mda

methods, data, analyses

Journal for Quantitative Methods and Survey Methodology

Volume 17, 2023 | 2

Frederick G. Conrad et al. — Video in Survey Interviews: Effects on Data Quality and Respondent Experience

Sebastian Kocar & Bernard Baffour — Comparing and improving the accuracy of nonprobability samples: Profiling Australian surveys

Zaza Zindel — Social Media Recruitment in Online Survey Research: A Systematic Literature Review

Henning Silber et al. — The Role of Public Opinion Research in the Democratic Process: Insights from Politicians, Journalists, and the General Public

Franziska Maria Locher et al. — Measuring Students‘ Reading Behavior with an Ambulatory Assessment – A Field Report on a Smartphone-Based Reading Diary Study

Natalia Waechter et al. — Large-Scale Comparative School-Based Survey Research: Challenges and Solutions for Sampling, Fieldwork and Informed Consent

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

Managing editor: Sabine Häder
GESIS – Leibniz Institute for the Social Sciences
PO Box 12 21 55
68072 Mannheim
Germany
Tel.: + 49.621.1246526
E-mail: mda@gesis.org
Internet: www.mda.gesis.org

Methods, data, analyses (mda) publishes research on all questions important to quantitative methods, with a special emphasis on survey methodology. In spite of this focus we welcome contributions on other methodological aspects. We especially invite authors to submit articles extending the profession‘s knowledge on the science of surveys, be it on data collection, measurement, or data analysis and statistics. We also welcome applied papers that deal with the use of quantitative methods in practice, with teaching quantitative methods, or that present the use of a particular state-of-the-art method using an example for illustration.
All papers submitted to mda will first be screened by the editors for general suitability and then double-blindly reviewed by at least two reviewers. Mda appears in two regular issues per year (January and July).

Layout: Bettina Zacharias (GESIS)
Print: Bonifatius Druck GmbH Paderborn, Germany

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

Content

RESEARCH REPORTS

135 Video in Survey Interviews: Effects on Data Quality and Respondent Experience
Frederick G. Conrad, Michael F. Schober, Andrew L. Hupp, Brady T. West, Kallan M. Larsen, Ai Rene Ong & Tianheao Wang

171 Comparing and improving the accuracy of nonprobability samples: Profiling Australian surveys
Sebastian Kocar & Bernard Baffour

207 Social Media Recruitment in Online Survey Research: A Systematic Literature Review
Zaza Zindel

249 The Role of Public Opinion Research in the Democratic Process: Insights from Politicians, Journalists, and the General Public
Henning Silber, Allyson L. Holbrook & Timothy P. Johnson

FIELD REPORTS

273 Measuring Students‘ Reading Behavior with an Ambulatory Assessment – A Field Report on a Smartphone-Based Reading Diary Study
Franziska Maria Locher, Verena Angelika Schnabel, Valentin Unger & Maximilian Pfost

303 Large-Scale Comparative School-Based Survey Research: Challenges and Solutions for Sampling, Fieldwork and Informed Consent
Natalia Waechter, Veronika Kalmus, Giovanna Mascheroni & Signe Opermann

325 Information for Authors

Video in Survey Interviews: Effects on Data Quality and Respondent Experience

Frederick G. Conrad[1], *Michael F. Schober*[2], *Andrew L. Hupp*[1], *Brady T. West*[1], *Kallan M. Larsen*[1], *Ai Rene Ong*[1] & *Tianheao Wang*[1]

[1] *Institute for Social Research, University of Michigan, Ann Arbor*

[2] *Department of Psychology, The New School for Social Research, New York*

Abstract

This study investigates the extent to which video technologies – now ubiquitous – might be useful for survey measurement. We compare respondents' performance and experience (n = 1,067) in live video-mediated interviews, a web survey in which prerecorded interviewers read questions, and a conventional (textual) web survey. Compared to web survey respondents, those interviewed via live video were less likely to select the same response for all statements in a battery (non-differentiation) and reported higher satisfaction with their experience but provided more rounded numerical (presumably less thoughtful) answers and selected answers that were less sensitive (more socially desirable). This suggests the presence of a live interviewer, even if mediated, can keep respondents motivated and conscientious but may introduce time pressure – a likely reason for increased rounding – and social presence – a likely reason for more socially desirable responding. Respondents "interviewed" by a prerecorded interviewer, rounded fewer numerical answers and responded more candidly than did those in the other modes, but engaged in non-differentiation more than did live video respondents, suggesting there are advantages and disadvantages for both video modes. Both live and prerecorded video seem potentially viable for use in production surveys and may be especially valuable when in-person interviews are not feasible.

Keywords: live video survey interviews, video mediated interviews, video web surveys, pre-recorded video interviews, survey satisficing, sensitive questions, disclosure of sensitive information

Since video capability has become standard on computers and smartphones, video communication has become ubiquitous–at least for those with access to the right equipment and connectivity. For many, two-way live video communication has become an indispensable option for remote personal and business communication. One-way video communication has also become commonplace, whether via live streaming (from baby monitors to video doorbells to security surveillance systems) or via the recorded video that has become a fixture of the environment, from television screens in countless public places to online instructional videos to personal videos recorded and posted by smartphone users.

To what extent might video technologies be useful for collecting survey data? Even before video was ubiquitous, survey methodologists investigated the potential of live video for interviewing (Anderson, 2008) and video recordings of interviewers embedded in self-administered questionnaires (e.g., Fuchs, 2009; Fuchs & Funke, 2007; Gerich, 2008; Krysan & Couper, 2003), or both (Jeannis et al., 2013). Since the proliferation of everyday video communication, investigators have compared data quality between traditional modes and live video in the laboratory (Endres et al., 2022) or between traditional modes and embedded recorded video in the field (Haan et al., 2017), concluding that survey data collection using video technologies is feasible and warrants further investigation.

In the current study, we compare two video "interviewing" modes, Live Video and Prerecorded Video (video recordings of an interviewer asking survey questions, embedded in a web survey), with each other and with a conventional web survey, focusing on data quality and respondents' experience completing the questionnaire. We see these comparisons as particularly important as the COVID-19 pandemic has introduced new health and safety concerns about in-person data collection, compounding in-person interviewing's continued challenges and increas-

Acknowledgements

We gratefully acknowledge financial support from NSF grants SES-1825113 and SES-1825194 (Methodology, Measurement, and Statistics program) and NIA grant P30 AG012846 (Michigan Center on the Demography of Aging), and access to participants through NCATS grant UL1TR002240 (Michigan Institute for Clinical & Health Research). We also thank Hanyu Sun, Pooja Varma-Laughlin, Raphael Nishimura, and nine interviewers from the University of Michigan Survey Research Center.

Replication data and documentation are available at
https://www.openicpsr.org/openicpsr/

Study pre-registered at Open Science Foundation:
https://osf.io/2vmx4/?view_only=c90cd24fb46a42d38b285f3453483a37

Direct correspondence to

Frederick G. Conrad, Institute for Social Research, University of Michigan,
Ann Arbor, MI
E-mail: fconrad@umich.edu

ing interest in alternatives (Schober, 2018)[1]. It is important to better understand how video interviews should be designed and implemented (Schober et al., 2020), how video technologies (live or prerecorded) might affect respondent participation, engagement, disclosure, rapport, or conscientiousness, and how video interviewing (live or recorded) might compare with data collection modes currently in use with respect to access, data quality, or cost.

Our strategy was to compare response quality for the same 36 survey questions in each of these three modes, with questions in the live and prerecorded video modes asked by the same 9 interviewers (a larger number than in prior studies). We examine data quality with four widely used measures of conscientious responding that presumably reflect respondents' thoughtfulness, i.e., the extent to which respondents are investing full effort in answering rather than taking mental shortcuts or "satisficing" (Krosnick, 1991; Krosnick & Alwin, 1987; C. Roberts et al., 2019; Simon, 1956) and honesty, i.e., providing socially undesirable and likely uncomfortable but also likely truthful answers (e.g., Schaeffer, 2000; Tourangeau & Smith, 1996). To measure thoughtfulness in answering objective factual questions that require numerical responses, we measured the prevalence of rounded ("heaped") answers, i.e., ending in a zero or a five; in general, unrounded answers are assumed to be more likely to result from deliberate, memory-based thought processes than from estimation (Brown, 1995; Conrad, Brown, & Cashman, 1998), and they have been shown to be more accurate in answers to these kinds of questions (Holbrook et al., 2014).

We measured thoughtfulness in answering multiple questions that use the same response scale, e.g., from "strongly favor" to "strongly oppose," by looking at the extent to which respondents selected the same option for all statements in a battery (Herzog & Bachman, 1981), on the assumption that at least some differentiation in the answers reflects more thoughtful responding (Krosnick, 1991; Roberts et al., 2019). We measured honest responding[2] through increased reporting of socially undesirable information such as more visits to pornography sites or more reports of not voting in local elections on the assumption that more embarrassing or stigmatized answers to survey questions are more likely to be true (e.g., Kreuter et al., 2008; Tourangeau & Yan, 2007; Turner et al., 1998). In addition, we use answering a greater proportion of sensitive questions, i.e., fewer refusals to answer them, as additional evidence of honesty.

[1] We use "in-person" for interviews with physically copresent participants rather than "face-to-face," as live video certainly involves faces, potentially amplifying their importance compared to in-person interactions.

[2] We use the term "honest" though we recognize that more socially desirable responding can occur for many reasons and does not necessarily involve a conscious intention to mislead (e.g., Schaeffer, 2000; Schober & Glick, 2011).

Our strategy for measuring respondent experience during data collection was to ask post-interview, online debriefing questions about how respondents had felt during the survey and (for the live and prerecorded video respondents) about any technical problems they may have experienced during the interview.

Features of the Modes and Implications for Response Quality

To develop expectations about how the quality of data collected in the three modes might differ, we have decomposed the modes into (at least some of) their features. This is presented in Table 1. The modes as we implemented them differ on several features, any of which or any combination of which could affect response quality and respondent experience. The values in the table suggest that live video interviews create social presence of the interviewer – a sense that a human interlocutor is present (Lind et al., 2013): respondents and live interviewers can engage in dialogue, and the interviewer's facial expressions can change based on the respondents' speech and behavior; the spoken questions and facial movement in prerecorded video may create a weaker sense of social presence. The web survey mode and prerecorded video are self-administered in the sense that the respondent controls the flow of the "interview"; self-administration likely creates a greater sense of privacy for respondents than is present in live video interviews (e.g., Kreuter, Presser & Tourangeau, 2008; Tourangeau & Smith, 1996).

Based on these features, how might live video interviewing affect response quality relative to a web survey? For *thoughtful responding*, the increased social presence of the interviewer in live video could lead respondents to feel more accountable for their answers or, from another perspective, less able to get away with low effort responding, which could lead to less non-differentiation than in a web survey. Endres et al. (2022) observed a similar result in comparing live video interviews to web surveys.

On the other hand, live video interviews could increase *rounding* by creating time pressure and thus quicker responses to avoid awkward silences as in everyday conversation (e.g., Jefferson, 1988; F. Roberts & Francis, 2013). More specifically, increased time pressure may push respondents to replace more time-consuming recall-and-count strategies with faster estimation processes that are more likely to result in rounded answers (Brown, 1995; Conrad et al., 1998; Holbrook et al., 2014).

With respect to *socially desirable responding*, live video could feel more intrusive and create more opportunity for respondents to feel judged than a web survey, potentially leading respondents to produce fewer socially undesirable (i.e., fewer honest) answers and refuse to answer more questions. Endres et al. (2022) also report more disclosure in a web survey than live video interviews.

Table 1 Features of the three modes that could plausibly affect response quality and respondent experience. Live video interviews and web surveys differ on all these features; prerecorded video shares some features with live video and some with web surveys.

Feature	Live Video	Prerecorded Video	Web Survey
Interview speaks question	Yes	Yes	NA*
Respondent speaks answer	Yes	No	No
Dialogue between interviewer and respondent	Yes	No	No
Facial representation of interviewer	Yes	Yes	No
Interviewer's facial expressions are responsive	Yes	No	NA
Questions are self-administered	No	Yes	Yes
Questions persist beyond respondent's first exposure to them	No	No	Yes
Question is re-presented when…	Interviewer re-reads (aloud) after respondent's request	Respondent re-plays video as needed	Respondent re-reads as needed
Interviewer has perceptual capability (can see and hear respondent)	Yes	No	NA
Interviewer has evaluative capability (can pass judgment on respondent's answers)	Yes	No	NA

*NA = Not applicable

As for respondent *subjective experience*, the same alternate possibilities are plausible. The increased social presence of the interviewer in live video data collection could lead respondents to be generally more satisfied due to establishing rapport and a sense of connection with interviewers, increasing their willingness to answer honestly (Sun et al., 2020) or it could feel intrusive and less private, reducing satisfaction.

Will prerecorded video feel to respondents more like live video, more like a web survey, or, given that it shares some features with both (Table 1), feel somewhere in between? The fact that the prerecorded interviewers speak the survey questions and that their faces are displayed visually and auditorily in the interface, moving as they speak, could activate the same kinds of social responses as might live video interviews, leading to less *non-differentiation* and more honest, i.e., less socially desirable, answers, as well as more positive subjective experiences than in the web survey. But the fact that there is no live interviewer to keep the respondent engaged and accountable or to potentially judge their answers could lead to the same patterns of responding we expect for web surveys[3]. If the latter pattern is observed in our data, it would be consistent with Haan et al.'s (2017) finding of similar levels of socially desirable responding in prerecorded video and web surveys.

Methods

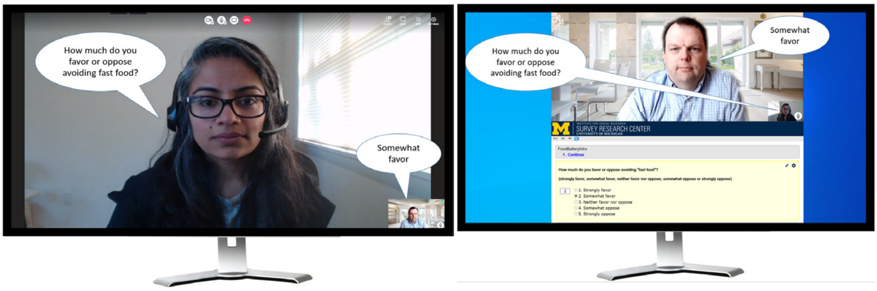
Mode Implementations

All three modes were implemented as a single Blaise 5.6.5 questionnaire which allowed alternate displays appropriate to each mode. Two-way video communication in the Live Video (LV) interviews was conducted via BlueJeans[4]. Except for those on mobile devices, BlueJeans users can join a call through a browser without downloading an app; we expected this to lower the barriers to participation for inexperienced video users. LV respondents were required to schedule the interview beforehand (as opposed to being "cold-called") using Calendly[5] software. LV interviews were conducted from a standard call center carrel with a neutral backdrop (see https://www.mivideo.it.umich.edu/media/t/1_1zoid4cu for an example). To give respondents the sense that the interviewer was looking at them while they were

[3] For a full list for each measure of how patterns of responses in PV could correspond to the patterns in LV and WS responding in this study, see our Open Science Foundation pre-registration,
https://osf.io/2vmx4/?view_only=c90cd24fb46a42d38b285f3453483a37

[4] Versions 2.15 to 2.18. We restricted the study to one platform in order to reduce operational complexity, aware that this might reduce participation among users unfamiliar with the platform (see Schober et al., 2020).

[5] Calendly is continuously updated, so is not identified by version number.


How much do you favor or oppose avoiding fast food?
Somewhat favor
How much do you favor or oppose avoiding fast food?
Somewhat favor
INSTITUTE FOR SOCIAL RESEARCH
SURVEY RESEARCH CENTER
UNIVERSITY OF MICHIGAN

FoodBatteryIntro
1. Continue

How much do you favor or oppose avoiding "fast food"?

