Electronic publication,
Published on August 3rd, 2025.
https://doi.org/10.33700/jhrs.4.1.137
eISSN: 2820-5480

Alma Mater Europaea University – ECM
Journal of Health and Rehabilitation Sciences, 2025 Aug 3; 4(1), 10–23
Psychology Research

Relative Bias in Health Estimates from Probability-Based Online Panels: Systematic Review and Meta-Analysis

Andrea IVANOVSKA[1],
Michael BOSNJAK[1,2],
Vasja VEHOVAR[1]
[1]Faculty of Social Sciences, University of Ljubljana, Slovenia
[2]Department of Psychology, University of Trier, Germany
Email: andreaivanovska00@gmail.com

Received: 07-Apr-2025
Revised: 22-May-2025
Accepted: 27-May-2025
Online first: 29-May-2025

Research synthesis

Abstract

Introduction: Health surveys require the highest data quality, especially when they inform public health policies. With recent technological developments, probability-based online panels (PBOPs) are becoming an attractive cost-effective alternative to traditional surveys. They are also beginning to be used for official health statistics. However, PBOPs still face concerns about bias, especially for health-related estimates.

Method: Using the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) approach, we conducted a systematic review and meta-analysis of PBOP health survey data quality, with relative bias (RB) of the estimates as the effect size. We analysed 137 health-related survey items from 14 studies and used a linear regression model to examine factors that moderate RB.

Results: The RB varied considerably across the subjects, and its overall median was 12.7%. The highest RBs were exhibited by disabilities (23.6%), mental illnesses (23.2%), personal mental health conditions (20.8%) and drug use (20.7%), and the lowest, by doctor's treatment (2.24%). The measurement levels with ordinal scales (25.8%) showed higher RB, and certain country effects were also observed.

Conclusion: This moderate bias of the health estimates raises concerns about the accuracy of PBOP estimates regarding sensitive health topics. Therefore, PBOP should be used cautiously for official health statistics; and when designing PBOP surveys for health subjects, the item and study characteristics should be included as methodological considerations.

Keywords: probability-based online panels, data quality, relative bias, meta-analysis

Citation: Ivanovska, A., Bosnjak, M., Vehovar V. (2025). Relative Bias in Health Estimates from Probability-Based Online Panels: Systematic Review and Meta-Analysis. *Journal of Health and Rehabilitation Sciences*, 2025 Aug 3; 4(1), 10–23. https://doi.org/10.33700/jhrs.4.1.137

Corresponding address:
Andrea IVANOVSKA
Faculty of Social Sciences, University of Ljubljana
1000 Ljubljana, Slovenia
Email:andreaivanovska00@gmail.com

1. Introduction

Health surveys have been established as an important tool for formulating public health policies, but the utility of their findings depends on the quality of the data they produce (Stevens et al., 2016). To serve this purpose, sampling, measurement, noncoverage and other survey errors need to be minimized. Probability-based online panels (PBOPs) aim to meet this need by using probability-based recruitment methods such as address-based sampling or face-to-face contact (Cornesse et al., 2022).

PBOPs primarily collect data online but may also use a mixed-mode survey. In fact, they often include offline populations by offering them paper surveys or internet access (Bosch & Maslovskaya, 2023). For instance, the GESIS Panel uses both online and paper questionnaires (Bosnjak et al., 2018). Although offline respondents have shown lower retention rates, their inclusion mitigates selection bias over time (Cornesse & Schaurer, 2021). Some PBOPs recruit participants from offline surveys, such as CRONOS, which recruited participants from the European Social Survey (Maslovskaya & Lugtig, 2022). A typical mixed-mode PBOP for health research is the Health in Germany panel, launched by the Robert Koch Institute (Lemcke et al., 2024), which supports online and paper participation. It mitigates potential bias in the estimates by employing a residents' registration sampling frame, incentives and multicontact strategies.

Compared with much cheaper nonprobability (i.e., access) panels, PBOPs generally yield higher data quality. For example, Mercer and Lau (2023) found that PBOP estimates had lower absolute errors (2.6%) than nonprobability panel estimates (5.8%), while Lavrakas et al. (2022) reported the superior accuracy and reliability of Australian PBOPs. However, PBOPs still have limitations, including coverage errors, relatively low recruitment and participation rates, and response heterogeneity (Hays et al., 2015). PBOPs also often overrepresent younger, well-educated males (Bosnjak et al., 2013). Health surveys in PBOPs also raise concerns about social desirability bias in self-reported measures of mental health, substance use and preventive health behaviours (Nayak & Narayan, 2019). Moreover, while self-administered surveys reduce interviewer effects, measurement errors may still be an issue (Cornesse et al., 2021).

To evaluate data quality, PBOP estimates are often benchmarked against official statistics or other external sources (e.g., administrative data) to assess their quality, using absolute bias (AB; Mercer & Lau, 2023), relative bias (RB; Pforr & Dannwolf, 2017), mean squared error (MacInnis et al., 2018) and some other difference measures (Bosnjak et al., 2013). This study focuses on RB, defined as the absolute difference between the PBOP estimate and the benchmark (i.e., AB), divided by the benchmark and expressed as a percentage (Eckman et al., 2015). RB is preferred for its accessibility and its ability to account for effect size. Unlike AB, it contextualises bias magnitude; for instance, holding the absolute bias constant at 2%, the relative bias is 40% when the benchmark is 5%, but only 4% when the benchmark is 50%.

Health data from PBOPs pose distinct methodological challenges due to their reliance on sensitive, self-reported measures like physical or mental health status, behaviors or medical history. These are especially prone to social desirability bias in online self-administered formats (Latkin et al., 2017). Despite this, PBOPs are increasingly used for health surveys because they offer rapid, cost-effective data collection. Some studies report lower AB for health variables (3.9%) than for secondary demographics (5.8%), though with greater variability (Lavrakas et al., 2022). Yeager et al. (2011) found AB in health estimates ranged from 2.6% to 7.0%.

The consequences of poor-quality survey data are not just methodological—they have real-world implications. Biased or inaccurate health estimates can distort public health planning, especially during crises. For instance, flawed data during emergencies may misguide scientific interpretation (do Nascimento et al., 2022), and improperly analysed health surveys, such as the Korean NHANES, have led to misleading conclusions (e.g., suggesting a protective effect of mercury on osteoporosis; Kim et al., 2013). Biased data can also obscure health disparities and shape policy: states with healthier electorates have been shown to spend up to 21.5% less on public health and Medicaid (Pacheco, 2020). These examples highlight why ensuring the quality of health survey estimates—including those from PBOPs—is a public health priority.

