 2022;28(4):696–711. https://doi.org/10.1111/spsr.12539.
13. Couper MP. Web Surveys: A Review of Issues and Approaches. Public Opin Q. 2000;64(4):464–94. https://doi.org/10.1086/318641.
14. Bethlehem J. Web Surveys in Official Statistics. In: Engel U, Jann B, Lynn P, Scherpenzeel A, Sturgis P, editors. Improving Survey Methods: Lessons from Recent Research. New York: Routledge; 2015. p. 156–69.
15. Leenheer J, Scherpenzeel AC. Does It Pay Off to Include Non-Internet Households in an Internet Panel? Int J Internet Sci. 2013;8(1):17–29.
16. Blom AG, Herzing JME, Cornesse C, Sakshaug JW, Krieger U, Bossert D. Does the Recruitment of Offline Households Increase the Sample Representativeness of Probability-Based Online Panels? Evidence From the German Internet Panel. Soc Sci Comput Rev. 2016;35(4):498–520. https://doi.org/10.1177/0894439316651584.
17. Cornesse C, Schauer I. The Long-Term Impact of Different Offline Population Inclusion Strategies in Probability-Based Online Panels: Evidence From the German Internet Panel and the GESIS Panel. Soc Sci Comput Rev. 2021;39(4):687–704. https://doi.org/10.1177/0894439320984131.
18. Eurostat. Households – level of internet access. 2023. https://ec.europa.eu/eurostat/databrowser/view/isoc_ci_in_h/default/table?lang=en. Accessed 06 Aug 2023.
19. Eurostat. Individuals – internet use. 2023. https://ec.europa.eu/eurostat/databrowser/view/ISOC_CI_IFP_IU/default/table?lang=en. Accessed 06 Aug 2023.
20. United States Census Bureau. Types of Computers and Internet subscriptions. 2023. https://data.census.gov/table?q=internet&tid=ACSST1Y2021.S2801. Accessed 06 Aug 2023.
21. Bethlehem J, Biffignandi S. Handbook of Web Surveys. Hoboken: Wiley; 2012.
22. Baker R, Blumberg SJ, Brick JM, Couper MP, Courtright M, Dennis JM, et al. Research Synthesis: AAPOR Report on Online Panels. Public Opin Q. 2010;74(4):711–81. https://doi.org/10.1093/poq/nfq048.
23. Meyer BD, Mok WK, Sullivan JX. Household Surveys in Crisis. J Econ Perspect. 2015;29(4):199–226. https://doi.org/10.1257/jep.29.4.199.
24. Czajka JL, Beyler A. Declining Response Rates in Federal Surveys: Trends and Implications (Background Paper). 2016. Technical Report Final Report – Volume I, Mathematica Policy Research.
25. Williams D, Brick JM. Trends in U.S Face-to-face Household Survey Nonresponse and Level of Effort. J Surv Stat Methodol. 2017;6(2):186–211. https://doi.org/10.1093/jssam/smx019.
26. de Leeuw E, Hox J, Luiten A. International Nonresponse Trends Across Countries and Years: An Analysis of 36 Years of Labour Force Survey Data. Surv Insights Methods Field. 2018;1–11.
27. Daikeler J, Bošnjak M, Lozar Manfreda K. Web Versus Other Survey Modes: An Updated and Extended Meta-Analysis Comparing Response Rates. J Surv Stat Methodol. 2020;8(3):513–39. https://doi.org/10.1093/jssam/smz008.
28. Groves RM, Peytcheva E. The Impact of Nonresponse Rates on Nonresponse Bias: A Meta-Analysis. Public Opin Q. 2008;72(2):167–89. https://doi.org/10.1093/poq/nfn011.
29. Adams J, White M. Health Behaviours in People Who Respond to a Web-based Survey Advertised on Regional News Media. Eur J Pub Health. 2007;18(3):335–8. https://doi.org/10.1093/eurpub/ckm100.
30. Tourangeau R, Conrad FG, Couper MP. The Science of Web Surveys. New York: Oxford University Press; 2013.
31. Schnell R, Noack M, Torregroza S. Differences in General Health of Internet Users and Non-users and Implications for the Use of Web Surveys. Surv Res Methods. 2017;11(2):105–23. https://doi.org/10.18148/srm/2017.v11i2.6803.
32. Braekman E, Charafeddine R, Demarest S, Drieskens S, Berete F, Gisle L, et al. Comparing Web-based Versus Face-to-face and Paper-and-pencil Questionnaire Data Collected Through Two Belgian Health Surveys. Int J Publ Health. 2020;1–12. https://doi.org/10.1007/s00038-019-01327-9.
33. Dutwin D, Buskirk TD. A Deeper Dive into the Digital Divide: Reducing Coverage Bias in Internet Surveys. Soc Sci Comput Rev. 2022. https://doi.org/10.1177/08944393221093467.