(strongly favor, somewhat favor, neither favor nor oppose, somewhat oppose or strongly oppose)

2
○ 1. Strongly favor
● 2. Somewhat favor
○ 3. Neither favor nor oppose
○ 4. Somewhat oppose
○ 5. Strongly oppose

A: Respondent's view

B: Interviewer's view

Figure 1 A) Respondent's screen: Interviewer video fills most of the BlueJeans application window. Respondent's self-view video thumbnail appears in the lower right corner. Speech bubbles contain text of a question the interviewer asked and a possible answer from the respondent.

B) Interviewer's screen: BlueJeans application window (filled primarily by respondent's video with interviewer's self-view video thumbnail in lower right corner) above Blaise instrument. Speech bubbles contain the text of a question that an interviewer asked and a possible answer from the respondent.

answering questions, we positioned the respondents' video window in the upper half of the interviewer's screen (above the Blaise questionnaire) so that by looking at the respondent the interviewer was looking in the direction of the camera (see Figure 1). In LV interviews, the interviewer read the question and response options out loud, manually entering answers in the Blaise questionnaire, as an in-person interviewer would do.

Respondents were able to participate on the device of their choice. The percentages of LV respondents who participated on a desktop/laptop computer versus mobile devices appear in Table 5. See Supplementary Appendix A, Figure 1 for screen images of both desktop/laptop and mobile implementations of LV.

The Prerecorded Video (PV) mode was implemented with video recordings[6] of the same nine interviewers reading the same survey questions embedded in the web display of the Blaise instrument (see https://www.mivideo.it.umich.edu/media/t/1_vjhtigaf for an example). The questions were spoken by the video-recorded interviewers without any textual presentation of the questions. The textually displayed response options appeared automatically on the screen after the video recording of the interviewer reading the question had finished playing. (The on-screen delivery of the response options in PV contrasts with their spoken delivery in LV interviews.)

[6] Recorded and edited using Camtasia version 2019.0.5.4959

In the desktop/laptop version, the prerecorded videos autoplayed to reduce the respondent's effort and to give the delivery of the questions an interviewer-administered character. In the mobile version, this was not possible because autoplay was not implemented in Blaise 5.6 for mobile devices. Thus, these respondents were instructed to click/tap the play button to play each video. Respondents were again able to participate on the device of their choice. The percentages of PV respondents who participated on a desktop/laptop computer versus mobile devices appear in Table 5.

All respondents in PV interviews entered their answers by selecting an option or typing, e.g., an open numerical response. They advanced to the next question by clicking/tapping "Next" (which they could do without answering). See Supplementary Appendix A, Figure 2 for screen images of both desktop/laptop and mobile implementations of PV.

The Web Survey (WS) mode was implemented in Blaise with textually presented questions and response options which appeared on the screen simultaneously (see https://www.mivideo.it.umich.edu/media/t/1_82z2zs7y for an example). The mobile implementation of the WS mode was designed to follow recommended practices for mobile web survey interfaces (Antoun et al., 2018; 2020). In particular, the mobile interface in the WS mode presented large response buttons and large font, fit content to the width of the screen so that horizontal scrolling was not needed, and chose design features that were simple and standard across mobile and desktop operating systems. (In designing the mobile interface for PV interviews, we followed the same design practices to the extent possible, but the screen real estate required us to limit the size of the font and led us to use radio buttons instead of large "clickable" buttons.) Respondents were again able to participate on the device of their choice. The percentages of WS respondents who participated on a desktop/laptop computer versus mobile devices appear in Table 5. See Supplementary Appendix A, Figure 3 for screen images in both desktop/laptop and mobile implementations of WS.

To promote comparability between modes, question batteries were always presented as a series of individual questions even though in the WS mode the batteries could have been implemented as grids. In the PV and WS modes, the display was optimized for screen size, for example using response buttons that included the text of the response within the button for devices with smaller screens, primarily smartphones, and radio buttons for devices with larger screens, primarily computers.

Comparing Data Quality Between Modes

We examine data quality in these three modes by measuring the extent to which respondents' answers were thoughtful, i.e., the extent to which respondents did not take mental shortcuts or "satisfice" (Krosnick, 1991; Krosnick & Alwin, 1987; C. Roberts et al., 2019; Simon, 1956), and the extent to which respondents were willing to disclose sensitive information. We measure thoughtful responding in two ways. First, for questions that require numerical responses we measure the *absence* of thoughtfulness as the prevalence of rounded responses, i.e, non-zero answers that ended in a 0 or a 5 and so were divisible by 5, quantified in two ways: the average percentage of respondents who rounded at least one answer and the average percentage of questions (out of seven) on which rounding is observed.

Second, we measure the absence of thoughtful responding to batteries of questions or statements that use the same response scale, e.g., from "strongly favor" to "strongly oppose," by classifying instances in which the respondent selected a single response option for all statements in a battery as non-differentiation, and instances in which the respondent selected at least two different responses for different statements in a battery as differentiation; our main dependent variable for measuring data quality was whether a respondent did or did not differentiate between the statements in at least one of the three batteries.

We use greater disclosure of sensitive information (e.g., more reported lifetime sexual partners, more reported alcohol use) as evidence of higher quality data, consistent with the evidence that more embarrassing or stigmatized answers are more likely to be true (e.g., Kreuter et al., 2008; Schaeffer, 2000; Tourangeau et al., 2000). We measured disclosure in two ways: the average rated sensitivity of responses to 12 questions concerning potentially sensitive topics and the average number of these questions for which a respondent's answers were sensitive. We quantified the sensitivity of each response to these 12 questions as the proportion of raters who judged that more than 50% of most people would be very or somewhat uncomfortable selecting that option (See Supplementary Appendix B for details).

Items

Main questionnaire. Questionnaire items from previously fielded government and social scientific surveys were selected to allow us to test the three main measures of data quality. Supplementary Appendix B lists the 36 items in the questionnaire along with the corresponding data quality indicator (rounding, non-differentiation, disclosure) that each was included to measure. Supplementary Appendix B also details the item selection procedure. Of the 12 items selected to measure disclosure, six were selected because the topics were rated as (1) very or somewhat uncomfortable for most people to be asked by 50% or more of the raters and (2) for which a sensitive response (i.e., which 50% or more of the raters judged would make most people feel very or somewhat uncomfortable) was likely to be selected for a high proportion of respondents based on response distributions from studies that previously used the questions. Six others were selected that concerned topics *not* rated as sensitive but for which a high proportion of respondents was likely to select a sensitive response, based on the same previous studies. The sensitivity of questions increased from the least (for measuring rounding) to most (for measuring disclosure) over the course of the questionnaire. This design was intended to promote completion of the questionnaire and to minimize missing data.

Measuring respondent experience. We quantified respondents' experience in two ways. First, because the amount of time required to complete a questionnaire has long been used as a measure of respondent burden (e.g., Bradburn, 1979; Hedlin et al., 2005; Office of Management and Budget, 2006; Yan, Fricker, & Tsai, 2020), we calculate mean and median interview duration for the three modes by device type. Second, after respondents completed the main questionnaire, they were directed to an online post-survey questionnaire that included a core set of eight questions about their subjective experience, irrespective of the mode in which they responded to the main questionnaire. This questionnaire included three questions about the interview, two of which were asked only to LV and PV respondents and one of which was asked only to LV respondents, and five questions asked to all respondents about their demographic characteristics. The post-survey questionnaire also included a question about prior use of live video on any device. Most of these items asked respondents to rate their experience on a 5-point scale, with 5 being most positive (see Supplementary Appendix C). Respondents in Live and Prerecorded video interviews were asked if they experienced any of nine technical problems[7]. Another source of data relevant to the experience of LV respondents was transmission logs automatically generated by Bluejeans containing technical information such as video and audio packet loss that might indicate blurred video or choppy audio.

Interviewers and Interviewer Training

Nine telephone interviewers (median years of interviewing experience = 3.5) conducted the LV interviews during their normal on-site work hours. The same nine

[7] We note that technical problems can occur for many reasons that are not under the researchers' control, including the respondent's device and its current level of performance, the respondent's connection speed, network stability and performance, and presumably internet and platform traffic. These can all be affected by the respondent's circumstances at the moment of the interview, for instance the number of simultaneous users on the respondent's network and the resource demands of the simultaneous tasks, ambient noise in the respondent's environment, and even the respondent's ability to troubleshoot technical problems on their own.

interviewers were video-recorded asking the questions; these recordings formed the basis of the PV mode. See Supplementary Appendix D for details about interviewer training, and Schober et al. (2020) for more general considerations about training live video interviewers. The interviewers were all trained in standardized interviewing techniques, designed to reduce interviewer variance by standardizing as much of the data collection as possible.

Respondent Recruitment

In August 2019 we tested the effectiveness of address-based sampling for all three modes but a low response rate in LV (so low that our budget would not allow recruiting the target number of respondents) led us to shift to opt-in, nonprobability sample sources. One potential downside of recruiting participants from online nonprobability sample sources is that panelists may be more technically proficient than the public in general, but this does not necessarily mean that our participants were any more likely at the time of data collection to have previously participated in live or prerecorded video survey interviews. In addition, it is not possible to fully calculate response rates for samples selected from opt-in, non-probability panels (Callegaro & DiSogra, 2008) because it is generally not known (and was not known to us) how many sample members were exposed to, i.e., read, the invitations sent by the sample vendor. Completion rates – recommended by Callegaro and DiSogra – are presented in Supplementary Appendix E.

The respondents were recruited from two opt-in sample sources, CloudResearch (https://www.cloudresearch.com/) and the Michigan Clinical Health Research (MICHR) (https://michr.umich.edu/), targeting estimated 2018 Current Population Survey (CPS) proportions for cross-classes defined by age, gender, race/ethnicity, and education level, and oversampling adults older than 65 years of age (doubling their proportions) to allow exploratory analyses (not reported here) for this age group. In the end, respondents whose highest level of education was high school or less were underrepresented in all cross-classes for LV; to account for the relatively high level of education in the sample, we adjusted statistically for education level in all mode comparisons. For the PV and WS modes, the CPS targets were reached (see Supplementary Appendix F). Sample members were invited to participate in the three modes at random, with substantially more invitations to participate in a live video interview (see Supplementary Appendix E for the number of invitations and completion rates in each mode for each sample source). We were unable to fulfill our quota for LV respondents from CloudResearch so recruited additional respondents from another opt-in sample source, the Michigan Clinical Health Research (MICHR) panel where we enlisted more LV than PV and WS respondents to compensate for the imbalance in Cloud Research (see Supplementary Appendix G for details about inviting sample members and assigning them to a survey mode).

To control for any confounding between sample source and mode we tested the interaction of mode and sample source in all our models; it was never significant, indicating that there was no confound (see Analytic Approach).

Data collection took place between November 2019 and March 2020. See Supplementary Appendix G for further details about recruitment and invitations, incentives, and scheduling constraints.

The total number of completed cases, i.e., cases for which both the main and debriefing questionnaires were submitted, was 1,067. Based on our early experience with Address Based Sampling, we expected sample members assigned to LV interviews to respond at a lower rate than those assigned to the other modes (see Supplementary Appendix E). The number of invitations and the final sample sizes in the three modes for both sample sources appear in Supplementary Appendix E. Note that because we recruited from non-probability, opt-in sample sources, it is not known how many invitations were seen by sample members and thus response rates cannot be calculated, nor can they be interpreted at least comparatively (Callegaro & DiSogra, 2008).

Figure 2 depicts the data collection flow for the full study from recruitment through debriefing and post-paid incentive. Note that LV respondents self-scheduled their interview which necessarily created a lag between screening-in to the study and answering questions; there was no such lag for PV and WS respondents as soon as they had screened in, they were automatically directed to the questionnaire (no scheduling was required because no live interviewers were involved). Thus, it is possible that attrition in LV interviews during the lag could have biased the characteristics of respondents in this mode compared to the other modes. To account for this possibility – and more generally for differences in the characteristics of the responding samples in the three modes – we control for respondent demographics and live video experience in all models (see Analytic Approach).

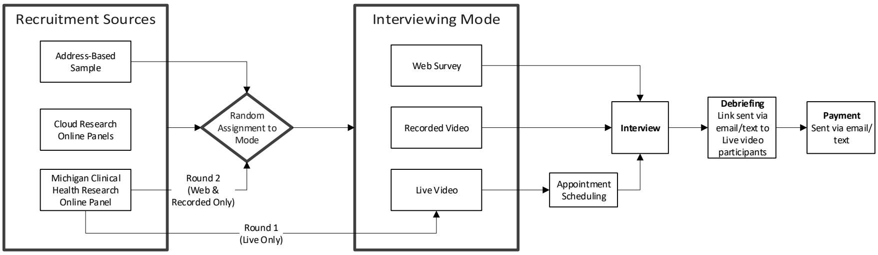

Recruitment Sources
Address-Based Sample
Cloud Research Online Panels
Michigan Clinical Health Research Online Panel
Random Assignment to Mode
Round 2 (Web & Recorded Only)
Round 1 (Live Only)
Interviewing Mode
Web Survey
Recorded Video
Live Video
Appointment Scheduling
Interview
Debriefing
Link sent via email/text to Live video participants
Payment
Sent via email/ text

Figure 2 Data collection flow for the full study from recruitment through debriefing and post-paid incentive.

Analytic Approach

Our analytic strategy involved fitting models to the variables of interest using GEE (with the xtgee function in Stata/SE 16.0), which allowed us to take interviewer clustering into account in order to compare data quality and respondent experience across modes[8]. For all analyses, we excluded cases (respondents) for which any data relevant to the analysis, e.g., responses to numerical questions for analyses of rounding, were missing.

For each outcome variable of interest, all models included mode as a predictor and all key demographic variables as covariates (respondent age, education, gender, and race), as well as prior respondent experience with live video, sample source (CloudResearch vs. MICHR), device type (desktop/laptop computer vs. smartphone vs. tablet), the two-way interaction of age and mode, and the two-way interaction of sample source and mode. Any variables other than mode, age and sample source that were not significant predictors in the first model were removed in the interest of parsimony, and the models were re-fitted iteratively to include mode, age, sample source, and the remaining significant predictors. Please see Supplementary Appendix H for the terms in all the final models.

The interaction of sample source and mode was included in the initial models to test the possibility that the mode differences were driven by differences between the two sample sources, specifically whether the greater proportion of MICHR than CloudResearch respondents in LV and the greater proportion of CloudResearch than MICHR respondents in PV and the WS modes might have been responsible for the patterns of rounding, non-differentiation, and disclosure. The interaction was not significant in any of the initial models, indicating that mode effects appeared to be robust across the sample sources; in the interest of parsimony, we therefore removed this interaction term from all subsequent models.

The interaction of mode x age was included to control for the possibility that older and younger respondents may have differed in how familiar and comfortable they were with the technology used in the three modes and thus have produced different patterns of data quality across the modes. This interaction was significant and thus included in the final models for all three data quality measures as well as for one battery in which non-differentiation was tested and five of the individual statements in the batteries.

We included the main effect of device in the initial models to control for any differences in data quality that might have originated in the device, such as screen

[8] While it is common to model interviewer effects using multilevel models that include random effects of the interviewers, our interest here was in accounting for possible clustering of responses by interviewers in the marginal comparisons between the three modes, not in estimating interviewer variance components. See West et al. (2022) for estimates of interviewer variance components in the data set on which the current article is based.

size or input method (e.g., touch versus mouse). The effect was significant for the overall disclosure models and for one of the battery-level models for non-differentiation; therefore the terms were retained in those models.

We measured rounding with two outcome variables. One such measure was an indicator of respondents rounding at least once (1 if rounded on at least one item and 0 if not); each model predicting this outcome treats it as binary and uses a logit link. A second measure was the count of rounded responses for the seven numerical items, which was treated as binomial with seven possible events for each respondent, and a logit model was fitted to these data. The outcome variable measuring non-differentiation is also treated as binary (1 if the respondent selected the same answer for all statements in at least one battery and 0 if the respondent never selected the same answer for all statements in a battery) and modeled using a logit link. One disclosure measure (mean sensitivity of responses to 12 items) followed a normal distribution and so the models treat the measures as numeric; the other disclosure measure (number of responses out of 12 for which the respondent provided a sensitive answer) is treated as binomial and modeled using a logit link.