In this study, we assessed the quality of PBOPs by systematically reviewing evaluations of data quality related to PBOP health estimates, with a focus on the bias of the estimates. While primary studies have examined bias or described PBOP methodology, the extent of RB in health estimates and how it varies has not been systematically reviewed. This study addressed that gap and identified potential moderators of bias in PBOP health data. We included PBOPs where online was the dominant data collection mode (i.e., that had >51% online respondents). Although various operationalisations of bias exist, we focused on RB because AB does not account for differences in estimate scales, limiting comparability (Eckman, 2015). RB enables more meaningful comparisons across metrics and populations by scaling differences relative to benchmark values. We posed the following research questions:

- RQ1: What is the overall magnitude of RB in health estimates from PBOPs?
- RQ2: To what extent is RB moderated by study and item characteristics (e.g., topic, level of measurement, sensitivity and country of data collection)?

2. Methods

2.1 Literature Selection

This research followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines (Page et al., 2021) for identifying, screening and selecting studies using English-language search terms. The search was conducted on the Digital Library of the University of Ljubljana (DiKUL)[1] academic search engine, which searches 155 databases, including Web of Science, Scopus and PubMed. We used English search terms related to PBOPs, as well as data quality, combined with Boolean operators ('OR' within groups and 'AND' between groups). The PBOP terms included 'probability panel', 'probability-based panel', 'probability online panel', 'probability-based online panel', 'probability web panel', 'probability-based web panel', 'probability internet panel' and 'probability-based internet panel'. The data quality terms included 'difference', 'evaluation', 'comparison', 'data quality', 'bias', 'error' and 'accuracy'. We also conducted citation analyses on references from eligible studies.

Studies were excluded if they:

1. were not related to PBOPs;
2. did not compare PBOP estimates with external benchmarks;
3. lacked empirical information for RB calculation; and
4. did not provide estimates for any health-related variables.

The last date search was 13 January 2025. A total of 216 records were identified using the DiKUL harvester, and an additional 293 records were identified from other sources (mainly citation search). The screening process is illustrated in Figure 1.

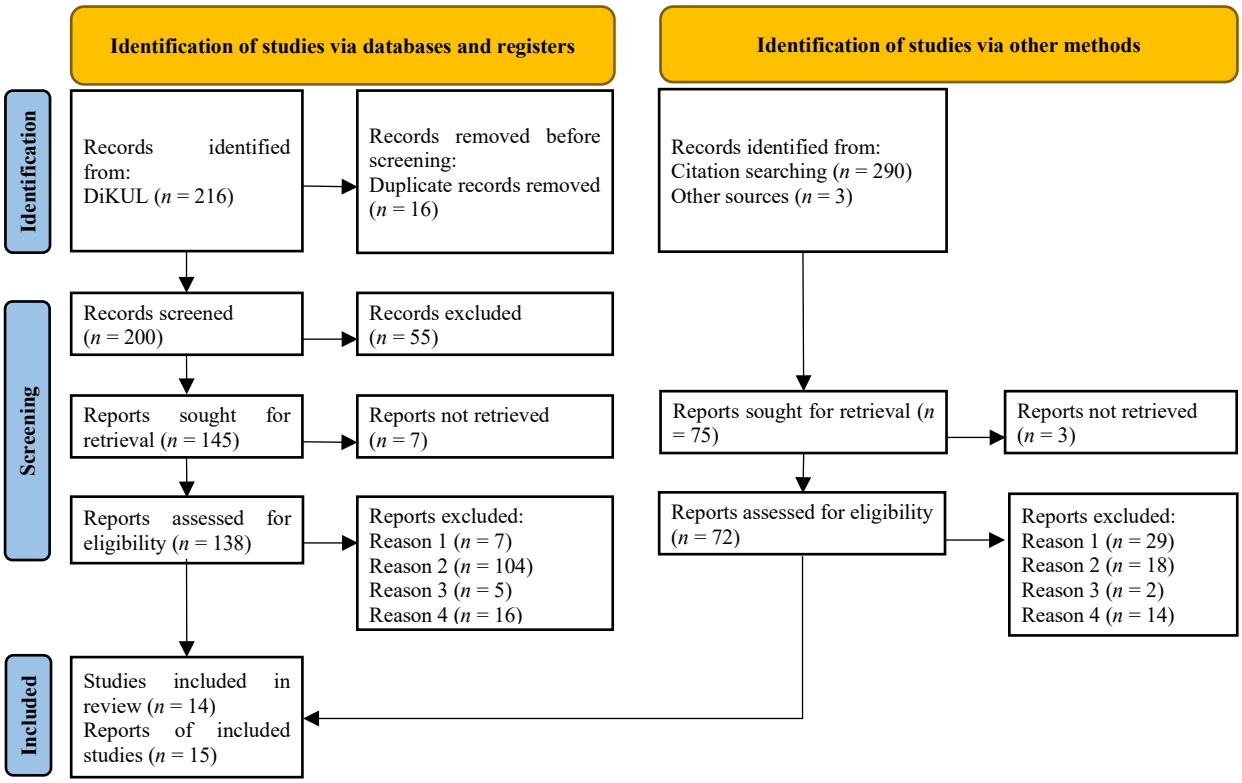

Fig. 1. Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) Diagram Showing the Literature Selection Process

2.2 Data Extraction

We extracted details on panel names, countries, measured variables, measurement levels (nominal, ordinal or interval), sensitivity, specific subject, bias operationalisations and RB estimates. Sensitivity and the specific subject were coded based on the Survey Quality Predictor (2017), which is a coding and predictive tool designed to assess the measurement quality of survey questions. Sensitivity was coded into the following three levels based on social desirability: *Not present* for items unlikely to provoke socially desirable responding (e.g., household size), *A bit* for items where mild desirability effects are plausible (e.g., income, illness), and *A lot* for highly sensitive topics (e.g., drug use, sexual behaviour, racism).

[1] https://viri.ctk.uni-lj.si/

Classifications were guided by the Survey Quality Predictor (2017) definitions and applied consistently across items using a standardised codebook. Each variable was classified into a specific subject (*Personal physical health condition*, *Personal mental health condition*, *Physical illnesses*, *Mental illnesses*, *Disabilities*, *Use of medicine*, *Use of drugs*, *Medical institutions and hospitals*, *Doctor's treatment* and *Other*). For studies that reported multiple bias measures, we recorded all but prioritised that which allowed the easiest RB calculation. Only health-related variables were included. The primary outcome was the reported or derived RB of PBOP health estimates compared to external benchmarks. We also examined how RB values varied across characteristics and methodological choices.