34. Helsper EJ, Reisdorf BC. The emergence of a "digital underclass" in Great Britain and Sweden: Changing reasons for digital exclusion. New Media Soc. 2017;19(8):1253–70. https://doi.org/10.1177/1461444816634676.

35. Zhou XH, Zhou C, Liu D, Ding X. Applied Missing Data Analysis in the Health Sciences. Hoboken: Wiley; 2014.

36. Little RJA, Rubin DB. Statistical Analysis with Missing Data. 3rd ed. Hoboken: Wiley; 2020.

37. Särndal CE, Swensson B, Wretman J. Model Assisted Survey Sampling. New York: Springer; 1992.

38. Särndal CE, Lundström S. Estimation in Surveys with Nonresponse. Chichester: Wiley; 2005.

39. Göritz A. Determinants of the Starting Rate and the Completion Rate in Online Panel Studies. In: Callegaro M, Baker RP, Bethlehem J, Göritz A, Krosnick JA, Lavrakas PJ, editors. Online Panel Research: A Data Quality Perspective. Hoboken: Wiley; 2014. p. 154–70.

40. Güllner M, Schmitt LH. Innovation in der Markt- und Sozialforschung: das forsa.omninet-Panel [Innovations in market research: The fosa.omninet-panel, in German]. Sozialwissenschaften Berufspraxis. 2004;27(1):11–22.

41. Gößwald A, Lange M, Dölle R, Hölling H. Die erste Welle der Studie zur Gesundheit Erwachsener in Deutschland (DEGS1): Gewinnung von Studienteilnehmenden, Durchführung der Feldarbeit und Qualitätsmanagement [The First Wave of the Study of Adult Health in Germany (DEGS1): Recruitment of Study Participants, Fieldwork Implementation, and Quality Management, in German]. Bundesgesundheitsbl Gesundheitsforsch Gesundheitsschutz. 2013;56(5). https://doi.org/10.1007/s00103-013-1671-z.

42. Kamtsiuris P, Lange M, Hoffmann R, Rosario AS, Dahm S, Kuhnert R, et al. Die erste Welle der Studie zur Gesundheit Erwachsener in Deutschland (DEGS1): Stichprobendesign, Response, Gewichtung und Repräsentativität [The first wave of the Study of Adult Health in Germany (DEGS1): sampling design, response, weighting, and representativeness., in German]. Bundesgesundheitsbl Gesundheitsforsch Gesundheitsschutz. 2013;56(5). https://doi.org/10.1007/s00103-012-1650-9.

43. RKI. Beiträge zur Gesundheitsberichterstattung des Bundes - Daten und Fakten: Ergebnisse der Studie Gesundheit in Deutschland aktuell 2012 [Contributions to federal health reporting – Facts and figures: Results of the study on current health in Germany 2012, in German]. Abteilung für Epidemiologie und Gesundheitsmonitoring. Berlin: Robert Koch-Institut; 2014.

44. Saß AC, Lange C, Finger JD, Allen J, Born S, Hoebel J, et al. Supplement: Fragebogen zur Studie 'Gesundheit in Deutschland aktuell': GEDA 2014/2015-EHIS [Supplement: Questionnaire for the study 'Current Health in Germany': GEDA 2014/2015-EHIS, in German]. J Health Monit. 2017;2(1):106–34.

45. Forschungsdatenzentrum ALLBUS. ALLBUS 2014 Fragebogendokumentation: Material zu den Datensätzen der Studiennummern ZA5240 und ZA5241 [ALLBUS 2014 questionnaire documentation: material on the data sets of study numbers ZA5240 and ZA5241, in German]. 2014.