For items about respondent experience the approach was the same as for data quality. However, technical problems that respondents may have experienced, there were sometimes too few cases for a model to converge. In these situations, we report raw means (i.e., which were not adjusted for covariates) and test comparisons with pairwise t-tests, applying the Bonferroni correction. For the question asked of respondents in only LV, we report raw means.

Results

Thoughtful Responding: Rounding

Respondents in LV interviews produced rounded answers, i.e., non-zero answers that ended in a 0 or a 5 and so were divisible by 5, more often than did WS respondents. As shown in the top two rows of Table 2, more respondents rounded at least once and the average number of rounded responses was greater in LV than WS, significantly so for the first measure. And LV respondents produced a (non-significantly) greater percentage of rounded responses than did WS respondents (Row 2).

Table 2 Rounding overall (average percentage of rounded responses and percentage of respondents rounding at least once) and for each of the seven items. Standard errors are in parentheses and p-values less than .05 are **bold**.

		Live Video (n = 278)	Web Survey (n = 403)	Prerecorded Video (n = 385)	Live Video vs. Web Survey	Live Video vs. Prerecorded Video	Prerecorded Video vs. Web Survey
					p-value		
Overall Mode Differences							
	% Rs rounding at least once	86.9% (2.5%)	82.0% (1.9%)	76.3% (2.2%)	**0.005**	**0.003**	0.773
	Average % rounded responses	28.6% (2.9%)	24.5% (2.2%)	20.9% (2.2%)	0.286	**0.031**	0.209
Item-Level Mode Differences							
Television Hours	% Rs who rounded	13.4% (2.0%)	17.9% (1.8%)	17.2% (1.0%)	0.121	0.119	0.726
MovieTheaterYear	% Rs who rounded	10.6% (1.7%)	10.5% (1.6%)	11.0% (1.2%)	0.284	0.472	0.633
MoviesYear	% Rs who rounded	61.0% (3.5%)	54.8% (2.5%)	25.9% (2.5%)	0.175	**<0.001**	**<0.001**
Restaurants Month	% Rs who rounded	30.5% (2.6%)	18.4% (2.0%)	16.8% (1.3%)	**<0.001**	**<0.001**	0.477
SpicyFood	% Rs who rounded	24.3% (2.1%)	21.4% (2.1%)	21.8% (1.4%)	0.366	0.340	0.888
GroceryStore	% Rs who rounded	31.7% (2.8%)	22.0% (2.1%)	26.3% (1.9%)	**0.009**	0.126	0.130
Drinking Water	% Rs who rounded	28.1% (3.1%)	26.8% (2.3%)	27.7% (2.5%)	0.752	0.923	0.797

Where did rounding by PV respondents fall relative to that of LV and WS respondents? By both measures, PV respondents rounded least of all. A significantly lower percentage of PV respondents rounded at least once than did LV respondents, although by these measures PV respondents did not round any more or less than did WS respondents.[9]

The overall pattern is less evident at the level of individual items (rows 3-9) but can be seen, nonetheless. For two of the seven items, a significantly larger percentage of LV respondents rounded their numerical answers than did WS respondents and the pattern was in the same direction for six of the seven items. For two of the seven items, a significantly larger percentage of LV than PV respondents rounded, and the same pattern was evident for six of the seven items.

These mode differences in rounding for individual items are potentially consequential. If the actual survey estimates had been the point of the study (as opposed to mode differences), e.g., mean number of movies watched in the last year, these would have been significantly different in LV than WS for two of the seven items, and different in PV than LV interviews for three of the items (see Supplementary Appendix I).

Thoughtful Responding: Non-differentiation

Respondents in LV interviews were less likely to select the same answer for all statements in any of the three batteries than were the respondents in the WS and PV modes. As Table 3 details, a significantly smaller proportion of respondents in LV interviews exhibited non-differentiation, even though our implementation of the questionnaire in WS (individual questions for each statement in a battery rather than a grid) may well have reduced non-differentiation among these respondents compared to what might well have resulted with a grid design (e.g., Mavletova et al., 2018). We observed this pattern for all three batteries, significantly so for the money battery;[10] aggregating the findings for the individual batteries makes the overall pattern (less non-differentiation in LV than in the two self-administered modes) more evident and suggests that LV respondents answered battery items more conscientiously than did respondents in either of the self-administered modes. And, as with rounding, the survey estimates that would have been derived for some

[9] The pattern of results is essentially the same if we define rounding as answers divisible by 10, rather than by 5. The mode comparison p-values are lower (now significant for PV versus WS responses) when rounding is defined as divisible by 10.

[10] The pattern of results is essentially the same if we relax the criterion for what counts as non-differentiation so that providing the same response for all or all but one statement in a battery is counted, although the effects are attenuated. LV interviews led to significantly less of this liberally defined non-differentiation than did the WS mode, and less (though not significantly less) of this behavior than in PV interviews.

Table 3 Non-differentiation (selecting the same option for all statements in a battery) and mode comparisons overall and for each of the three batteries. Standard errors are in parentheses, and p-values less than .05 are **bold**.

	Live Video (n = 278)	Web Survey (n = 403)	Prerecorded Video (n = 385)	Live Video vs. Web Survey	Live Video vs. Prerecorded Video	Prerecorded Video vs. Web Survey
					p-value	
Mode differences overall						
% Rs who selected same option for all questions/ statements in at least one battery	1.7% (1.1%)	7.7% (1.3%)	6.0% (1.3%)	**0.018**	**0.027**	0.634
Food Battery (Q8-Q14)						
% Rs who selected same option for all questions	0.5% (0.5%)	1.0% (0.5%)	1.4% (0.6%)	0.470	0.329	0.697
Money Battery (Q15-Q20)						
% Rs who selected same option for all statements	1.1% (0.7%)	4.7% (1.0%)	2.6% (0.7%)	**0.040**	0.254	0.079
Sports Battery (Q26-Q29)						
% Rs who selected same option for all statements	1.1% (0.8%)	3.7% (0.9%)	3.1% (0.9%)	0.110	0.187	0.608

statements within each battery – had that been the point of the study – differed significantly by mode, presumably due at least in part to mode differences in non-differentiation (see Supplementary Appendix J). The estimates differed significantly by mode for four of seven statements in the food battery and marginally for a fifth, two of six statements in the money battery, and one of four statements in the sports battery and marginally for two additional statements.

Honest Responding: Disclosure

By one measure, respondents in LV interviews disclosed significantly less than did WS respondents. As Table 4 shows, across the 12 items selected to measure disclosure, responses in LV were on average less sensitive, i.e., a smaller proportion of judges rated these items as very or somewhat uncomfortable for respondents to select (row 1). By our second measure (row 2), the number of items out of 12 for which the response was rated as very or somewhat uncomfortable by more than 50% of the raters, LV responses were also less sensitive than WS responses, but not significantly so. At the item level, mode differences in the proportion of responses rated very or somewhat uncomfortable to give were significant for four of the twelve items. Disclosure as measured by average response sensitivity for each item appears in Supplementary Appendix K; the pattern of mode differences by this measure closely parallels the pattern for items in Table 4.

PV respondents disclosed significantly more than did respondents in LV when disclosure is measured by mean response sensitivity for the 12 items (row 1) and they disclosed more (but not significantly so) than WS respondents by the same measure. By the other measure (number of items out of 12 for which the response was rated as sensitive) neither mode difference was significant (row 2). For individual items, significantly more PV respondents provided a sensitive answer than did LV respondents for four items and marginally for a fifth.

As with the other data quality measures, the different modes led to significantly different survey estimates (percent of respondents selecting the most sensitive answers) between modes for five items and marginally different estimates for two items (see Supplementary Appendix L). These mode differences in estimates may well be due to how different modes affect disclosure of sensitive information.

Table 4 Mode differences in disclosure overall and for each of 12 sensitive items. Comparisons of the mean sensitivity (percent of raters judging the response as very or somewhat uncomfortable for most people to give) for all 12 responses (row 1) and number of responses out of 12 independently rated as sensitive (very or somewhat uncomfortable for most people to give) by > 50% of online raters (row 2). The values for individual items are the proportion of responses in the current study that were independently rated as sensitive (very or somewhat uncomfortable to give) by > 50% of online raters. Standard errors appear in parentheses, and p-values less than .05 are **bold**.

		Live Video (n = 271)	Web Survey (n = 396)	Prerecorded Video (n = 377)	Live Video vs. Web Survey	Live Video vs. Prerecorded Video	Prerecorded Video vs. Web Survey
						p-value	
Mode differences overall							
	Mean sensitivity of all 12 responses	0.564 (0.004)	0.581 (0.003)	0.588 (0.003)	**0.001**	**<0.001**	0.110
	Mean number of questions out of 12 for which response is sensitive	3.16 (0.034)	3.20 (0.023)	3.53 (0.029)	0.936	0.498	0.435
Mode differences for each item							
Credit Card Balance	Proportion of sensitive responses	37.0% (2.9%)	40.5% (2.5%)	34.3% (1.9%)	0.097	0.381	0.302
Religious Attendance	Proportion of sensitive responses	66.9% (3.6%)	70.2% (2.3%)	74.8% (2.6%)	0.444	0.078	0.180
Bus Seat	Proportion of sensitive responses	36.4% (3.9%)	33.0% (2.3%)	41.8% (3.3%)	0.469	0.299	**0.025**
Volunteer Work	Proportion of sensitive responses	36.5% (3.7%)	46.3% (2.4%)	55.3% (3.0%)	**0.036**	**<0.001**	**0.019**

		Live Video (n = 271)	Web Survey (n = 396)	Prerecorded Video (n = 377)	Live Video vs. Web Survey	Live Video vs. Prerecorded Video	Prerecorded Video vs. Web Survey
						p-value	
Help Homeless	Proportion of sensitive responses	38.0% (3.9%)	45.3% (2.5%)	54.3% (3.4%)	0.140	**0.003**	**0.032**
Local Elections	Proportion of sensitive responses	20.2% (2.4%)	28.7% (2.3%)	26.0% (2.0%)	**0.019**	0.086	0.352
Sex Partners Year	Proportion of sensitive responses	49.3% (3.0%)	44.5% (2.5%)	48.4% (2.2%)	0.243	0.814	0.229
Female Sex Partner	Proportion of sensitive responses	100.0%	100.0%	100.0%			
Male Sex Partner	Proportion of sensitive responses	100.0%	100.0%	100.0%			
Sex Frequency	Proportion of sensitive responses	100.0%	100.0%	100.0%			
Sex Partner Gender	Proportion of sensitive responses	74.0% (2.4%)	69.7% (1.5%)	73.6% (1.8%)	0.160	0.891	0.090
Porn Frequency	Proportion of sensitive responses	28.6% (2.2%)	38.0% (2.2%)	38.3% (1.6%)	**0.005**	**0.001**	0.901

Honest Responding: Item Nonresponse

Levels of item nonresponse (missing answers in cases that completed the debriefing questionnaire) were low overall (0.08% of responses across all modes[11]), but there was significantly more item nonresponse in LV interviews (4.3% of respondents skipped one or more items in this mode) than in the WS (0.5%) mode (two-sided Fisher's exact test odds ratio 9.01, p < .001). This appears to have been driven by two questions on highly sensitive topics (sex frequency and frequency of visiting a pornography site). The missing data rate for PV interviews (1.8%) was significantly less than the rate for LV interviews (two-sided Fisher's exact test odds ratio 2.43, p < 0.05) and was marginally greater than for the WS (two-sided Fisher's exact test odds ratio 3.71, p = 0.08).

Respondent Experience

Interview Duration

Our first measure of respondent experience is interview duration, as possible evidence of respondent burden (see Table 5). The WS durations are substantially shorter than the durations for the other two modes (t(957) = 17, p < 0.001) and were particularly brief when respondents participated on their smartphones (t(422)=18, p < 0.001), contrary to prior research indicating longer durations for smartphones (Couper & Peterson, 2017).

[11] Responses for 30 out of 39,479 possible responses were missing.

Table 5 Mean Duration (Mins) and Number of Interviews* by Mode and Device

Device		Live Video	Web Survey	Prerecorded Video	Overall
Computer	Avg. Duration	9.84	7.80	12.43	10.10
	Median	9.38	6.69	10.81	9.08
	# Iws	186	187	206	579
	% Within Mode	(66.7%)	(46.5%)	(53.5%)	(54.3%)
Smartphone	Avg. Duration	9.93	5.75	13.79	9.48
	Median	9.47	4.97	11.88	8.83
	# Iws	89	190	155	434
	% Within Mode	(31.9%)	(47.3%)	(40.3%)	(40.7%)
Tablet	Avg. Duration	9.55	6.87	13.42	10.04
	Median	9.73	7.01	10.83	8.96
	# Iws	4	25	24	53
	% Within Mode	(1.4%)	(6.2%)	(6.2%)	(5.0%)
Total # Iws		279	402	385	1066
Avg. Duration		9.87	6.77	13.04	9.85
Median Duration		9.46	5.85	11.25	9.00

*One case (in Web survey, Computer) was excluded as an outlier; its duration was four times that of the next highest case.

Devices

Device use – which was controlled statistically in the data quality models – varied somewhat by survey mode. See Supplementary Appendix A, Figures 1-3 for screen images of both desktop/laptop and mobile implementations. As shown in Table 5, more respondents participated in LV interviews on a desktop/laptop computer (66.7%) than on a mobile device (31.9% smartphone, 1.4% tablet). It is possible that because LV respondents scheduled an interview for a future day and time and were thus aware of the mode in which they would be interviewed, they chose to participate on a relatively big screen more often than on a mobile device for which screens are generally smaller. PV participants responded on smartphones and tablets somewhat more than LV respondents and WS participants responded on smartphones and computers about equally often. In the two self-administered modes it is unlikely respondents chose their devices based on the interview mode as the screener and interview were continuous in these modes: whatever device these participants used to follow the invitation link was almost certainly the mode in which they were interviewed.

Satisfaction

After the primary data collection, respondents in all three modes completed an online (self-administered) debriefing questionnaire about their experience participating in the study. LV respondents were significantly more “satisfied with the survey” and a higher proportion were “very satisfied” than were participants in the two self-administered modes (see Table 6), which did not differ from each other. Consistent with this, in response to a question asked only of the LV respondents (results not in the table), 58.5% reported that they “thoroughly enjoyed” their interaction with the interviewer (mean = 4.4 on a 5-point scale). Only two LV respondents (0.7%) reported not enjoying the interview at all. Comparing just the Live and Prerecorded Video modes, LV respondents reported having felt significantly more connected and more comfortable with the interviewer. The higher satisfaction with LV interviews cannot be attributed to greater familiarity with this mode: substantially *fewer* LV respondents (12.3%) reported using live video “weekly or more” than respondents in the WS (27.5%) and PV (24.2%) modes.

Table 6 Mode differences in respondent experience of the interview, measured in an online post-interview debriefing survey (questions are presented in the order they appeared to respondents). Rows marked with # are items for which there were too few cases for the models to converge, and so raw means and test results from pairwise t-tests and Bonferroni correction are reported.