2.3 Meta-Analytic Procedure

RB was calculated as defined in the introduction. We relied on published PBOP estimates, assuming the data were already weighted in accordance with standard reporting and publication practices. When weighted data were not available, we used the unweighted estimates.

Moreover, differences among alternative weighting approaches are typically minor and do not eliminate most of the bias arising from nonresponse or noncoverage (Callegaro et al., 2015; Tourangeau et al., 2013).

The target effect size was the reported or calculated RB of the PBOP health estimates compared to external benchmarks. Traditionally, effect sizes are weighted against their sampling error to account for precision differences across studies (Dettori et al., 2022).

However, in over 95% of studies, the sampling error was negligible (Table 1), making weighting unnecessary. Thus, we aggregated unweighted RB estimates within each subject.

Table 1. Effect Sizes and Corresponding Sampling Variances per Article

Article	Mdn (RB)	M (RB)	Mdn (Sampling variance)	M (Sampling variance)	Count
Bradley et al. (2021)	3.26	3.14	0.0004	0.000433	6
Dever et al. (2021)	5.73	6.64	0.0027	0.00277	3
Herman et al. (2024)	9.09	17.0	0.0024	0.0039	11
Kaczmirek et al. (2019)	8.18	10.3	0.00255	0.00272	4
Kennedy et al. (2016)	20.0	20.0	0.00065	0.00065	2
Kocar & Baffour (2023)	11.3	12.4	0.00265	0.00288	4
Kocar & Biddle (2023)	11.4	12.2	0.0027	0.0029	4
MacInnis et al. (2018)	3.61	5.21	0.0012	0.00159	7
Mercer & Lau (2023)	13.0	29.1	0.0018	0.00359	9
Pennay et al. (2018)	9.80	12.3	0.011	0.0111	4
Spijkerman et al. (2009)	62.5	78.9	0.0081	0.0240	0
Struminskaya et al. (2014)	12.7	12.7	0.0068	0.0068	1
Unangst et al. (2020)	21.1	23.0	0.00375	0.00476	8
Yeager et al. (2011)	8.97	11.4	0.00405	0.0047	4

***Note**: Mdn = median; M = mean; RB = relative bias.*

Because of the low number of observations in several subjects, we consolidated them into broader topics for the threshold analysis and regression. *Mental illnesses* and *Personal mental health condition* were merged into *Mental health*, and *Physical illnesses*, *Personal physical health condition* and *Disabilities*, into *Physical health*. *Use of drugs* was kept separate, while *Use of medicine*, *Doctor's treatment* and *Other* were grouped into *Health practices and Other Health issues*.

We examined RB prevalence exceeding 5%, 10%, 15% and 20% across these groups.

Effect sizes were computed using a linear regression model in R Statistical Software (v4.3.3; R Core Team 2021) with tidyverse (Wickham et al., 2019) and lmerTest (Kuznetsova et al., 2017). We subjected the RB to a log +1 transformation to address skewness, which served as the outcome variable, while *Country*, *Topic*, *Level* and *Sensitivity* were included as moderators to account for potential variations in bias. Nominal and ordinal variables may introduce subjective interpretations and social desirability effects, whereas interval variables are prone to rounding errors and nonresponse patterns (Lalla, 2017). Cultural differences affect response behaviours, with collectivist cultures showing higher social desirability bias and individualist cultures favouring independent responses (Schwarz et al.,

2008), and some cultures leaning towards extreme or socially desirable responses (Matsumoto & van de Vijver, 2012). Sensitivity also moderates bias, as people tend to overreport desirable and underreport undesirable behaviours, and while online surveys reduce interviewer effects, they can lead to higher nonresponse on sensitive questions (Bosch & Maslovskaya, 2023).

Country and *Topic* were assigned sum contrasts. The sampling variances of all the studies were below 0.1, except for two estimates from Spijkerman et al. (2009), which had higher variances because they were based on benchmarks with prevalence rates below 1%. As mentioned, the sampling error was negligible, making precision weighting unnecessary. However, a sensitivity analysis was conducted to address cases with small prevalence rates.

Description of Included Report

The systematic review included 15 reports of 14 studies of PBOPs from the USA, Australia, Germany and the Netherlands (Table 2). All the studies assessed a single panel, except for three studies, which examined multiple panels. Some reports did not disclose the names of the panels. In cases where several reports used the same data, only the earliest report was included (i.e., Struminskaya et al., 2014 instead of Struminskaya et al., 2015; and Pennay et al., 2018 instead of Kaczmirek et al., 2019 for the ANU Poll). A total of 137 items were analysed, with RB estimates available for 136 of them. AB was the most common metric used for data quality assessment, whereas only one study (Kocar & Biddle, 2023) utilised RB.

Table 2. Identified Studies Included in the Systematic Review

Report	*Measures used*	*Panel*	*Country*
Bradley et al. (2021)	*AB*	*Axios-Ipsos*	*USA*
Dever et al. (2021)	*AB*	*National Internet Flu Survey*	*USA*
Herman et al. (2024)	*Percentages*	*KnowledgePanel*	*USA*
Kaczmirek et al. (2019)	*AB*	*Life in Australia*	*Australia*
		*ANU Poll**	*Australia*
Kennedy et al. (2016)	*AB, β coefficients*	*American Trends Panel*	*USA*
Kocar & Baffour (2023)	*AB*	*Life in Australia*	*Australia*
Kocar & Biddle (2023)	*AB, aggregated RB*	*Life in Australia*	*Australia*
MacInnis et al. (2018)	*RMSE*	*Knowledge Networks*	*USA*
Mercer & Lau (2023)	*AB*	*Unnamed PBOP*	*USA*
		Unnamed PBOP	*USA*
		Unnamed PBOP	*USA*
Pennay et al. (2018)	*AB*	*ANU Poll*	*Australia*
Spijkerman et al. (2009)	χ^2	*Dutch online panel of Survey Sampling International LLC*	*Netherlands*
Struminskaya et al. (2014)	χ^2, *β coefficients*	*GESIS Online Panel Pilot*	*Germany*
Struminskaya et al. (2015)	χ^2, *standardised mean difference effect sizes*	*GESIS Online Panel Pilot**	*Germany*
Unangst et al. (2020)	*AB*	*Unnamed PBOP*	*USA*
		Unnamed PBOP	*USA*
Yeager et al. (2011)	*AB*	*Unnamed PBOP*	*USA*

3. Results

Table 3 presents the characteristics of the analysed survey items. Almost 75% of them were measured first at the nominal level, and then, at the ordinal and interval levels. As the items were health-related, more than half of them had high sensitivity, indicating a strong potential for social desirability bias. *Use of drugs* was the most frequently assessed subject (almost half of all cases), followed by *Physical illnesses*, *Personal physical health condition* and *Use of medicine*. Less common subjects (<5%) included *Mental illnesses*, *Disabilities*, *Doctor's treatment* and *Other*.