46. Destatis. Statistik und Wissenschaft: Demographische Standards Ausgabe 2010 [Statistics and Science: Demographic Standards Edition 2010, in German]. Wiesbaden; 2010.

47. Wasserstein RL, Schirm AL, Lazar NA. Moving to a World Beyond "p<0.05". Am Stat. 2019;73(sup1):1–19. https://doi.org/10.1080/00031305.2019.1583913.

48. Flury BK, Riedwyl H. Standard Distance in Univariate and Multivariate Analysis. Am Stat. 1986;40(3):249–51. https://doi.org/10.1080/00031305.1986.10475403.

49. Bickel DR. Genomics Data Analysis: False Discovery Rates and Empirical Bayes Methods. Boca Raton: CRC Press; 2020.

50. Callegaro M, Manfreda KL, Vehovar V. Web Survey Methodology. Los Angeles: Sage; 2015.

51. Toepoel V. Doing Surveys Online. London: Sage; 2016.

52. Potter F, Zheng Y. Methods and Issues in Trimming Extreme Weights in Sample Surveys. JSM Proc Surv Res Methods Sect. 2015;2707–2719.

53. Chen Q, Elliott MR, Haziza D, Yang Y, Ghosh M, Little RJA, et al. Approaches to Improving Survey-Weighted Estimates. Stat Sci. 2017;32(2):227–48.

54. Elliott MR. Model Averaging Methods for Weight Trimming. J Off Stat. 2008;24(4):517–40.

55. Cohen J. Statistical Power Analysis for the Behavioral Sciences. 2nd ed. Hillsdale: Erlbaum; 1988.

56. Christmann P, Gummer T, Hähnel S, Wolf C. Does the mode matter? An experimental comparison of survey responses between face-to-face and mixed-mode surveys. Unpublished presentation at the 8th Conference of the European Survey Research Association, Zagreb (Croatia). 2019. Available at https://www.europeansurveyresearch.org/conf2019/uploads/393/3000/71/ESRA_Christmann_et_al..pdf. Accessed 17 July 2019.

57. Haziza D, Beaumont JF. Construction of Weights in Surveys: A Review. Stat Sci. 2017;32(2):206–26. https://doi.org/10.1214/16-STS608.

58. Schonlau M, van Soest A, Kapteyn A. Are 'Webographic' or attitudinal questions useful for adjusting estimates from Web surveys using propensity scoring? Surv Res Methods. 2007;1(3):155–63. https://doi.org/10.18148/srm/2007.v1i3.70.

59. DiSogra C, Cobb C, Chan E, Dennis JM. Calibrating Non-Probability Internet Samples with Probability Samples Using Early Adopter Characteristics. In: Proceedings of Joint Statistical Meetings (JSM). Alexandria: American Statistical Association, Section on Survey Research Methods; 2011. p. 4501–4515.

60. Gelman A, Little TC. Poststratification into many categories using hierarchical logistic regression. Surv Methodol. 1997;23(2):127–35.

61. Rosenbaum PR, Rubin DB. The Central Role of the Propensity Score in Observational Studies for Causal Effects on JSTOR. Biometrika. 1983;70(1):41–55.

62. Lee S, Valliant R. Estimation for Volunteer Panel Web Surveys Using Propensity Score Adjustment and Calibration Adjustment. Sociol Methods Res. 2009;37(3):319–43. https://doi.org/10.1177/0049124108329643.

63. Bruch C, Felderer B. Applying multilevel regression weighting when only population margins are available. Commun Stat Simul Comput. 2022. https://doi.org/10.1080/03610918.2021.1988642.

64. Copas A, Burkill S, Conrad F, Couper MP, Erens B. An evaluation of whether propensity score adjustment can remove the self-selection bias inherent to web panel surveys addressing sensitive health behaviours. BMC Med Res Methodol. 2020;20(1):1–10. https://doi.org/10.1186/s12874-020-01134-4.

65. Si Y. On the Use of Auxiliary Variables in Multilevel Regression and Poststratification. arXiv. 2019. https://doi.org/10.48550/arXiv.2011.00360.

66. Hanretty C. An Introduction to Multilevel Regression and Post-Stratification for Estimating Constituency Opinion. Political Stud Rev. 2019;18(4):630–45. https://doi.org/10.1177/1478929919864773.

Publisher's Note