		Live Video (n = 278)	Web Survey (n = 403)	Prerecorded Video (n = 385)	Live Video vs. Web Survey	Live Video vs. Prerecorded Video	Prerecorded Video vs Web Survey
						p-value	
Overall, how satisfied were you with this survey?	Mean rating	4.66 (0.096)	4.27 (0.077)	4.20 (0.069)	**<0.001**	**<0.001**	0.332
	Percent Rs “very satisfied”	73.9% (28.1%)	52.1% (19.7%)	52.5% (19.2%)	**<0.001**	**<0.001**	0.904
Overall, how comfortable were you with the interviewer?	Mean rating	4.70 (0.108)	-	4.42 (0.098)	-	**0.004**	-
	Percent Rs “very comfortable”	80.9% (23.7%)	-	69.2% (19.8%)	-	**0.018**	-
How personally connected did you feel to the interviewer?	Mean rating	4.59 (0.17)5	-	3.96 (0.150)		**<0.001**	
	Percent Rs “connected” (4 or 5 on 5-point scale)	88.6% (34.0%)	-	71.4% (18.9%)		**<0.001**	
How often did you feel that you were able to answer the questions honestly?	Mean frequency	1.132 (0.048)	1.158 (0.024)	1.138 (0.018)	0.582	0.885	0.470
	Percent “always”	89.5% (20.5%)	87.1% (15.1%)	89.4% (12.0%)	0.296	0.948	0.244

		Live Video (n = 278)	Web Survey (n = 403)	Prerecorded Video (n = 385)	Live Video vs. Web Survey	Live Video vs. Prerecorded Video	Prerecorded Video vs Web Survey
						p-value	
Imagine you had been asked the survey questions in person, that is, in a face-to-face interview. Did the survey you just completed feel more private, the same, or less private than being asked the questions face-to-face?	Percent “more private”	23.4% (15.8%)	58.3% (10.6%)	56.7% (10.4%)	**<0.001**	**<0.001**	0.655
	Percent “same”	75.0% (15.3%)	45.1% (13.3%)	45.5% (14.4%)	**<0.001**	**<0.001**	0.898
Did anyone nearby affect the way you answered the questions?	Percent “Yes, the people nearby affected my answers”#	2.5% (0.9%)	3.0% (0.8%)	1.3% (0.6%)	1.000	0.890	0.338
	Percent “No, no one was around”	59.6% (24.2%)	51.0% (21.1%)	57.1% (20.1%)	**0.038**	0.522	0.083
How sensitive did you feel the survey questions were?	Percent R’s “very sensitive”	16.5% (26.6%*)	19.5% (19.9%*)	21.8% (16.5%)	0.469	0.128	0.540
How often do you participate in live video calls on any device?	Percent Rs weekly or more	11.7% (25.8%*)	26.7% (13.9%)	23.2% (10.0%)	**<0.001**	**<0.001**	0.148
Were you doing something else during the interview?	Percent Rs “yes”#	4.7% (1.3%)	7.7% (1.3%)	4.7% (1.1%)	0.295	1.000	0.212

Privacy

More than half of the LV respondents (56.7%) reported that the survey had felt about as private as an in-person interview would have felt in which the interviewer asked the same questions. An additional 26.7% reported that LV felt *more* private than an in-person interview. In contrast, nearly two thirds of the respondents in the self-administered modes reported that the survey had felt more private than an in-person interview; this evidence is consistent with the general assumption that self-administration increases respondents' sense of privacy (e.g., Tourangeau & Smith, 1996).[12] Nonetheless, respondents in the three modes did not differ significantly in the extent to which they reported that their answers had been affected by nearby others.

Technical Problems

More than half of the LV respondents (52.7%) experienced no problems, and of those who experienced any problems, many (45.5%) experienced only one type of problem. As Supplementary Appendix M shows, each of the 11 types of problems was reported rarely, occurring in 2.5% (Volume too soft) to 18.3% (Interrupted speech – interviewer and respondent were speaking at the same time) of interviews; of those reporting any problems, the median number of reported problems was 2.

Follow-up questions about whether and how these technical problems had been resolved (see Supplementary Appendix M) indicated that more problems resolved themselves than with additional intervention by the respondent, interviewer, or others. In whatever way these problems were resolved (or not), the evidence suggests that they were unrelated to respondent satisfaction with the interview; mean respondent satisfaction was not significantly lower (on a 5 point scale) in interviews that had at least one problem (4.52) than in interviews that had none (4.64), $t(253) = -2.01$, $p = 0.1$. The evidence thus suggests that technical problems were not a major factor in the LV interviews. It is not entirely clear what the technological origins of these problems were, as there was no evidence that the problems in the BlueJeans transmission logs – which were rare – corresponded to respondents' self-reported technical problems.

[12] A small percentage of respondents in all three modes (69 of 1067) reported that this survey had felt less private than an in-person interview (10.4% in LV, 6.0% in PV, 4.2% in the web survey). We can only speculate about why these respondents might see any of these modes as less private than an in-person interview, but perhaps the fact that they are technology-mediated raises the possibility for respondents that their answers may not be secure or that the data collection itself might be subject to surveillance.

Discussion

These findings demonstrate significant advantages – and disadvantages – for data quality and respondents' experience in both video modes relative to a conventional web survey, depending on the data quality measure. More specifically, respondents in LV interviews exhibited higher quality data with respect to non-differentiation – they were less likely to select the same answer for all statements in any of the batteries than respondents assigned to the WS mode – but they exhibited lower data quality by rounding more, disclosing less information that was sensitive, and leaving more sensitive questions unanswered. LV respondents reported significantly higher satisfaction with their experience completing the survey than respondents in either of the self-administered modes.

In our view, the overarching explanation for this pattern of findings concerns the presence or absence of a live interviewer. Live interviewers elicited more conscientious responding (less non-differentiation) than was observed in the self-administered modes but seem to have introduced time pressure leading respondents to provide more rounded numerical answers. And the visual and audio presence of a live interviewer who was clearly thinking and reacting in real time very likely led to the lower levels of disclosing sensitive information than in the two self-administered modes. Although a prerecorded video of an interviewer asking questions seemed to evoke a type of social reaction among participants, e.g., feeling comfortable with and connected to the prerecorded interviewer (although less than respondents in LV interviews felt comfortable with and connected to live interviewers), the mode differences seemed largely driven by whether a live human interviewer asked the questions and interacted with the respondent. Data quality and respondent experience did not differ nearly as much between the two self-administered modes.

Similarity of Live Video Interviewing to In-person Interviewing

Based on the component features of the three modes displayed in Table 1, one would expect the results from LV interviews to be similar to those for in-person interviews (if we had been able to conduct interviews in this mode, despite its greater cost due to interviewer travel expense and the generally higher salaries of field than centralized interviewers). Had we included in-person interviewing in the table, the pattern would have been virtually the same as the pattern for live video interviewing[13]. The primary difference between the features of live video and in-person interviews is that the former mode is mediated and the latter is not, i.e., in in-person interviews, the respondent and interviewer are physically co-present. Yet it is possible that

[13] See Schober et al., in press, for such an analysis of in-person interviews.

these two modes could differ in how they affect responses and subjective experience. To explore this, we look at published mode comparisons involving in-person interviewing and web surveys, as well as the few studies that compare the results from in-person and live video interviewing.

We observed less non-differentiation in LV interviews than in either the PV or WS modes. This closely mirrors the finding by Heerwegh and Loosveldt (2008) that respondents in in-person interviews exhibited less non-differentiation than those in web surveys, and suggests that the involvement of an interviewer, whether physically or virtually present, motivated respondents to attend to all items in the batteries compared to modes in which respondents self-administer batteries of items.

Similarly, our finding of more rounding in LV interviews than in either the PV or WS modes is analogous to the finding by Liu and Wang (2015) of more rounding when respondents answered feeling thermometer questions in person, i.e., when an interviewer asked the questions, than in web surveys. The authors attributed the greater amount of rounding in in-person interviews to greater time pressure in the former mode than in the web survey – the same mechanism we proposed could lead to more rounding in LV than in the self-administered modes.

The disclosure results underscore how socially present the live video interviewer is despite being mediated; as in in-person interviews, this presence seems to inhibit reporting sensitive information compared to self-administered modes such as CASI and ACASI (e.g., Tourangeau and Smith 1996, and many others) and web surveys (e.g., Burkill et al., 2016; Kreuter et al., 2008). It seems to matter to respondents how they are perceived by the LV interviewer, much as it does in person, even though the video interviewers are not physically co-present. There is to our knowledge one reported comparison of data quality in in-person and live video survey interviews, and it is consistent with our impression that the two modes likely produce data of similar quality: Endres et al. (2022) report no differences between these modes for feeling thermometer items, both of which elicited more socially undesirable (colder) responses that did an online (self-administered) questionnaire. The Endres et al. (2022) finding further supports the conclusion that live video and in-person interviews affect respondents in much the same way and are more similar to each other than to online (self-administered) modes. Certainly, the details of how live video and in-person interviewing affect disclosure across a range of topics should be a top priority in future investigations.

There is evidence that LV respondents' subjective experience may resemble that of in-person respondents in other studies. Looking first at rapport, the one study that has compared rapport in live video and in-person interviews (Sun et al., 2020) found no difference between the modes in how respondents rated rapport with interviewers. With respect to perceived privacy, our own results indicate that 56.7% of respondents who had participated in a LV interview rated their experienced privacy as being "the same" as in a hypothetical in-person interview.

Finally, it is possible that much as interviewers in in-person interviews are known to introduce error variance, i.e., to create interviewer effects (e.g., Davis, et al., 2010; West & Blom, 2017), the LV interviewers in the current study may have introduced interviewer effects. While we cannot compare the IICs from the current study to those from in-person interviews, it does not appear that the LV interviewers introduced more error variance than is typically observed in in-person interviews: West et al. (2022) analyzed the data collected by interviewers in LV – as well as in PV – and report that interviewer variance (IICs) was low overall, with all IICs less than 0.02.

Similarity of the two Self-administered Modes

The two self-administered modes are similar to each other in many ways, but as is evident in Table 1, they also differ on several features, primarily those having to do with the presence of an interviewer's facial and vocal attributes in the PV mode. There is a suggestion in the data that the presence of an interviewer, albeit clearly recorded and asynchronous, may help improve data quality by some measures: while there was less rounding in the two self-administered modes than in LV interviews, rounding was reduced even further in the PV than WS data (the former group of respondents rounded on fewer items than the latter group). It is possible that a video-recorded interviewer may amplify respondents' willingness to engage in the generally more effortful recall and count process (the likely origin of reduced rounding) than when the interface is entirely textual (i.e., no facial or vocal representation of an interviewer). Similarly, the greater levels of disclosure for several items (Bus Seat, Volunteer, Help Homeless) in PV interviews than in the WS data may also reflect the interviewers' presence despite their inanimacy. The idea that respondents might react socially to a video recorded interviewer is consistent with Reeves and Nass's (1996) Computers are Social Actors framework. It is possible that such social engagement might be strengthened and thus disclosure further increased as the feel of a live, two-way interview is approximated. For example, it may be possible to enable respondents in prerecorded video interviews to speak their answers rather than just entering them by clicking and typing (Höhne, 2021). The challenge will be to stop short of reintroducing human-like attributes to the extent that they promote socially desirable responses.

While the data collected in the LV and PV modes were high quality by some measures, the WS mode never produced the highest quality data. In fact, the only measure in which the WS respondents outperformed those in the other two modes is the brevity of data collection sessions. This could be due to inherent properties of the modes, e.g., reading questions may take less time than does the delivery of spoken questions, or to our implementation, in particular allowing respondents in PV to enter their answers only after the video had finished playing. Whatever the origin of the shorter WS sessions, this did not lead to higher satisfaction with the experience, as one would expect if duration were a key determinant of respondent burden (e.g., Bradburn, 1979). Instead, the LV respondents reported greater satisfaction than in the other modes despite significantly longer interview sessions.

Considerations in Fielding Live Video Interviews

It is possible the preference for LV interviewing is due to the relative novelty of live video communication in general, at least at the time these data were collected when a significantly smaller percentage of respondents in the LV mode reported frequently using live video (weekly or more often) than in the two self-administered modes (see Pew Research Center, 2021). If this is the case, then the preference for live video data collection could fade as the mode becomes widely used in everyday communication. Alternatively, some respondents may just prefer interacting with a live, albeit mediated, interviewer to self-administering survey questions. Yet, for at least some LV respondents the experience was subtly different than in-person interviews: about a quarter reported that they experienced their interview to be more private than a hypothetical in-person interview, consistent with the suggestion that video mediation can provide a "protective barrier," as observed in training psychologists (Miller & Gibson, 2004). This could bode well for disclosure of sensitive information in live video interviews.

Although LV respondents' interview experience was quite positive, recruiting sample members to participate in this mode was challenging, particularly from one of the online sample vendors. One consequence of this challenge was that a higher proportion of participants in LV interviews were recruited from the medical research panel than in the two self-administered modes. Might this have accounted for any of our findings? We examined this by testing the interaction of mode and sample source in the initial models developed for all our analyses. This interaction was not significant in any of the models, indicating that the effects of mode were unrelated to the panel from which participants were recruited, supporting the interpretation that the results were in fact due to mode differences.

The combination of greater difficulty recruiting LV respondents and a more positive experience for those who ultimately completed the study in this mode suggests that live video interviews may not be for everyone but are quite appealing to some. It could be that as of now live video interviews fit better into a mixed mode, longitudinal research design, or ongoing panel, where sample members are familiar with and presumably trust the research organization than a stand-alone, cross-sectional study. For example, researchers might initially collect data in a mode with which sample members are familiar, e.g., online, on the telephone, or in an in-person interview, after which researchers would invite sample members to participate in future data collection in a mode of their choice (Conrad et al., 2017) where the choices include live video interviews.

Both live and prerecorded video might be combined in a multimodal data collection platform that takes advantage of the strengths of each mode. For example, researchers might extend the ACASI approach to video interviewing by administering non-sensitive questions in a live video interview in which interviewer and respondent are visible and audible to one another when the questions are not sensitive, but when they are sensitive the questions could be administered in a prerecorded video interview.

Before such hybrid approaches can be developed and deployed with confidence, many questions remain about using video – live or prerecorded – in survey data collection. Will the patterns of findings observed here replicate in other samples, with other recruitment methods, with different survey questions and measures of data quality? Will they replicate with different implementations of these modes? Will the cost saving of live video interviews due to the elimination of travel expenses for in-person interviews be sufficient to offset the additional effort – especially in recruitment – that this mode might entail? Will sample members' willingness to participate in live video interviews increase as their comfort with live video communication increases (Schober et al., in press), their access to necessary hardware and software increases, and their familiarity with self-scheduling appointments – not just survey interviews – increases? Are there groups of people who might be more likely to participate in a live video interview than in other modes, e.g., those unwilling to invite an interviewer into their home or who live in areas not easily accessible for in-person interviewers? Whatever the answers to these questions, our findings demonstrate that both live and prerecorded video – at least as we implemented them – are viable survey modes with advantages and disadvantages, worth considering as video communication becomes ever more available – and for many people – central to daily life.