Table 3. Characteristics of the Analysed Items

Characteristic	Item count	Percent
Measurement level		
Nominal	102	74.5
Ordinal	29	21.2
Interval	6	4.38
Sensitivity		
A lot	77	56.2
A bit	38	27.7
Not present	22	16.1
Subject		
Use of drugs	67	48.9
Physical illnesses	24	17.5
Personal physical health condition	16	11.7
Use of medicine	14	10.2
Personal mental health condition	6	4.38
Doctor's treatment	5	3.65
Mental illnesses	2	1.46
Disabilities	2	1.46
Other	1	0.73

The median RB across all items was 12.7 (Table 4; distribution in Figure 2). Additionally, the RB varied by subject (Figure 3). The highest RB was found for *Disabilities* (23.6%), *Mental illnesses* (23.2%), *Personal mental health condition* (20.8%) and *Use of drugs* (20.7%). *Doctor's treatment* had the lowest RB (2.24%).

Table 4. Relative Bias (RB) Within Subjects

Subject	Median (RB)	MAD	Minimum RB	Maximum RB	Item count
Disabilities	23.6	2.06	22.2	25	2
Mental illnesses	23.2	21.5	8.70	37.7	2
Personal mental health condition	20.8	2.20	13.6	22.9	6
Use of drugs	20.7	28.1	0	260	67
Physical illnesses	13.0	14.3	1.61	50	24
Personal physical health condition	10.3	3.88	1.99	24.6	14
Use of medicine	7.48	8.11	0	71.1	16
Other	3.61	0.00	3.61	3.61	1
Doctor's treatment	2.24	1.32	1.28	4.26	5
All subjects	12.7	15.3	0	260	137

Note. RB = relative bias; MAD = median absolute deviation.