References

Anderson, A. H. (2008). Video-mediated interactions and surveys. In F. G. Conrad & M. F. Schober (Eds.), *Envisioning the survey interview of the future* (pp. 95–118). Wiley. https://doi.org/10.1002/9780470183373.ch5

Antoun, C., Katz, J., Argueta, J., & Wang, L. (2017). Design heuristics for effective smartphone questionnaires. *Social Science Computer Review*, *36*(5), 557–574. https://doi.org/10.1177/0894439317727072

Antoun, C., Nichols, E., Olmsted-Hawala, E., & Wang, L. (2020). Using buttons as response options in mobile web surveys. *Survey Practice*, *13*(1), 1–10. https://doi.org/10.29115/sp-2020-0002

Bradburn, N. (1979). Respondent burden. In L. G. Reeder (Ed.), *Health Survey Research Methods: Second Biennial Conference, Williamsburg, VA*. US Government Printing Office. http://www.asasrms.org/Proceedings/papers/1978_007.pdf

Brown, N. R. (1995). Estimation strategies and the judgment of event frequency. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*(6), 1539–1553. https://doi.org/10.1037/0278-7393.21.6.1539

Burkill, S., Copas, A., Couper, M. P., Clifton, S., Prah, P., Datta, J., Conrad, F., Wellings, K., Johnson, A. M., & Erens, B. (2016). Using the web to collect data on sensitive behaviours: A study looking at mode effects on the British national survey of sexual attitudes and lifestyles. *PLoS ONE*, *11*(2). https://doi.org/10.1371/journal.pone.0147983

Callegaro, M., & Disogra, C. (2008). Computing response metrics for online panels. *Public Opinion Quarterly*, *72*(5), 1008–1032. https://doi.org/10.1093/poq/nfn065

Conrad, F. G., Brown, N. R., & Cashman, E. R. (1998). Strategies for estimating behavioural frequency in survey interviews. *Memory*, *6*(4), 339–366. https://doi.org/10.1080/741942603

Conrad, F. G., Schober, M. F., Antoun, C., Yan, H. Y., Hupp, A. L., Johnston, M., Ehlen, P., Vickers, L., & Zhang, C. (2017). Respondent mode choice in a smartphone survey. *Public Opinion Quarterly*, *81*(S1), 307–337. https://doi.org/https://doi.org/10.1093/poq/nfw097

Couper, M. P., & Peterson, G. J. (2017). Why do web surveys take longer on smartphones? *Social Science Computer Review*, *35*(3), 357–377. https://doi.org/10.1177/0894439316629932

Davis, R. E., Caldwell, C. H., Couper, M. P., Janz, N. K., Alexander, G. L., Greene, S. M., Zhang, N., & Resnicow, K. (2013). Ethnic identity, questionnaire content, and the dilemma of race matching in surveys of African Americans by African American interviewers. *Field Methods*, *25*(2), 142–161. https://doi.org/10.1177/1525822x12449709

Endres, K., Hillygus, D. S., DeBell, M., & Iyengar, S. (2022). A randomized experiment evaluating survey mode effects for video interviewing. *Political Science Research and Methods*, 1–16. https://doi.org/10.1017/psrm.2022.30

Fuchs, M. (2009). Gender-of-interviewer effects in a video-enhanced web survey: Results from a randomized field experiment. *Social Psychology*, *40*(1), 37–42. https://doi.org/10.1027/1864-9335.40.1.37

Fuchs, M., & Funke, F. (2007). Video web survey - Results of an experimental comparison with a text-based web survey. In M. Trotman, T. Burrell, L. Gerrard, K. Anderton, G. Basi, M. Couper, K. Morris, K. Birks, A. Johnson, R. B. (Market Strategies), M. R. (PSI), S. T. (Inputech), & A. W. (Survey & S. Computing) (Eds.), *Proceedings of the Fifth International Conference of the Association for Survey Computing: The Challenges of a Changing World* (pp. 63–80).

Gerich, J. (2008). Real or virtual? Response behavior in video-enhanced self-administered computer interviews. *Field Methods*, *20*(4), 356–376. https://doi.org/10.1177/1525822X08320057

Haan, M., Ongena, Y. P., Vannieuwenhuyze, J. T. A., & De Glopper, K. (2017). Response behavior in a video-web survey: A mode comparison study. *Journal of Survey Statistics and Methodology*, *5*(1), 48–69. https://doi.org/10.1093/jssam/smw023

Hedlin, D., Dale, T., Haraldsen, G., & Jones, J. (2005). *Developing methods for assessing perceived response burden: A joint report of Statistics Sweden, Statistics Norway and the UK Office for National Statistics.* https://ec.europa.eu/eurostat/documents/64157/4374310/10-DEVELOPING-METHODS-FOR-ASSESSING-PERCEIVED-RESPONSE-BURDEN.pdf/1900efc8-1a07-4482-b3c9-be88ee71df3b

Heerwegh, D., & Loosveldt, G. (2008). Face-to-face versus web surveying in a high-internet-coverage population: Differences in response quality. *Public Opinion Quarterly*, *72*(5), 836–846. https://doi.org/10.1093/poq/nfn045

Herzog, A. R., & Bachman, J. G. (1981). Effects of questionnaire length on response quality. *Public Opinion Quarterly*, *45*(4), 549–559. https://doi.org/10.1086/268687

Höhne, J. K. (2021). Are respondents ready for audio and voice communication channels in online surveys? *International Journal of Social Research Methodology*, 1–8. https://doi.org/10.1080/13645579.2021.1987121

Holbrook, A. L., Anand, S., Johnson, T. P., Cho, Y. I., Shavitt, S., Chavez, N., & Weiner, S. (2014). Response heaping in interviewer-administered surveys: Is it really a form of satisficing? *Public Opinion Quarterly*, *78*(3), 591–633. https://doi.org/10.1093/poq/nfu017

Jeannis, M., Terry, T., Heman-Ackah, R., & Price, M. (2013). Video interviewing: An exploration of the feasibility as a mode of survey application. *Survey Practice*, *6*(1), 1–5. https://doi.org/10.29115/sp-2013-0001

Jefferson, G. (1988). Preliminary notes on a possible metric which provides for a "standard maximum" silence of approximately one second in conversation. In D. Roger & P. Bull (Eds.), *Conversation: An interdisciplinary pespective* (pp. 166–196). Multilingual Matters.

Kreuter, F., Presser, S., & Tourangeau, R. (2008). Social desirability bias in CATI, IVR, and web surveys: The effects of mode and question sensitivity. *Public Opinion Quarterly*, *72*(5), 847–865. https://doi.org/10.1093/poq/nfn063

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*, *5*(3), 213–236. https://doi.org/10.1002/acp.2350050305

Krosnick, J. A., & Alwin, D. F. (1987). An evaluation of a cognitive theory of response-order effects in survey measurement. *Public Opinion Quarterly*, *51*(2), 201–219. https://www.jstor.org/stable/2748993

Krysan, M., & Couper, M. P. (2003). Race in the live and the virtual interview: Racial deference, social desirability, and activation effects in attitude surveys. *Social Psychology Quarterly*, *66*(4), 364. https://doi.org/10.2307/1519835

Lind, L. H., Schober, M. F., Conrad, F. G., & Reichert, H. (2013). Why do survey respondents disclose more when computers ask the questions? *Public Opinion Quarterly*, *77*(4), 888–935. https://doi.org/10.1093/poq/nft038

Liu, M., & Wang, Y. (2015). Data collection mode effect on feeling thermometer questions: A comparison of face-to-face and Web surveys. *Computers in Human Behavior*, *48*, 212–218. https://doi.org/https://doi.org/10.1016/j.chb.2015.01.057

Mavletova, A., Couper, M. P., & Lebedev, D. (2018). Grid and item-by-item formats in PC and mobile web surveys. *Social Science Computer Review*, *36*(6), 647–668. https://doi.org/10.1177/0894439317735307

Miller, R. J., & Gibson, A. M. (2004). Supervision by videoconference with rural probationary psychologists. *International Journal of Innovation in Science and Mathematics Education*, *11*(1).

Office_of_Management_and_Budget. (2006). *Office of Management and Budget (OMB) Policy on Surveys*. https://www2.usgs.gov/customer/page9.html

Pew Research Center, September 2021, "The Internet and the Pandemic." https://www.pewresearch.org/internet/2021/09/01/the-internet-and-the-pandemic/

Reeves, B., & Nass, C. I. (1996). *The media equation: How people treat computers, television, and new media like real people and places*. Center for the Study of Language & Information; Cambridge University Press.

Roberts, C., Gilbert, E., Allum, N., & Eisner, L. (2019). Satisficing in surveys: A systematic review of the literature. *Public Opinion Quarterly*, *83*(3), 598–626. https://doi.org/10.1093/poq/nfz035

Roberts, F., & Francis, A. L. (2013). Identifying a temporal threshold of tolerance for silent gaps after requests. *The Journal of the Acoustical Society of America*, *133*(6), EL471–EL477. https://doi.org/10.1121/1.4802900

Schaeffer, N. C. (2000). Asking questions about threatening topics: A selective overview. In A. A. Stone, J. S. Turkkan, C. A. Bachrach, J. B. Jobe, H. S. Kurtzman, & V. S. Cain (Eds.), *The science of self-report: Implications for research and practice* (pp. 105–121). Lawrence Erlbaum Associates.

Schober, M. F. (2018). The future of face-to-face interviewing. *Quality Assurance in Education*, *26*(2), 293–302. https://doi.org/10.1108/QAE-06-2017-0033

Schober, M. F., Conrad, F. G., Hupp, A. L., Larsen, K. M., Ong, A. R., & West, B. T. (2020). Design considerations for live video survey interviews. *Survey Practice*, *13*(1). https://doi.org/10.29115/SP-2020-0014

Schober, M. F., & Glick, P. J. (2011). Self-deceptive speech: A psycholinguistic view. In C. Piers (Ed.), *Personality and psychopathology: Critical dialogues with David Shapiro* (pp. 183–200). Springer. https://doi.org/10.1007/978-1-4419-6214-0_8

Schober, M. F., Okon, S., Conrad, F. G., Hupp, A. L., Ong, A. R., & Larsen, K. M. (n.d.). Predictors of willingness to participate in survey interviews conducted by live video. *Technology, Mind, and Behavior.*

Simon, H. (1956). Rational choice and the structure of the environment. *Psychological Review*, *63*(2), 129–38. https://doi.org/10.1037/h0042769

Sun, H., Conrad, F. G., & Kreuter, F. (2020). The relationship between interviewer-respondent rapport and data quality. *Journal of Survey Statistics and Methodology*, 1–20. https://doi.org/10.1093/jssam/smz043

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response*. Cambridge University Press. https://doi.org/10.1017/CBO9780511819322

Tourangeau, R., & Smith, T. W. (1996). Asking sensitive questions: The impact of data collection mode, question format, and question context. *Public Opinion Quarterly*, *60*(2), 275–304. https://doi.org/10.1086/297751

Tourangeau, R., & Yan, T. (2007). Sensitive questions in surveys. *Psychological Bulletin*, *133*(5), 859–883. https://doi.org/10.1037/0033-2909.133.5.859

Turner, C. F., Ku, L., Rogers, S. M., Lindberg, L. D., Pleck, J. H., & Sonenstein, F. L. (1998). Adolescent sexual behavior, drug use, and violence: Increased reporting with computer survey technology. *Science*, *280*(5365), 867–873. https://doi.org/10.1126/science.280.5365.867

West, B. T., & Blom, A. G. (2017). Explaining interviewer effects: A research synthesis. *Journal of Survey Statistics and Methodology*, *5*(2), 175–211. https://doi.org/10.1093/jssam/smw024

West, B. T., Ong, A. R., Conrad, F. G., Schober, M. F., Larsen, K. M., & Hupp, A. L. (2022). Interviewer effects in live video and prerecorded video interviewing. *Journal of Survey Statistics and Methodology*, *10*(2), 317–336. https://doi.org/10.1093/jssam/smab040

Yan, T., Fricker, S., & Tsai, S. (2020). Response burden: What is it and what predicts it? In P. Beatty, D. Collins, L. Kaye, J. L. Padilla, G. Willis, & A. Wilmot (Eds.), *Advances in questionnaire design, development, evaluation and resting* (pp. 193–212). John Wiley & Sons, Inc. https://doi.org/10.1002/9781119263685.ch8

methods, data, analyses | Vol. 17(2), 2023, pp. 171-206 DOI: 10.12758/mda.2023.04

Comparing and Improving the Accuracy of Nonprobability Samples: Profiling Australian Surveys

Sebastian Kocar[1] *& Bernard Baffour*[2]

[1] *Institute for Social Change, University of Tasmania*

[2] *School of Demography, The Australian National University*

Abstract

There has been a great deal of debate in the survey research community about the accuracy of nonprobability sample surveys. This work aims to provide empirical evidence about the accuracy of nonprobability samples and to investigate the performance of a range of post-survey adjustment approaches (calibration or matching methods) to reduce bias, and lead to enhanced inference. We use data from five nonprobability online panel surveys and compare their accuracy (pre- and post-survey adjustment) to four probability surveys, including data from a probability online panel. This article adds value to the existing research by assessing methods for causal inference not previously applied for this purpose and demonstrates the value of various types of covariates in mitigation of bias in nonprobability online panels. Investigating different post-survey adjustment scenarios based on the availability of auxiliary data, we demonstrated how carefully designed post-survey adjustment can reduce some bias in survey research using nonprobability samples. The results show that the quality of post-survey adjustments is, first and foremost, dependent on the availability of relevant high-quality covariates which come from a representative large-scale probability-based survey data and match those in nonprobability data. Second, we found little difference in the efficiency of different post-survey adjustment methods, and inconsistent evidence on the suitability of 'webographics' and other internet-associated covariates for mitigating bias in nonprobability samples.

Keywords: nonprobability sampling, volunteer online panels, post-survey adjustment, calibration, matching methods, benchmarking

It has become increasingly evident that traditional surveys face challenges in measuring and understanding emerging and complex social issues, since they often fail to accurately measure individual behavior, attitudes and perceptions on various issues (Baker et al. 2010; Malhotra & Krosnick 2007; Tourangeau et al. 2014). Recent notable failures of polls to predict the outcomes of referenda and elections have shown that the way in which data are collected from the population must be responsive to people's dynamic lifestyles, choices, and attitudes (e.g., Goot 2021; Kennedy et al. 2018; Wang et al. 2015). Further, the widespread availability of and access to the internet and social media leads to a quick diffusion of ideas that may rapidly shift social attitudes and behaviors (e.g., Wang et al. 2021).

Compared to traditional (probability-based) survey methods which usually include offline data collection (mail, telephone, face to face (f2f)) and have been proven to be inadequate to capturing new to emerge, quick to change events, web-based surveys are advantageous given their convenience, quick turn-around times, and relatively low respondent costs (Baker et al. 2013). Additionally, nonprobability online panel surveys allow tests for consistency and reliability to be performed in a timelier manner than telephone and interviewer administered surveys. While there are web-based surveys that are probability-based (for instance push-to-web surveys with a 'population' frame of emails (Cornesse et al. 2020)), the majority of online surveys rely on being quick and efficient through reaching potentially millions of internet users which comes at the expense of being representative of the population (Bethlehem & Biffignandi 2012; Baker et al. 2013). We focus on nonprobability online (web-based) panel[1] surveys in this research, although the findings can be applied to other types of nonprobability surveys.

There are four main issues associated with nonprobability online panel surveys, which are related to type of sampling, sampling frame, nonresponse, and coverage. First, respondents are not selected based on probability sampling. Even though they may be 'randomly' selected, it is often not possible to work out their chance of being selected into the survey. Consequently, it is unknown what respondents with a non-zero chance of being selected comprise the population that the sample is selected from, and so the reliability of those sample survey estimates can-

[1] Nonprobability online panels are also known as volunteer, opt-in or access online panels.

Acknowledgements

We would like to acknowledge the contributions of the Social Research Centre for collecting and providing access to their survey data (Online Panels Benchmarking Study and Life in Australia™) and the Australian Government Department of Education, Skills and Employment for PhD program funding.

Direct correspondence to

Sebastian Kocar, Institute for Social Change, University of Tasmania
E-mail: sebastian.kocar@utas.edu.au

not be assessed with confidence (Callegaro & DiSogra 2008b). This is associated with the second issues, which is that there is no general population online sampling frame for the internet (e.g., Couper 2000). While virtually all internet users have an email address, there is no list comprising all of these email addresses that could be used to draw a random sample. Third, online survey respondents have been found to have different characteristics and behaviors to respondents from more traditional surveys. Online surveys generally have higher levels of item and unit nonresponse (e.g., Couper 2000; Daikeler et al. 2020), which has potential to introduce more nonresponse bias. Four, the internet does not have universal coverage: in Australia in early-2022, it was estimated that 9% of people did not use the internet (DataReportal 2022)[2]. This can introduce (under)coverage error, since this lack of access is concentrated amongst those of older age, rural location, Indigenous ethnicity, and lower education levels. Those are groups which are increasingly important to policymakers, hence limiting the utility of internet-based surveys. Collectively, these limitations mean that the data collected online with nonprobability panels are less reliable than those gathered by traditional survey methods, since they are generally more prone to the above-mentioned sampling, nonresponse, and undercoverage bias (which is, at the same time, challenging to estimate). Hence, the existing evidence suggests that we cannot be confident that the results from nonprobability panels accurately represent trends in the general population.