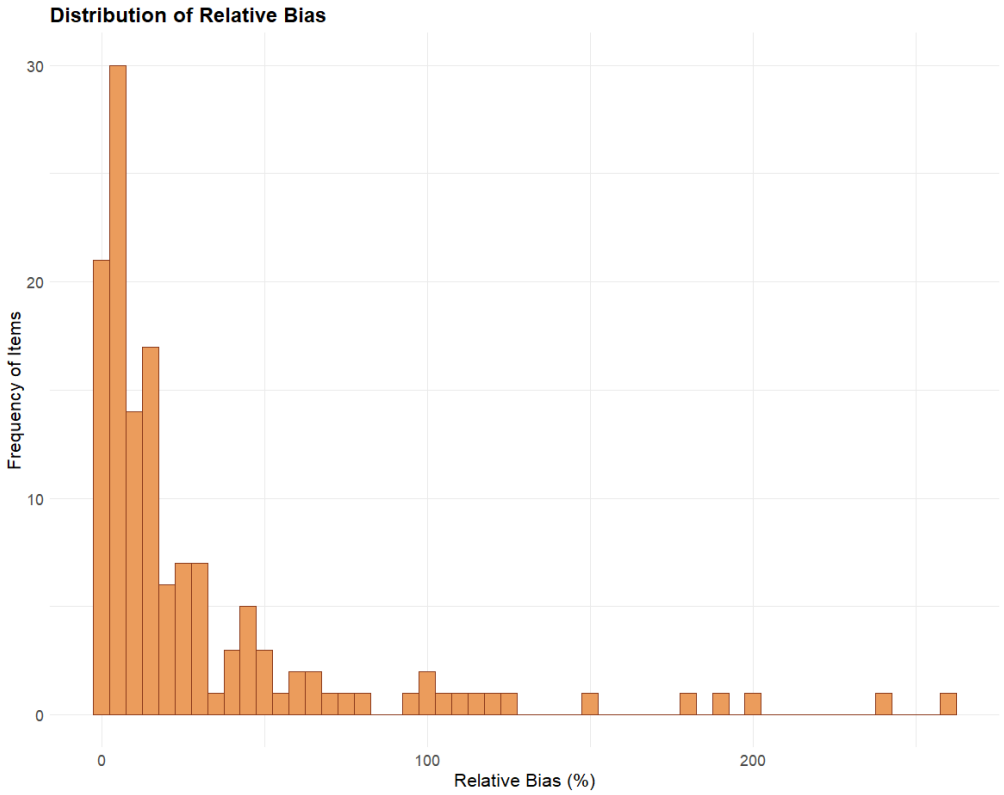

Fig. 2. Distribution of RB Across Items

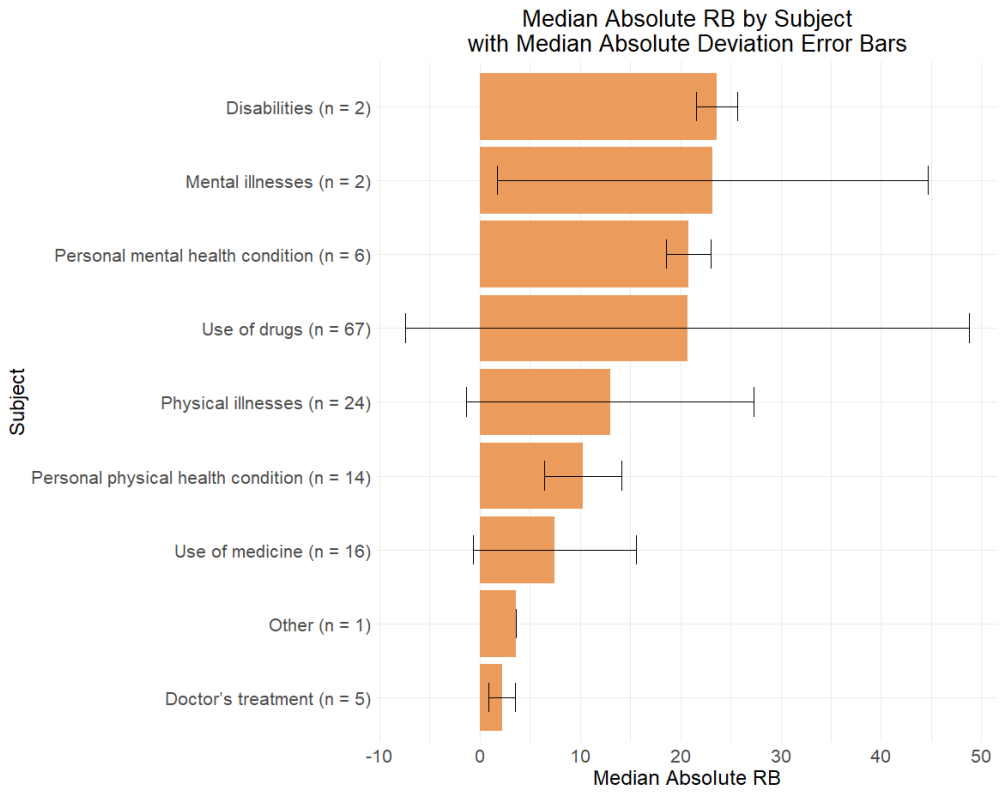

Fig. 3. Distribution of RB by Subject

Table 5 presents the shares of the items that exceeded the RB thresholds across the topics. *Mental health* had the highest RB across all thresholds, with 100% surpassing the 5% RB, while *Health practices and other health issues* had the lowest RB share across all thresholds.

Table 5. Shares of Items Exceeding RB Thresholds (in Percent)

Topic	**>5**	**>10**	**>15**	**>20**
Mental health	100.0	90.0	60.0	50.0
Physical health	76.7	53.5	37.2	34.9
Use of drugs	72.5	57.5	46.2	42.5
Health practices and other health issues	50.0	22.7	18.2	13.6
Total	72.3	53.5	40.6	36.8

The model intercept (α = 2.57 [1.77–3.37], $p < 0.001$) indicated an estimated RB of 12.1% for the reference categories (Table 6). Items from the Netherlands (β = 1.43 [0.63–2.23], $p = 0.001$) exhibited a significantly higher RB of 53.6%. The ordinal-level measurements (β = 0.72 [0.13–1.31], $p = 0.017$) also showed a higher RB of 25.88%. Additionally, a crossed model (Appendix A) was tested to account for variability across the panels and reports. However, the structure was not justified, as it demonstrated a low ICC (0.07) and identified the same effects as the simpler model.

Table 6. Linear Model Results for the Log-Transformed RB

Predictor	**Estimate**	**CI**	***p***
Intercept	2.57	1.77 – 3.37	**<0.001**
Country			
Australia	-0.61	-1.35 – 0.13	0.106
Germany	-0.66	-2.45 – 1.13	0.464
Netherlands	1.43	0.63 – 2.23	**0.001**
Level			
Interval	0.60	-0.49 – 1.68	0.280
Ordinal	0.72	0.13 – 1.31	**0.017**
Sensitivity			
A bit	0.28	-0.40 – 0.96	0.413
A lot	-0.73	-1.67 – 0.20	0.124
Topic			
Use of drugs	0.43	-0.29 – 1.15	0.236
Physical health	-0.00	-0.42 – 0.41	0.985
Mental health	0.32	-0.43 – 1.07	0.402
Observations	136		
R^2 / R^2 adjusted	0.285 / 0.227		

Note. CI = confidence interval; R^2 = coefficient of determination. Bold = statistically significant.

A second model was fitted for a sensitivity analysis, excluding seven estimates with low prevalence (<1%) and, subsequently, those with a sampling error above 0.1. The results remained consistent, confirming an increased RB for items from the Netherlands and for those measured at the ordinal level (Appendix B).

4. Discussion

This analysis provided insights into the data quality of health-related estimates from PBOPs. The descriptive results reveal that most of the items were measured at the nominal level, while fewer items were assessed at the ordinal or interval levels. The measurement levels moderated the RBs, with the ordinal items exhibiting higher RBs than the nominal items. This may be attributed to factors such as increased cognitive burden, ambiguity or response patterns (e.g., midpoint or extreme selection) associated with ordinal scales (Keusch & Yang, 2018).

Overall, in relation to RQ1, the analysis revealed a median RB of 12.7% across all the health-related items, indicating a moderate level of bias in the PBOP estimates despite methodological rigour. Notably, 40% of the estimates had RBs higher than 15%—a concerning level of bias, particularly because a quarter of national statistical institutes are either adopting or preparing to use PBOPs for official statistics (Vehovar et al., 2023), and biased estimates could mislead policymakers and cause resource misallocation. Therefore, given the observed bias, the use of PBOPs in official health statistics warrants caution.

Regarding moderators (RQ2), we found no strong evidence that sensitivity moderates RB.

Most health items are inherently prone to social desirability bias, potentially limiting additional effects of sensitivity. While survey design elements such as question order can mitigate this bias, they were not considered in the current analysis (Schwarz et al., 2008). Furthermore, social desirability bias may be lower in Web surveys than in interviewer-administered modes (Berzelak & Vehovar, 2018), complicating direct comparisons with government survey benchmarks.

At the measurement level, ordinal scales exhibited substantially higher RB than nominal scales. This may be due to greater interpretive variability inherent in ordinal measures (Lalla et al., 2017).

Furthermore, ordinal scales can distort statistical inference, as the spacing between values is not necessarily meaningful. Assuming ordinal properties when scale characteristics are unclear is problematic, as the measures may not even meet the minimal criteria for ordinal data (Kemp & Grace, 2021).

Estimates from the Netherlands showed slightly higher RB, potentially due to the panel design, the topics addressed (i.e., drugs) or cultural response patterns (Matsumoto & van de Vijver, 2012). Future research with more moderators and cross-national comparisons could clarify these differences. Subject-level analysis showed that topics related to *Disabilities*, *Mental illnesses*, *Personal mental health condition* and *Drug use* had the highest RB. These findings are consistent with previous concerns about stigma and social desirability bias in self-reported data, while legal and social implications may influence drug use estimates (Latkin et al., 2017). Disabilities may also be underreported due to societal stigma and self-stigma (Ali et al., 2012).