Our research aims are three-fold. First, we quantify the differences in survey estimates obtained from the same survey administered through a probabilistic sampling framework in contrast with those collected from a non-probabilistic framework. Second, we compare and contrast the performance of different post-survey adjustment methods on reducing bias in nonprobability-based online panel surveys. Third, we compare how the inclusion of different external data sources (such as Census) and covariates (such as non-demographics[3]) in post-survey adjustment affect the accuracy of survey estimates. This investigation adds value to the existing literature on approaches to mitigate bias in nonprobability surveys. As such, it provides valuable evidence to survey practitioners using samples from nonprobability online panels of better quality (e.g., those with ESOMAR or ISO accreditation), as well as survey researchers interested in implementing other types of nonprobability surveys.

[2] The last official statistics estimate for Australian households with no internet access at home was from 2016-17, i.e., 14.0%. The same estimate for households without children under 15 was even higher, i.e., 18.1% (Australian Bureau of Statistics 2018a).

[3] We define non-demographics as attitudinal, behavioral, knowledge and factual questions that do not ask about person's socio-demographic characteristics (see Yeager et al. 2011).

Background and Literature Review

With probability sampling we ensure that every unit in the population has a known, and non-zero, chance of being selected into the sample. This randomisation is a key design attribute of probability sampling, and enables the calculation of standard errors, confidence intervals, and making generalized inferences regarding the target population of interest from the sample (Hade & Lemeshow 2011). However, while most (probability) surveys have known selection probabilities, whether people respond cannot be controlled for, in spite of all the best efforts of survey practitioners. Rivers (2013) argues that it is the probability of sample inclusion not selection that matters, since whether people cooperate in probability surveys cannot be controlled for, and low response rates introduce skews similar to those in volunteer panels. Trends of high nonresponse rates with a large proportion of probability-based surveys reporting response rates of under 10% (Kennedy & Hartig 2019), and the associated nonresponse biases may lead to flawed results and problems in statistical inference (Baker et al. 2010; Baker et al. 2013). However, the fact that the selection probabilities for a sample are unknown does not imply that they cannot be estimated or adjusted for in a nonprobability sample, just as adjustments are used in probability-based surveys to compensate for issues around coverage and response (Rivers 2013).

Opportunities to Improve Accuracy of Nonprobability Samples

There is a whole gamut of online nonprobability-based surveys, from the opt-in click-through unsolicited surveys which are advertised on websites, to more structured recruitment of a panel of respondents; as a result of these idiosyncratic designs which make it difficult to work out the rates of contact, response, and (non)coverage, it is almost impossible to make reasonable statistical inferences from data obtained with nonprobability-based surveys (Rivers 2013). However, the characteristics of the nonprobability online panel sample may closely resemble the population being studied and identifying the conditions under which valid statistical inferences can be made using the realized sample is important (Mercer et al. 2017). This selection bias - which leads to the sample misrepresenting the population - can be controlled for using several different approaches, underpinned by an existing framework based on causal inference used in numerous fields such as epidemiology, political science and economics (Heckman 1979; Hug 2003; Rothman et al. 2008).

Valliant (2020) and Elliott and Valliant (2017) showed that it is not necessary and sufficient that (i) every unit in the population has some probability of being included in the sample, and that (ii) there is a structural model based on the observed sample which can be used to describe the variables we are interested in measuring, meaning that you do not need both conditions to hold. This implies that reweighting or matching schemes can be used to (a) estimate the probability of response and (b) calibrate to known benchmark population totals, to correct for any selection biases in the estimates derived from nonprobability sample surveys (Matei 2018). We distinguish between matching approaches and reweighting approaches, and the key goal of both approaches is to ensure that there is no (or little) bias in the observed data, meaning that the empirical distribution of the observed data is similar to the population (Baker et al. 2013; Elliott & Valliant 2017; Mercer et al. 2017; Mercer et al. 2018; Valliant 2020).

Post-Survey Adjustments in Nonprobability Samples

Post-survey adjustments correct for the unequal probabilities of selection and are common in both nonprobability and most probability surveys: virtually no probability sample uses simple random sampling. As such, in both probability and nonprobability samples, the objective for inference is to ensure that the composition of the sampled units with respect to the observed characteristics either matches or can be adjusted to match the population of interest. Post-survey adjustments have the dual purpose of reducing the bias and producing more accurate population estimates (Elliott & Valliant 2017; Mercer et al. 2017).

There are several approaches which have been proposed to improve accuracy and inference for data collected under a nonprobability sample. These approaches are predicated from the issues facing probability samples caused by differences in response and coverage of surveys. To cope with these issues, statistical adjustments typically correct for any systematic biases, including in nonprobability samples (Cornesse et al. 2020; Elliott 2009; Lehdonvirta et al. 2021; Rivers 2007).

This study compares six primary methods of reweighting and matching survey data: raking, generalized regression estimation (GREG), propensity score weighting (PSW), multilevel regression and poststratification (MRP), Mahalanobis distance matching (MDM) and coarsened exact matching (CEM). Reweighting methods directly adjust the sample distribution to the target population distribution, to achieve the desired sample composition in the presence of nonresponse and/or other factors. Matching methods attempt to create a balanced nonprobability sample which closely resembles the characteristics of a probability sample from the 'true' population (when compared with a selected array of auxiliary, often non-demographic, characteristics) (Bethlehem 2016; Cornesse et al. 2020). Assessing performance of different post-survey adjustment methodology is important as all methods come with certain limitations – for example, raking was reported to be less effective to mitigate bias in nonprobability online panel samples than in probability samples (Mercer et al. 2018), the GREG estimator becomes less precise the larger the number of benchmarks (Deville et al. 1993), MRP requires knowledge of the joint distribution of the poststratification variables in the target population (Deville & Särndal 1992), and matching methods cannot be used with all types of data.

In the next paragraphs, we provide more information about each of the post-survey adjustment methods investigated in this study.

Raking

Raking, also known as iterative proportional fitting, is the most common weighting method and is simple to implement as it relies on knowing the marginal distribution of population covariates. As part of the procedure, the weights for each individual are repeatedly adjusted until the sample distribution is perfectly aligned with the population distribution for the selected set of variables. As the utility of a large set of weighting covariates diminishes, using key socio-demographic variables is often sufficient to reduce the selection bias in probability samples (Kalton & Flores-Cervantes 2003).

Generalized Regression Estimation (GREG)

Generalized regression estimation (GREG) is a calibration[4] approach where the sampling weights are adjusted to make certain the survey estimators match to the set of known population totals (benchmarks). In contrast to raking which repeatedly reweights the sample to the marginal distributions of the known population totals, the GREG estimator is based on the minimizing the distance measure between the sample and the benchmark information and it is supposedly more efficient and provides more accurate population estimates (Deville & Särndal 1992).

Propensity Score Weighting (PSW)

In the simplest version of probability-based sampling, survey respondents are assumed to have a non-zero chance of being included in the sample and weighting each sample individual by the inverse of its sample selection probability removes any selection bias (Cochran 1977). When data are collected through a nonprobability-based sample, we can use the same ideas, and although selection probabilities from a nonprobability sample are unknown, it does not mean that they cannot be estimated (Rivers 2013). In PSW, a synthetic population assumed to "represent" the full target population is created by using external high-quality data representative of the population. Then pseudo-inclusion probabilities are estimated using binary (i.e., probit or logistic) regression modeling, which leads to a probability-based (or

[4] Calibration is a general framework for weighting in which the following conditions for adjustment weights have to be satisfied: (1) the weights have to be as close to 1 as possible, (2) after calibration, the sample distribution of the auxiliary variables should match the population distribution (Bethlehem 2008). Deville et al. (1993) distinguish between complete post-stratification, generalized raking, and GREG as calibration methods.

synthetic) reference sample which is combined with the nonprobability sample (Schonlau & Couper 2017; Valliant 2020). Like in calibration, PSW is efficient in bias reduction if the weighting variables and the propensity of response in the nonprobability sample are (strongly) associated with outcome variables (Rosenbaum & Rubin 1983; Valliant & Dever 2011).

Multilevel Regression and Poststratification (MRP)

The MRP approach (Gelman 2007; Gelman & Little 1997) is based on assuming the existence of a super-population model which can be fitted to the analytic survey variables and can be used to project the observed sample to the full population. The key assumption here is that sampled and non-sampled data are driven by an underlying model (for the analysis variables) and this model can be revealed by analyzing the sample responses. In the presence of nonresponse, this model also specifies the relationship between the observed units and the unobserved data (Brick 2013). Poststratification, which includes creating a set of post-strata and estimating the mean value by fitting mixed effects (multilevel) model in the case of MRP, requires knowledge of the joint distribution of the poststratification variables in the target population unlike other reweighting methods (Deville & Särndal 1992), except for interactions between covariates. In political science, this approach is useful in obtaining state-level predictions based on relatively small national samples (for example, Bon et al. 2019; Park et al. 2004; Park et al. 2006; Wang et al. 2015).

Mahalanobis Distance Matching (MDM)

MDM is a distance matching method which creates groups containing one or more observations from both the reference sample and the nonprobability sample that are similar on a set of auxiliary variables believed to be associated with the probability of selection. In MDM, we measure the distance between a pair of observations, y_i and y_j, with the Mahalonobis distance calculated as presented in Equation 1:

$$M(y_i, y_j) = \sqrt{(y_i - y_j)^T S^{-1} (y_i - y_j)} \tag{1}$$

where S is the sample covariance matrix of y. Two observations are matched if they have the minimum distance out of a set of pairs, e.g., through nearest neighbour matching. Since the population of possible match-pairs exponentially increases as the nonprobability sample size increases, usually some procedure is used to remove pairs that are unreasonably distant through defining calipers which are chosen cutoffs for which the maximum distance is allowed (Stuart & Rubin 2008).

Coarsened Exact Matching (CEM)

Coarsened exact matching (CEM) is a matching method like the MDM, but the key difference is that it is a stratification-based method (Sizemore & Alkurdi 2019), and calipers are not required to remove unreasonably bad matches (Iacus et al. 2011). In CEM, units with the same values of the selected covariates (in contract to exact matching, they can be coarsened, i.e., recategorized into fewer groups) are placed in a single stratum. Within each stratum, the units in the nonprobability sample are weighted to be equal to the number of units in the reference sample. Strata without at least a single nonprobability sample or reference sample unit, are given a zero weight which effectively prunes them from the dataset. By removing unmatched units, the inference is generally improved because it achieves a better balance between the empirical distributions of reference sample and the nonprobability sample (Iacus et al. 2009; Stuart 2010).

Scope of this Study

Following from Mercer et al. (2017), we use the general framework which emphasizes the characteristics of the realized sample (regardless of how it was generated), and therefore correct for any self-selection bias in survey inference (Groves 2006; Keiding & Louis 2016; Little & Rubin 2002). The authors identify three components that determine whether the presence of self-selection ultimately leads to biased survey estimates: exchangeability, positivity, and composition (Mercer et al. 2017). These components of self-selection bias are not fundamentally different for nonprobability samples, but what differs between probability and nonprobability samples are the underlying assumptions which lead to individuals becoming members of nonprobability samples (Kennedy et al. 2016; MacInnis et al. 2018; Pfeffermann et al. 2015).

Notwithstanding, this can be useful in investigating if there is (a) improved inference of sample data from a nonprobability survey, and (b) through comparing different post-survey adjustment methods under different external data sources scenarios we can ascertain their suitability/performance under various conditions. There have been a number of authors – for instance, DiSogra et al. (2011), Baker et al. (2013), Mercer et al. (2017), Mercer et al. (2018), and Valliant (2020) – who have undertaken similar research into the performance of different methods, and also discussed the requirements with respect to the external data sources for the various approaches.

Therefore, we will examine a range of survey estimates against two categories of population benchmarks: secondary demographics (such as citizenship and employment status), and non-demographics (such as alcohol consumption and life satisfaction), as well as against both categories combined. First, we compare the accuracy of probability and nonprobability samples from two Australian survey projects, by presenting updated evidence. Second, we investigate the performance of different post-survey adjustments to improve accuracy of nonprobability samples. We do that under four realistic scenarios which differ in terms of the nature of the auxiliary data that is available for use in post-survey adjustment for nonprobability surveys. The scenarios under which we are assessing performance of adjustment methods are the following:

- *Scenario 1 – availability of census aggregated statistics utilized to improve accuracy in nonprobability samples*
  Under this scenario, aggregated[5] population census data matching to *primary demographics*[6] from a nonprobability sample are used to adjust the sample distribution for those key auxiliary variables to match the population distribution (e.g., for *sex, age*, and *education*).
- *Scenario 2 – availability of additional census aggregated statistics utilized to improve accuracy in nonprobability samples*
  Under this scenario, aggregated population census data matching to *primary* and, additionally, *secondary demographics*[7] from a nonprobability sample are used to adjust the sample distribution for those selected auxiliary variables to match the population distribution (e.g., besides for *sex*, *age*, and *education*, *employment status* covariate can be included in the post-survey adjustment).
- *Scenario 3 – availability of census aggregated statistics and a representative source of non-demographic benchmarks (i.e., a large national survey) utilized to improve accuracy in nonprobability samples*
  Under this scenario, besides the aggregated population census data from Scenario 1, we can use *secondary demographics* and *non-demographics* from a large probability-based national survey (e.g., *household composition* and *health status* from a government survey on health) that are matching to those covariates in the nonprobability sample. This time, microdata[8] are a source of *secondary demographics* and *non-demographics*.

[5] Aggregated or tabular data are produced by grouping information into categories. Within these categories, values are combined (e.g., a count of respondents of particular age). They are also known as macrodata (Australian Bureau of Statistics n.d.-b).

[6] Primary demographics as defined by Pennay et al. (2018) are socio-demographic variables which were used in post-stratification weighting.

[7] In contrast to primary demographics, secondary demographics as defined and used by Pennay et al. (2018) were additional socio-demographic variables which were not included in post-stratification weighting but rather in accuracy calculations only (such as Indigenous status or voluntary work).

[8] Microdata, also known as unit record files, are a type of data including unit records containing detailed information about analytical units such as persons or organizations. They often include individual responses to survey questions or from administrative forms (Australian Bureau of Statistics n.d.-b).

- *Scenario 4 – availability of census aggregated statistics, and a smaller scale probability-based survey data utilized to improve accuracy in nonprobability samples*
  Under this scenario, besides the aggregated population census data from Scenario 1, we can use *non-demographics* from a smaller-scale non-government survey that are matching to selected covariates in the nonprobability sample. While we apply a less representative external data source to improve accuracy, there are additional non-demographic covariates which could be used to balance the samples as noted in the literature. An example of those non-demographics is 'webographic' variables, which are available in a microdata form. Webographic variables are attitudinal or lifestyle variables accounting the difference between web survey participants and those who do not do surveys online (Baker et al. 2013). Different authors considered different questions as 'webographic' questions, such as: feeling alone, eagerness to learn new things, willingness to take chances, lifestyle questions (on travelling, participation in sports, reading a book), opinions on what is a violation of privacy, knowing a 'lesbian, gay, bisexual, transgender, and queer or questioning' (LGBTQ) person (Schonlau et al. 2007), early-adopter items (DiSogra et al. 2011; Dutwin & Buskirk 2017) or media use (Baker et al. 2013). On the other hand, Mercer et al. (2018) used political attitude variables in post-survey adjustments. In our study, besides early-adopter items, we also consider internet connection, access and use, and number of surveys completed as 'webographic' variables or, simpler, 'webographics' (see Table 10 in the Appendix for more information).

The difference between Scenarios 3 and 4 is the type and the source of auxiliary survey data available for post-survey adjustment. Under Scenario 3, we have access to a large-scale nationally representative survey (large sample, e.g., 20,000+, with higher accuracy), such as the National Drug Strategy Household Survey. Under Scenario 4, we can use a smaller probability-based sample (e.g., about n=600), but with an ability to collect tailor-made data including key covariates which could help mitigate bias after matching or propensity scoring weighting (e.g., 'webographics'); data collectors attempting to improve the accuracy of their nonprobability samples could conduct a smaller-scale probability-based survey, e.g., a probability-based sample from Online Panels Benchmarking Study, to improve inference in opt-in panel samples.