The sensitivity analyses confirmed the robustness of this study's findings, showing that they were not driven by low-prevalence estimates or sampling errors. However, the validity of the benchmarks is concerning. While government surveys are generally of high quality, they are still subject to nonresponse and measurement biases (Bialik, 2018). In fact, some benchmarks may be less accurate than PBOP estimates. Moreover, differences between government surveys and PBOPs in terms of question phrasing, survey mode and context may affect the comparability of their outcomes.

This study had some limitations. We did not assess publication bias, which could have provided additional insights. For instance, some reports lacked transparency, having omitted panel names or key characteristics, which might have led to overlaps between the unidentified panels.

Second, methodological differences between the studies, such as in their sample sizes, response rates and weighting techniques, could also have moderated RB. Thus, alternative multilevel meta-analyses that account for these factors should be explored. Third, few of the identified panels estimated health data, which may be more prone to social desirability bias due to question sensitivity (Nayak & Narayan, 2019). Fourth, including other domains would enable comparisons to determine whether RB is specific to health data or represents a broader characteristic of PBOPs, while also clarifying the role of sensitivity as a moderator and validating the findings across different domains.

To our knowledge, this is the first systematic review and meta-analysis to synthesise RB in health-related estimates from PBOPs. Although primary studies have reported bias or explored PBOP methodology, none has comprehensively examined how item- or study-level characteristics moderate RB.

This contribution is relevant across public health, psychology, and social sciences, where survey data guide policy, interventions, and service planning. By identifying when and where bias is most likely to occur, our findings can improve survey design, interpretation, and benchmarking practices.

The implications extend beyond survey methodology, supporting more accurate interpretation and use of health data derived from PBOPs.

5. Conclusion

The In summary, this study highlights that while PBOPs offer a methodologically rigorous approach to health data collection, notable levels of bias exist, especially for sensitive topics and specific item formats. These biases, if unaccounted for, pose risks to evidence-based decision-making in public health and related fields. By synthesising existing evidence and identifying bias moderators, this study underscores the need for cautious interpretation of PBOP health estimates and for additional corrective measures when identified moderators are present.

Conflict of interests

The authors have no conflicts of interest to declare.

Conflict of interests

Funding: This work was supported by the Slovenian Research Agency [grant numbers P5-0399, J5-3100, and J5-50159].

References

Ali, A., Hassiotis, A., Strydom, A., & King, M. (2012). Self stigma in people with intellectual disabilities and courtesy stigma in family carers: A systematic review. *Research in Develop mental Disabilities*, *33*(6), 2122–2140. https://doi.org/10.1016/j.ridd.2012.06.013

Berzelak, J., & Vehovar, V. (2018). Mode effects on socially desirable responding in web surveys compared to face-to-face and telephone surveys. *Metodološki Zvezki*, *15*(2), 21–43. From http://www.dlib.si

Bialik, K. (2018, December 6). *How asking about your sleep, smoking or yoga habits can help pollsters verify their findings*. Pew Research Center. https://www.pewresearch.org/short-reads/2018/12/06/how-asking-about-your-sleep-smoking-or-yoga-habits-can-help-pollsters-verify-their-findings/

Bosch, O. J., & Maslovskaya, O. (2023, May 26). *GenPopWeb2: The utility of probability-based online surveys – literature review*. National Centre for Research Methods. From https://www.ncrm.ac.uk/

Bosnjak, M., Dannwolf, T., Enderle, T., Schaurer, I., Struminskaya, B., Tanner, A., & Weyandt, K.W. (2018). Establishing an open probability-based mixed-mode panel of the general population in Germany: The GESIS Panel. *Social Science Computer Review*, *36*(1), 103–115. http://dx.doi.org/10.1177/0894439317697949

Bosnjak, M., Haas, I., Galesic, M., Kaczmirek, L., Bandilla, W., & Couper, M. P. (2013). Sample composition discrepancies in different stages of a probability-based online panel. *Field Methods*, *25*(4), 339–360. https://doi.org/10.1177/1525822X12472951

Bradley, V. C., Kuriwaki, S., Isakov, M., Sejdinovic, D., Meng, X. L., & Flaxman, S. (2021). Unrepresentative big surveys significantly overestimated US vaccine uptake. *Nature*, *600*(7890), 695–700. https://doi.org/10.1038/s41586-021-04198-4

Callegaro, M., Manfreda, K. L., & Vehovar, V. (2015). *Selected topics in web survey implementation* (Chapter 5). In *Web survey methodology* (pp. 191–230). SAGE Publica tions.

Cornesse, C., & Blom, A. G. (2023). Response quality in nonprobability and probability-based online panels. *Sociological Methods & Research*, *52*(2), 879–908. https://doi.org/10.1177/0049124120914940

Cornesse, C., Felderer, B., Fikel, M., Krieger, U., & Blom, A. G. (2022). Recruiting a probability-based online panel via postal mail: Experimental evidence. *Social Science Computer Review*, *40*(5), 1259–1284. https://doi.org/10.1177/08944393211006059

Cornesse, C., Krieger, U., Sohnius, M., Fikel, M., Friedel, S., Rettig, T., Wenz, A., Juhl, S., Lehrer, R., Möhring, K., Naumann, E., Reifenscheid, M., & Blom, A. G. (2021). From German Internet Panel to Mannheim Corona Study: Adaptable probability-based online panel infrastructures during the pandemic. *Journal of the Royal Statistical Society, Series A (Statistics in Society)*, *185*, 773–797.

Dettori, J. R., Norvell, D. C., & Chapman, J. R. (2022). Fixed-effect vs random-effects models for meta-analysis: 3 points to consider. *Global Spine Journal*, *12*(7), 1624–1626. https://doi.org/10.1177/21925682221110527

Dever, J. A., Amaya, A., Srivastav, A., Lu, P. J., Roycroft, J., Stanley, M., Stringer, M. C., Bostwick, M. G., Greby, S. M., Santibanez, T. A., & Williams, W. W. (2021). Fit for purpose in action: Design, implementation, and evaluation of the National Internet Flu Survey. *Journal of Survey Statistics and Methodology*, *9*(3), 449–476. https://doi.org/10.1093/jssam/smz050

Digital Library of University of Ljubljani (DiKUL). http://dikul.uni-lj.si

do Nascimento, I. J. B., Pizarro, A. B., Almeida, J. M., Azzopardi-Muscat, N., Gonçalves, M. A., Björklund, M., & Novillo-Ortiz, D. (2022). Infodemics and health misinformation: A systematic review of reviews. *Bulletin of the World Health Organization, 100*(9), 544–561. https://doi.org/10.2471/BLT.22.288002