This study will address the following research question: *How accurate are nonprobability online samples in comparison to probability samples and to what extent can inference be improved by using post-survey adjustment methods under different scenarios?*

Methods

Data

Original Online Panel Benchmarking Study (2015 OPBS)

The 2015 Online Panels Benchmarking Study (OPBS, Pennay et al. 2016[9]) was conducted in June 2015 and administered the same questionnaire to eight samples, made up of three probability samples and five nonprobability online panel samples. Each sample aimed to achieve approximately six hundred completed interviews; in the end, the smallest sample comprised of 538 respondents (Pennay et al. 2018), as presented in Table 1. The design was similar to the US study by Yeager et al. (2011) which compared the accuracy of seven online samples and two probability samples. The main objective of OPBS was to inform the debate in Australia on the issues pertaining to inference from nonprobability online panel surveys.

Life in Australia™ – Probability-Based Online Panel: OPBS Replication (2017 OPBS)

Life in Australia™ is a probability-based internet panel for the Australian general adult population, and in January-February 2017, all active Life in Australia™ panellists were asked to participate in the replication of the OPBS. Social Research Centre administered the same questionnaire used for the original 2015 OPBS to determine the accuracy of their probability-based online panel (Kaczmirek et al. 2019). This was the second wave of Life in Australia, referred to as the Online Panel Benchmarking Study Replication or 2017 OPBS (Pennay & Neiger 2020[10]).

Life in Australia™ panellists were recruited in 2016 via their landline or mobile phones to take part in incentivized monthly surveys, and the final sample of registered panellists was 3,322 individuals (overall recruitment rate, AAPOR RR3: 15.5%). Since the recruitment of panellists was through probability-based dual-frame sampling, the results from the surveys are generalizable to the Australian population. Life in Australia™ is a mixed-mode probability online panel, and to take into account the population with no access to the internet, the study also contacted panel members who happened to be offline via phone (representing 13.6% of Wave 2 sample) (Kaczmirek et al. 2019).

Population, Sampling and Samples

Both the 2015 and 2017 OPBS surveys collected information from an in-scope population of all Australians aged 18 years and over. The studies were carefully designed to assess accuracy of nonprobability online panel samples relative to prob-

[9] Data DOI: 10.4225/87/FSOYQI

[10] Data DOI: 10.26193/YF8AF1

ability-based surveys using different probabilistic sampling methodology through applying the same data collection instrument to provide data on the demographic, social characteristics and wellbeing of people in Australia (Kaczmirek et al. 2019; Pennay et al. 2018).

As previously explained, the OPBS 2015 study data comprised of eight samples, three of which were probability-based samples: (i) an address-based sampling (A-BS) survey with Geocoded National Address File (G-NAF) as a sampling frame (survey mode: hard copy/mail, online, telephone), (ii) a standalone dual-frame Random Digit Dialing (RDD) survey sample (survey mode: telephone), and (iii) a RDD end-of-survey recruitment sample (survey mode: telephone, online, hard/copy) (Pennay et al. 2018), also known as 'piggybacking' survey sample (Tourangeau & Smith 1985). For the purpose of the 2015 OPBS study, five Australian nonprobability online panels collected data from about 600 of their panellists each. Four of five nonprobability online panels complied with all ESOMAR's questions to help online research buyers[11] and three of the five were with ISO 26362 accreditation[12] (Pennay et al. 2018). We will analyze accuracy of the whole nonprobability sample combined[13] (n=3,058) and for two purposely selected nonprobability samples, the most and the least accurate.

The OPBS Replication 2017 survey comprised of one probability-based mixed-mode (online and telephone) sample. The cumulative response rate (CUMRR1), which is a product of overall recruitment (RECR x PROR) and survey completion rates (COMR)[14], was 12.2% (AAPOR RR3). A total of 2,580 Life in Australia™ panellists completed Wave 2 questionnaire (Kaczmirek et al. 2019).

[11] ESOMAR's *Questions to help buyers of online samples* include questions on company profile (such as *What experience does your company have in providing online samples for market research?*), sample sources and recruitment (such as *Is the recruitment process 'open to all' or by invitation only?*), sampling and project management (such as *Do you employ a survey router or any yield management techniques?*), data quality and validation (such as *How often can the same individual participate in a survey?*), policies and compliance (such as *How can participants provide, manage and revise consent for the processing of their personal data?*) and metrics (*Which of the following [metrics] are you able to provide to buyers, in aggregate and by country and source?*). For more information, see ESOMAR (2021).

[12] ISO 26362:2009 developed criteria and specified terms, definitions and service requirements for organisations managing online panels, including on sampling, fieldwork, and data management. It has since been revised by ISO 20259:2019 standard (International Organisation for Standardisation 2022).

[13] Combining data from several volunteer panels can increase their overall accuracy (Cornesse et al. 2020), can be thus considered a solution to mitigate representation bias in nonprobability surveys, and is as such a subject of this study. We were particularly interested in the effectiveness of post-survey adjustment on combined data from different nonprobability sources, in comparison to individual volunteer panel samples.

[14] Recruitment rate, completion rate, and cumulative response rate were introduced by Callegaro and DiSogra (2008a) for calculation of response rates in online panels.

Table 1 Studies and subsamples analyzed

Study	Subsample	Response rate (AAPOR RR3)	n[a]
Online Panels Benchmarking Study (2015 OPBS)	Address-based sampling	26.2%	538
	Standalone RDD (dual-frame)	14.7%	600
	RDD "piggybacking" (dual-frame)	9.8%	560
	5 volunteer panel samples[b]	2.6%-15.4%[c]	3,058
Online Panels Benchmarking Study Replication (2017 OPBS)	Life in Australia™ Wave 2	recruitment rate: 15.5%, Wave 2 survey completion rate: 78.6%, cumulative response rate: 12.2%	2,580

[a] We have to acknowledge the fact that with relatively small sample (n=about 600), sampling variance as a component of sampling error is larger. In practice this means that estimates from surveys with smaller samples can be less accurate in benchmarking studies by chance in comparison to those from larger surveys.
[b] Besides the combined nonprobability sample, we will analyze data separately for the most accurate panel (Panel 3, n=601) and the least accurate panel (Panel 1, n=601) (based on the results from Kaczmirek et al. 2019, p. 25). We will not analyze data for all 5 nonprobability panels separately due to space constraints. However, through comparing the best and worst performing nonprobability panel, we can get an indication of the variation in the bias and accuracy of different panel providers.
[c] For nonprobability samples, response rates cannot be calculated and some authors (e.g., Pennay et al. 2018) report sample yields instead.

Generally speaking, there were notable differences in response between the subsamples listed in Table 1, which might result in different levels of nonresponse error. The hope is that we can mitigate against this in our analysis through effective post-survey adjustment procedures applied to nonprobability data.

Benchmarks

Assessing quality in surveys requires an objective standard to which the survey estimates can be compared, such as population benchmarks. Differences between estimates from survey response and population benchmarks can occur through bias or variance, where the bias term captures the systematic (selection) errors that are shared by nonprobability samples. The variance term captures the sampling variation and accounts for the variation due to the differences in survey protocols, statistical modeling or weighting adjustments.

To replicate benchmarking analysis from Pennay et al. (2018)[15] and Kaczmirek et al. (2019), we use the same benchmarks but from updated data sources collected closer in time to 2015 OPBS and 2017 OPBS studies. We primarily use information from the Australian quinquennial Census (Australian Bureau of Statistics 2016) as benchmarks since censuses offer universal coverage of the population by definition. For some instances we use administrative record data and information drawn from large government surveys as benchmarks. Those are electoral registration information from the Australian Electoral Commission, and social and health characteristics from the government funded surveys which are considered as the best quality sources of nationally representative benchmarks in Australia with the highest validity (e.g., Australian Bureau of Statistics 2018b).

Benchmarks will be divided into primary (for post-survey adjustment only), secondary demographics, and substantive items (see Pennay et al. 2018). Table 2 provides a description of the benchmarks used in the study.

[15] The findings presented in Pennay et al. (2018) were further explored and published by Lavrakas et al. (2022).

Table 2 Benchmarking data sources and nationally representative benchmarks

Study	Data collection mode	Sample size	Benchmarks *(modal response category in brackets)* [a]*primary demographics,* [b]*secondary demographics,* [c]*substantive items/non-demographics*
Australian Census 2016 (Australian Bureau of Statistics 2016)	self-administered online, F2F	n=23,401,892 persons	Age, in categories [a] Gender[a] State[a] Residence in state capital city[a] Country of birth[a] Australian citizenship[b] *(Australian citizen)* Employment status[b] *(currently employed)* Home ownership[b] *(with a mortgage)* Indigenous status[b] *(not Indigenous)* Language other than English[b] *(speak only English)* Living at last address 5 years ago[b] Most disadvantaged quintile for area-based socio-economic score[b] Resident of a major city[b] Voluntary work[b]
National Drug Strategy Household Survey (NDSHS) 2016 (Hewitt 2017)	self-administered paper-based or online, CATI	n=23,749 persons	Household status[b] *(couple with dependent children)* Smoking status[c] *(daily smoker)* Alcoholic drink of any kind in the past 12 months[c] (y*es, consumed alcohol)*
National Health Survey 2014-15	F2F	n=19,259 persons	Psychological distress (Kessler 6)[c] *(low distress)* General health[c] *(very good)* Private health insurance[c] *(yes, has insurance)* Wage and salary income[b] *(income $1000-1249 pw)*
General Social Survey 2014	F2F	n= 12,932 persons	Life satisfaction[c] *(8 out of 10)*
Australian Electoral Commission (2015)	administrative data	n=16,405,465 persons	Enrolled to vote[b] *(yes, enrolled)*

F2F – face-to-face; CATI – Computer-assisted telephone interviewing

Data Analysis

Benchmarking Analysis

To carry out our benchmark analysis, we need to balance against variance and bias in the final estimates. There are a wide variety of measures estimating the bias, such as the number of statistically significant differences from the benchmarks, the average absolute error (AAE) (including measures of uncertainty of the AAE, such as the standard deviation of the AAE or the range and ranking) (see Dutwin & Buskirk 2017; MacInnis et al. 2018; Yeager et al. 2011). To provide a measure of the variance, we compute the mean squared error which is a function of both the bias and the variance, and as such it is a good measure of the overall accuracy of the different approaches; it is usual practice to take the square root of the mean square error (RMSE) which is more sensitive to large errors than AAE. The aim of the study is to find the approach which is robust under the different scenarios. As such we present results using the AAE and RMSE to give an absolute measure of the error and the variability measure of the error, respectively.[16]

The AAE was used by Yeager et al. (2011) to compare impact of different weighting approaches for probability and nonprobability surveys in the US. The same measure was used by Pennay et al. (2018) and Kaczmirek et al. (2019), who replicated the study design in Yeager et al. (2011) for Australia.

Our study follows all three of these previous studies, and the AAE is calculated as presented in Equation 2:

$$AAE = \sum_{j=1}^{k} \frac{|\widehat{y_j} - y_j|}{k} \tag{2}$$

where $\widehat{y_j}$ is the j-th estimate (of a survey item) and y_j is the value for a corresponding (population) benchmark. And similarly, the RMSE is computed as presented in Equation 3:

$$RMSE = \sqrt{\frac{\sum_{j=1}^{k} (\widehat{y_j} - y_j)^2}{k}} \tag{3}$$

where k is the number of benchmarks, $\widehat{y_j}$ is again the j-th estimate from either OPBS surveys, and y_j is the value for a corresponding benchmark. In our study, the estimates ($\widehat{y_j}$) represented proportion estimates for modal response for items with corresponding benchmarks; this is consistent with the approach from the Australian benchmarking studies (Kaczmirek et al. 2019; Pennay et al. 2018) and the US studies described in the literature (e.g., Yeager et al. 2011).

[16] When we computed Relative Absolute Bias (see Dutwin & Buskirk 2021) as a relative measure, we reached the same conclusions about the accuracy of probability and nonprobability samples as when computing AAE as an absolute measure.

To explore the generalizability of these findings, we calculate AAE and RMSE for 12 secondary demographics, 6 substantive items, and all 18 survey items with corresponding benchmarks combined. Most probability and nonprobability surveys apply adjustment for primary benchmarks as a standard approach, and for the majority of surveys the differences between the sample and population for primary benchmarks is expected to be minimal (Cornesse et al. 2020; Mercer et al. 2017). The analysis was facilitated by the statistical coding environment and language R (R Core Team 2020) to carry out all data processing, post-survey adjustments, imputation of missing values[17] and benchmarking analyses. Besides R base or stats packages, the following packages were used: *Hmisc* (for data processing, Harrell et al. 2020), *missForest* (for imputation of missing values, Stekhoven 2013), *fastDummies* (to create dummy variables for MDM, Kaplan 2020), *anesrake* (to perform raking, Pasek 2018), *sjstats* (for data processing, Lüdecke 2020), *questionr* (for data processing, Barnier et al. 2020), *MatchingFrontier* (to perform MDM, King et al. 2015), *cem* (to perform CEM, Iacus et al. 2020), and *rstanarm* (to conduct dominance analysis, Goodrich et al. 2020).

Post-Survey Adjustment Approaches and Parameters

Methods. To improve inference in nonprobability samples, we will test a number of post-survey adjustment methods and techniques:

- raking[18]
- generalized regression estimation (GREG)
- multilevel regression and poststratification (MRP)
- coarsened exact matching (CEM)
- Mahalanobis distance matching (MDM)
- propensity score weighting (PSW).

PSW, MDM and CEM selection/weighting will be later adjusted to match primary demographic benchmarks from Australian Census 2016. This means that those methods will be combined with raking not to introduce bias due to any socio-demographic sample imbalance after the initial adjustment. For more information about each of these methods, see Subsection 2.2, and for post-survey adjustment details from this study, see Table 3.

[17] We imputed missing values using random forest imputation algorithm, which is suitable for both continuous and categorical variables. Missing values were imputed for calibration and matching purposes only, and not for estimation, which means that only valid values of items with corresponding benchmarks were used in calculations of estimates.

[18] In probability samples, a two-stage process can be used for weighting, first calculating a design weight (for the unequal probability of sample members being selected) and second raking (to reduce possible nonresponse). As the same process cannot be used for weighting nonprobability samples, and as the findings on the accuracy of nonprobability samples would not change (see Kaczmirek et al. 2019), we used a consistent one-stage raking approach across all samples (and calibration methods).

Table 3 Post-survey methods, covariates, and parameters

Method	Scenario	Type of covariates	Covariate selection mechanism	Source of covariates	Other post-survey adjustment characteristics
Raking	Scenario 1	Primary demographics(1)	Weighting variables from the original benchmarking studies	Australian Census 2016	We applied weight trimming to ensure that the maximum weight after post-survey adjustment was 5.
	Scenario 2	Primary demographics(1) Secondary demographics	Weighting variables from the original benchmarking studies Covariates with the largest absolute error relative to Census benchmarks	Australian Census 2016	
	Scenario 3	Primary demographics(1) Non-demographics	Weighting variables from the original benchmarking studies All matching additional covariates from a large-scale survey	Australian Census 2016 NDSHS 2016	
GREG	Scenario 1	Primary demographics(1)	Weighting variables from the original benchmarking studies	Australian Census 2016	
	Scenario 2	Primary demographics(1) Secondary demographics	Weighting variables from the original benchmarking studies Covariates with the largest absolute error relative to Census benchmarks	Australian Census 2016	
	Scenario 3	Primary demographics(1) Secondary demographics & non-demographics	Weighting variables from the original benchmarking studies All matching additional covariates from a large-scale survey	Australian Census 2016 NDSHS 2016	
MRP	Scenario 1	Primary demographics(1)	Weighting variables from the original benchmarking studies	Australian Census 2016	

Method	Scenario	Type of covariates	Covariate selection mechanism	Source of covariates	Other post-survey adjustment characteristics
CEM	Scenario 3	Secondary & non-demographics	All matching covariates from a large-scale survey	NDSHS 2016	Pruning(3) of maximum 50% of all nonprobability sample units; adjusted to match primary demographic(1) benchmarks
	Scenario 4	Non-demographics, including 'webographics')	Selected with dominance analysis(2) out of all available matching covariates (excluding those with corresponding benchmarks)	OPBS 2017 Replication sample	
MDM	Scenario 3	Secondary & non-demographics	All matching covariates from a large-scale survey	NDSHS 2016	The same matched sample sizes as for CEM; adjusted to match primary demographic(1) benchmarks
	Scenario 4	Non-demographics, including 'webographics')	All available matching covariates (excluding those with corresponding benchmarks)	OPBS 2017 Replication sample	
PSW	Scenario 3	Secondary & non-demographics	All matching covariates from a large-scale survey	NDSHS 2016	Adjusted to match primary demographic(1) benchmarks (the whole approach could also be called “doubly-robust”)
	Scenario 4	Non-demographics, including 'webographics')	All available matching covariates (excluding those with corresponding benchmarks)	OPBS 2017 Replication sample	

(1) Primary demographics from Australian Census 2016 (and used in the original benchmarking studies by Pennay et al. 2018 and Kaczmirek et al. 2019) were gender, country of birth, and interaction effects between age and education, and state and capital city in state.
(2) Dominance analysis is used to compare the relative importance of predictors in regression models by comparing R2 or Pseudo R2 coefficient with different ranges of selected predictors (Budescu 1993). In practice, with dominance analysis we can select the covariates which distinguish probability and nonprobability samples the most in a multivariate setting (logistic regression was used in our case).
(3) Removing those units from nonprobability data which cannot be matched with any unit from a probability sample.