Eckman, S. (2015). Does the inclusion of non-Internet households in a Web panel reduce coverage bias? *Social Science Computer Review*, *34*(1), 41–58. https://doi.org/10.1177/0894439315572985

Groves, R. M., & Lyberg, L. (2010). Total survey error: Past, present, and future. *Public Opinion Quarterly*, *74*(5), 849–879. https://doi.org/10.1093/poq/nfq065

Hays, R. D., Liu, H., & Kapteyn, A. (2015). Use of Internet panels to conduct surveys. *Behavior Research Methods*, *47*(3), 685–690. https://doi.org/10.3758/s13428-015-0617-9

Herman, P. M., Slaughter, M. E., Qureshi, N., Azzam, T., Cella, D., Coulter, I. D., DiGuiseppi, G., Edelen, M. O., Kapteyn, A., Rodriguez, A., Rubinstein, M., & Hays, R. D. (2024). Comparing health survey data cost and quality between Amazon's Mechanical Turk and Ipsos' KnowledgePanel: Observational study. *Journal of Medical Internet Research*, *26*, e63032. https://doi.org/10.2196/63032

Kaczmirek, L., Phillips, B., Pennay, D. W., Lavrakas, P. J., & Neiger, D. (2019). *Building a probability-based online panel: Life in Australia™* (CSRM & SRC Methods Paper No. 2/2019). ANU Centre for Social Research & Methods. https://csrm.cass.anu.edu.au/research/publications/building-probability-based-online-panel-life-australia

Kemp, S., & Grace, R. C. (2021). Using ordinal scales in psychology. *Methods in Psychology, 5*, 100054. https://doi.org/10.1016/j.metip.2021.100054

Kennedy, C., Mercer, A., Keeter, S., Hatley, N., McGeeney, K., & Gimenez, A. (2016, May 2). *Evaluating online nonprobability surveys: Vendor choice matters; widespread errors found for estimates based on blacks and Hispanics*. Pew Research Center. https://www.pewresearch.org/methods/2016/05/02/evaluating-online-nonprobability-surveys/

Keusch, F., & Yang, T. (2018). Is satisficing responsible for response order effects in rating scale questions? *Survey Research Methods*, *12*(3), 259–270. https://doi.org/10.18148/srm/2018.v12i3.7263

Kim, Y., Park, S., Kim, N.-S., & Lee, B.-K. (2013). Inappropriate survey design analysis of the Korean National Health and Nutrition Examination Survey may produce biased results. *Journal of Preventive Medicine and Public Health*, *46*(2), 96–104. https://doi.org/10.3961/jpmph.2013.46.2.96

Kocar, S., & Baffour, B. (2023). Comparing and improving the accuracy of nonprobability samples: Profiling Australian surveys. *Methods, Data, Analyses*, *2*(2023). https://doi.org/10.12758/MDA.2023.04

Kocar, S., & Biddle, N. (2023). Do we have to mix modes in probability-based online panel research to obtain more accurate results? *Methods, Data, Analyses*, *16*(1), 93–120. https://doi.org/10.12758/mda.2022.11

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13), 1–26. https://doi.org/10.18637/jss.v082.i13

Lalla, M. (2017). Fundamental characteristics and statistical analysis of ordinal variables: A review. *Quality & Quantity*, *51*(1), 435–458. https://doi.org/10.1007/s11135-016-0314-5

Latkin, C. A., Edwards, C., Davey-Rothwell, M. A., & Tobin, K. E. (2017). The relationship between social desirability bias and self-reports of health, substance use, and social network factors among urban substance users in Baltimore, Maryland. *Addictive Behaviors*, *73*, 133–136. https://doi.org/10.1016/j.addbeh.2017.05.005

Lavrakas, P. J., Pennay, D., Neiger, D., & Phillips, B. (2022). Comparing probability-based surveys and nonprobability online panel surveys in Australia: A total survey error perspective. *Survey Research Methods*, *16*(2), 241–266. https://doi.org/10.18148/srm/2022.v16i2.7907

Lemcke, J., Loss, J., Allen, J., Öztürk, I., Hintze, M., Damerow, S., Kuttig, T., Wetzstein, M., Hövener, C., Hapke, U., Ziese, T., Scheidt-Nave, C., & Schmich, P. (2024). Health in Germany: Establishment of a population-based health panel. *Journal of Health Monitoring*, *9*(Suppl 2), 2–21. https://doi.org/10.25646/11992.2

MacInnis, B., Krosnick, J. A., Ho, A. S., & Cho, M.-J. (2018). The accuracy of measurements with probability and nonprobability survey samples: Replication and extension. *Public Opinion Quarterly*, *82*(4), 707–744. https://doi.org/10.1093/poq/nfy038

Martinsson, J., & Riedel, K. (2015). *Postal recruitment to a probability based web panel: Long term consequences for response rates, representativeness and costs* (LORE working paper 2015:1). University of Gothenburg. https://gup.ub.gu.se/publication/222612?lang=en

Maslovskaya, O., & Lugtig, P. (2022). Representativeness in six waves of cross-national online survey (CRONOS) panel. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 185*(3), 851–871. https://doi.org/10.1111/rssa.12801

Matsumoto, D., & van de Vijver, F. J. R. (2012). Cross-cultural research methods. In H. Cooper (Ed.-in-Chief), P. M. Camic, D. L. Long, A. T. Panter, D. Rindskopf, & K. J. Sher (Eds.), *APA handbook of research methods in psychology, Volume 1: Foundations, planning, measures, and psychometrics* (pp. 85–102). American Psychological Association. https://doi.org/10.1037/13619-000

Mercer, A., & Lau, A. (2023). *Comparing two types of online survey samples.* Pew Research Center. https://www.pewresearch.org/methods/2023/09/07/comparing-two-types-of-online-survey-samples/

Nayak, S. D. P., & Narayan, K. A. (2019). Strengths and Weaknesses of Online Surveys. *IOSR Journal of Humanities and Social Sciences, 24*, 31-38.

Pacheco, J. (2020). The policy consequences of health bias in political voice. *Political Research Quarterly, 73*(4), 935–949. https://doi.org/10.1177/1065912919874256

Page, M. J., Moher, D., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., … McKenzie, J. E. (2021). PRISMA 2020 explanation and elaboration: Updated guidance and exemplars for reporting systematic reviews. *BMJ*, *372*, n160). https://doi.org/10.1136/bmj.n160

Pennay, D. W., Neiger, D., Lavrakas, P. J., & Borg, K. (2018). The Online Panels Benchmarking Study: A Total Survey Error comparison of findings from probability-based surveys and non-probability online panel surveys in Australia (CSRM Methods Series No. 2/2018). Centre for Social Research and Methods. https://csrm.cass.anu.edu.au/research/publications/online-panels-benchmarking-study-total-survey-error-comparison-findings

Pforr, K., & Dannwolf, T. (2017). What do we lose with online-only surveys? Estimating the bias in selected political variables due to online mode restriction. *Statistics, Politics and Policy*, *8*(1), 105–120. https://doi.org/10.1515/spp-2016-0004

R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Schwarz, N., Knäuper, B., & Oyserman, D. (2008). The psychology of asking questions. In E. de Leeuw, J. Hox, & D. Dillman (Eds.), *International handbook of survey methodology* (pp. 18–34). Taylor & Francis.

Spijkerman, R., Knibbe, R., Knoops, K., Van De Mheen, D., & Van Den Eijnden, R. (2009). The utility of online panel surveys versus computer-assisted interviews in obtaining substance-use prevalence estimates in the Netherlands. *Addiction*, *104*(10), 1641–1645. https://doi.org/10.1111/j.1360-0443.2009.02642.x

Stevens, G. A., Alkema, L., Black, R. E., Boerma, J. T., Collins, G. S., Ezzati, M., Grove, J. T., Hogan, D. R., Hogan, M. C., Horton, R., Lawn, J. E., Marušić, A., Mathers, C. D., Murray, C. J., Rudan, I., Salomon, J. A., Simpson, P. J., Vos, T., & Welch, V. (The GATHER Working Group). (2016). Guidelines for accurate and transparent health estimates reporting: The GATHER statement. *The Lancet*, *388*(10062), E19–E23. https://doi.org/10.1016/S0140-6736(16)30388-9

Struminskaya, B., de Leeuw, E., & Kaczmirek, L. (2015). Mode system effects in an online panel study: Comparing a probability-based online panel with two face-to-face reference surveys. *Methods, Data, Analyses*, *9*(1), 3–56. https://doi.org/10.12758/mda.2015.001

Struminskaya, B., Kaczmirek, L., Schaurer, I., & Bandilla, W. (2014). Assessing representativeness of a probability-based online panel in Germany. In M. Callegaro, R. Baker, J. Bethlehem, A. S. Göritz, J. A. Krosnick, & P. J. Lavrakas (Eds.), *Online panel research: A data quality perspective* (Chapter 3). Wiley. https://doi.org/10.1002/9781118763520.ch3

Survey Quality Predictor. (2017). *SQP coding instructions*. Universitat Pompeu Fabra. http://sqp.upf.edu/media/files/sqp_coding_instructions.pdf

Tourangeau, R., Conrad, F. G., & Couper, M. P. (2013). *Measurement error on the web and in other modes of data collection* (Chapter 7). In *The science of web surveys* (pp. 129–150). Oxford University Press.

Unangst, J., Amaya, A. E., Sanders, H. L., Howard, J., Ferrell, A., Karon, S., & Dever, J. A. (2020). A process for decomposing total survey error in probability and nonprobability surveys: A case study comparing health statistics in US Internet panels. *Journal of Survey Statistics and Methodology*, *8*(1), 62–88. https://doi.org/10.1093/jssam/smz040

Vehovar, V., Čehovin, G., & Praček, A. (2023). *The use of probability web panels in national statistical institutes* [Elaborat, predštudija, studia]. Faculty of Social Sciences, Centre for Social Informatics. https://repozitorij.uni-lj.si/IzpisGradiva.php?lang=slv&id=145313

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., ... Yutani, H. (2019). Welcome to the tidyverse. *The Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686

Yeager, D. S., Krosnick, J. A., Chang, L., Javitz, H. S., Levendusky, M. S., Simpser, A., & Wang, R. (2011). Comparing the accuracy of RDD telephone surveys and internet surveys conducted with probability and non-probability samples. *Public Opinion Quarterly*, *75*(4), 709–747. https://doi.org/10.1093/poq/nfr020

Appendix A

Predictor	Estimate	CI	*p*
Intercept	2.54	1.71 – 3.37	**<0.001**
Country			
Australia	-0.58	-1.41 – 0.25	0.166
Germany	-0.66	-2.49 – 1.16	0.473
Netherlands	1.48	0.54 – 2.42	**0.002**
Level			
Interval	0.65	-0.44 – 1.74	0.243
Ordinal	0.69	0.10 – 1.28	**0.023**
Sensitivity			
A bit	0.25	-0.45 – 0.95	0.486
A lot	-0.49	-1.49 – 0.50	0.329
Topic			
Use of drugs	0.18	-0.63 – 0.99	0.659
Physical health	0.06	-0.37 – 0.48	0.792
Mental health	0.31	-0.45 – 1.06	0.422
Random Effects			
σ^2	1.25		
$\tau_{00Panel}$	0.07		
$\tau_{00Article}$	0.03		
ICC	0.07		
N_{Panel}	17		
$N_{Article}$	14		
Observations	136		
Marginal R^2 / Conditional R^2	0.249 / 0.304		

Note. CI = confidence interval; R^2 = coefficient of determination. Bold = statistically significant.

Appendix B

Predictor	Estimate	CI	*p*
Intercept	2.47	1.69 – 3.26	**<0.001**
Country			
Australia	-0.51	-1.24 – 0.21	0.165
Germany	-0.57	-2.32 – 1.18	0.522
Netherlands	1.14	0.34 – 1.95	**0.006**
Level			
Interval	0.60	-0.47 – 1.66	0.269
Ordinal	0.72	0.14 – 1.30	**0.015**
Sensitivity			
A bit	0.28	-0.38 – 0.94	0.403
A lot	-0.73	-1.65 – 0.18	0.117
Topic			
Use of drugs	0.43	-0.27 – 1.14	0.226
Physical health	-0.00	-0.41 – 0.40	0.984
Mental health	0.32	-0.42 – 1.05	0.392
Observations	129		
R^2 / R^2 adjusted	0.216 / 0.150		

Note. CI = confidence interval; R^2 = coefficient of determination. Bold = statistically significant.