Based on the literature review from Subsection 2.2, we selected all post-survey adjustment methods appropriate for use with particular types of data. While raking, GREG and MRP can be used with tabular data and estimates from survey microdata (or both at the same time), weighting schemes should include estimates from nationally representative data sources producing known population totals (Kalton & Flores-Cervantes 2003, p. 82); hence, raking, GREG and MRP are not analyzed under Scenario 4, i.e., with a smaller scale probability survey data producing rough estimates of population totals. Also, a disadvantage of MRP is the requirement of the joint distribution of the poststratification variables, and CEM, MDM and PSW can only be used with microdata, i.e., under Scenarios 3 and 4.

Covariates. The theory explains that the selection of covariates for post-survey adjustment should be based on the relationship with nonresponse and non-coverage (e.g., Battaglia et al. 2009). Kalton and Flores-Cervantes (2003) pointed out that the precision of estimates can be increased by benchmarking to external sources with covariates that are closely related to key survey variables. The literature on post-survey adjustment in nonprobability samples (e.g., Dutwin & Buskirk 2017) suggests using covariates that are associated with participation in nonprobability samples/online panels, in attempt to primarily reduce errors associated with coverage, and adjust for inherent selection bias. We will follow these general recommendations/principles by:

- selecting secondary demographic covariates with the largest absolute error relative to Census benchmarks under Scenario 2 – our assumption is that those socio-demographic differences are directly associated with undercoverage (and nonresponse) bias in nonprobability online panels;
- selecting all matching health-related items (besides a secondary demographic item) to reduce error of other health-related items under Scenario 3 – if adjustment covariates are closely related to the target outcome variables, bias could be mitigated
- selecting non-demographic covariates which were previously discussed in the literature as effective in reducing coverage error in non-probability samples, so-called ‘webographic’ variables, under Scenario 4;
- identifying a limited number of ‘webographic’ covariates which distinguish nonprobability and probability samples the most, to be used with CEM under Scenario 4.

At the same time, validity of the sample has to be preserved by including core demographics like age and gender; in the case of calibration, we also have to have in mind that selecting too many covariates can lead to significant variance inflation and inability for raking algorithm to converge (Battaglia et al. 2009).

For details on the final selection of covariates, applied with different methods and under different scenarios, please see *Final selection of post-survey adjustment covariates* section and Table 10 in the Appendix.

Results

Accuracy of Nonprobability Online Panels

The results in this section provide updated evidence regarding the accuracy of nonprobability online samples in comparison to probability samples (with more recent benchmarks, for original results see Kaczmirek et al. 2019). We will use the identified gap in accuracy as a reference for assessment of effectiveness of post-survey adjustments (see Section 4.2).

Table 5 presents the results on the accuracy of OPBS 2015 and OPBS 2017 Replication surveys. The results confirm the findings from Pennay et al. (2018) and Kaczmirek et al. (2019) on the accuracy of nonprobability-based online panels in comparison to probability samples, as well as that raking as a post-survey adjustment method improves the quality of estimates from probability surveys more effectively than for nonprobability-based online panels. While nonprobability panel samples are similarly accurate in measuring secondary demographics as probability samples (AAE: nonprobability samples 4.7-5.4, probability samples 4.2-5.3, all raked), they are less accurate in measuring non-demographics than probability surveys (AAE: nonprobability samples 6.6-9.9, probability samples 3.7-5.4, all raked), which is also confirmed by RMSE measures. We would particularly like to reduce the non-demographic bias with various post-survey adjustments.

Assessment of Effectiveness of Post-Survey Adjustment Methods for Improving Inference in Nonprobability Samples

In this section, we will show if the difference in accuracy between probability and nonprobability samples, i.e., representation bias, can be reduced using different post-survey adjustment methods. The results will be presented by scenarios based on the availability of external data and, as previously explained, not all methods can be used with all data types. Importantly, we will use Life in Australia™ Wave 2 sample as a reference sample for post-survey adjustment efficiency. This sample has been selected as it is similarly accurate as the OPBS 2015 probability samples (see online Appendix Table 5), yet with a much larger sample size (smaller sampling variance) and greater comparability with nonprobability samples in terms of the survey mode (online: 86.4% in Life in Australia™ Wave 2, 100% in volunteer samples). We will use AAE for the raked[19] Life in Australia™ sample, and no further post-survey adjustment will be carried out with this probability sample. Fundamentally, we will assess (i) the efficiency of post-survey adjustments with nonprobability samples relative to (ii) the accuracy of probability-based online panel estimates normally reported in practice (i.e., calibrated using primary demographics).

Scenario 1: Availability of Census Aggregated Statistics, and Only Primary Demographics were Collected from the Nonprobability Sample

To illustrate the effectiveness of post-survey adjustments (i.e., raking, GREG and MRP) using primary demographics (i.e., performing 'basic calibration'), we are presenting results for unweighted and weighted data for the nonprobability online samples in Figure 1 (see Table 6 from the Appendix for more detailed results). The presented evidence shows how basic weighting post-survey adjustments improve the quality of estimates, but the improvement is only slight on average (AAE combined reduction between 0.4 [GREG, Panel 3] and 0.7 [MRP, Panel 1]). We can confirm our previous finding on how raking improves the accuracy of nonprobability samples to a lesser extent than those from probability samples. We can also extend this finding to other calibration methods studied in this article – GREG and MRP.

The improvement in accuracy is more apparent for all 18 survey items combined than for six substantive items combined, which indicates that calibration using primary demographic more consistently improves the quality of secondary demographic estimates than non-demographic estimates. Moreover, the results from Figure 1 show how calibration can deteriorate substantive item estimates from nonprobability samples, especially the least accurate one, but also the combined volunteer panel sample. This is consistent across all calibration methods, with MRP performing just slightly better than GREG and raking. On the other hand, weighting improved accuracy of the most accurate nonprobability panel in a similar fashion for both secondary demographics and non-demographics.

We have to note that the differences in item-level results (not only at the AAE level, see Table 6 in the Appendix) are almost non-existent for raking and GREG and very little between the first two calibration methods and MRP. Based on this finding, as well as due to the limitations of MRP (i.e., requiring a joint distribution), we will only assess the efficiency of the first two calibration methods under Scenario 2. Also, the results for basic raking from Scenario 1 are included as a reference method for Scenarios 2, 3 and 4 (see Figures 2, 3 and 4).

[19] By gender, age group*education (interaction), country of birth, state*capital city in state (interaction)

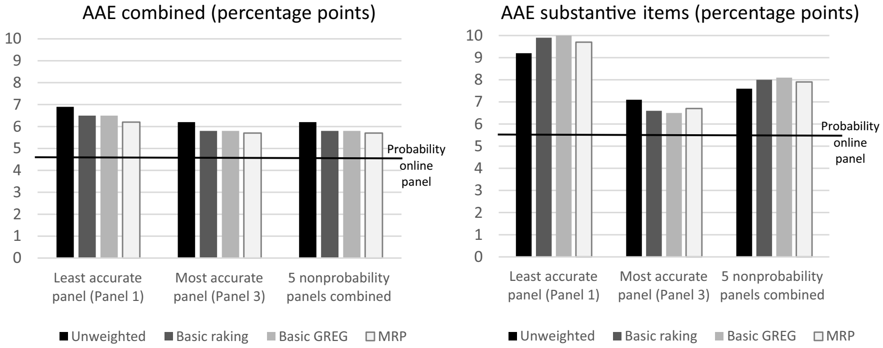


Figure 1 Accuracy of post-survey adjusted nonprobability panel samples for Scenario 1 - average absolute error (AAE) for all sample estimates (see Table 6), un- and weighted (raking, GREG, MRP)*

* AAE for secondary demographics and all RMSE calculations (combined, secondary demographics, and substantive items) are presented in the tables in the Appendix.

Scenario 2: Availability of Census Aggregated Statistics, Both Primary and Secondary Demographics were Collected from the Nonprobability Sample

To illustrate how including new covariates in calibration further improves the accuracy of nonprobability samples, additional socio-demographic items with corresponding census benchmarks were added[20] and 'expanded' calibration 1 (e.g., expanded raking 1) was performed (see Figure 2). The presented evidence suggests that expanded raking and GREG predominantly improved secondary demographic estimates and, in some cases, estimates from substantive items (see Table 7 from the Appendix for more detailed results). For the most and the least accurate online panel, as well as all panels combined, we can see a slight improvement in the combined AAE and RMSE. Generally speaking, we can again report almost negligible differences between estimates adjusted with expanded raking and expanded GREG. We also did not notice a significant increase of design effect compared to basic raking.

Moreover, this time calibration did not increase AAE for substantive items for the least accurate panel and five panels combined. Including three secondary demographic covariates seemed to eliminate the negative effect of raking with primary demographics only. Moreover, we can notice a notable improvement in accuracy of substantive items after using an expanded raking scheme for the most

[20] For more information on selection of additional covariates under Scenarios 2, 3 and 4 (e.g., employment status, language other than English, and voluntary work under Scenario 2), see *Post-survey adjustment approaches and parameters* section (Methods) and Table 10 (Appendix).

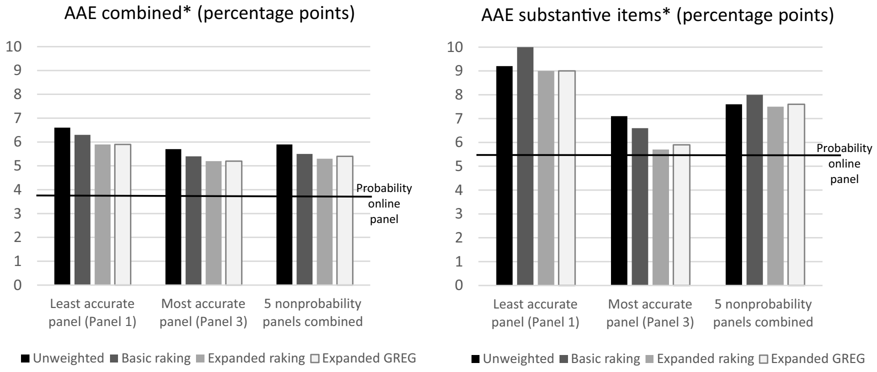


Figure 2 Accuracy of post-survey adjusted nonprobability panel samples for Scenario 2 - average absolute error (AAE) for all sample estimates (see Table 7), unweighted and weighted (raking, GREG)

*AAE were calculated for all items excluding the secondary demographics included in an expanded calibration scheme (employment status, language other than English (LOTE), and voluntary work, see Table 7 in the Appendix for more information)

accurate nonprobability panel (AAE: unweighted 7.1, raking 5.7, GREG 5.9). The selected secondary demographic items seem to be more associated with representation bias in the most accurate nonprobability online panel than our core/primary demographics.

The evidence from Figures 1 and 2 suggests that the highest-quality nonprobability online panels are not only the most accurate for unweighted estimates, but they also respond better to various calibration adjustments.

Scenario 3: Availability of Census Aggregated Statistics and One Other Representative Source of Benchmarks

To illustrate potential added value of having access to an additional external high-quality data source with non-demographic matching covariates, we are presenting results for ‘expanded’ calibration 2, CEM, MDM, and PSW in Figure 3. The presented evidence shows how including new non-demographic covariates in post-survey adjustment improves the accuracy of nonprobability samples fairly similarly to including new secondary demographic covariates. However, the improvement seems to be more substantial under Scenario 3 – an increase in accuracy measured with AAE combined ranges from 0.4 (Panel 1, MDM) to 1.8 (Panel 1, CEM).

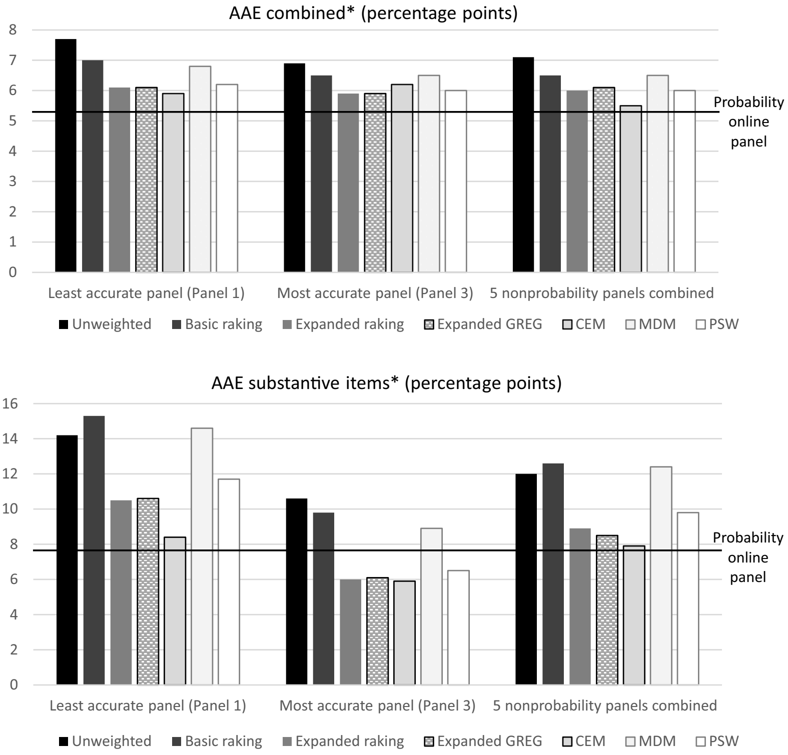


Figure 3 Accuracy of post-survey adjusted nonprobability panel samples for Scenario 3 - average absolute error (AAE) for all sample estimates (see Table 8), unweighted and adjusted post-survey (raking, GREG, CEM, MDM, PSW)

*AAE were calculated for all items excluding the covariates in an expanded post-survey adjustment scheme (household status, frequency of smoking, and drinking alcohol, see Table 8 in the Appendix for more information)

In comparison to the efficiency of calibration under Scenario 2, including non-demographic covariates improved the accuracy of substantive items[21] to a greater extent. The decrease in that AAE (substantive items) was as high as 5.8 (Panel 3, CEM). Generally speaking, post-survey adjustment with a limited number of covariates was more efficient with calibration (raking, GREG) and CEM than distance-based models, i.e., PSW and especially MDM. While CEM seems to com-

[21] The remaining three substantive items for benchmarking were from National Health Survey 2014-15 and General Social Survey 2014 (see Table 2).

pare favourably to other methods using covariates from a large-scale survey, we noticed a larger design effect than for expanded raking 2.

All in all, post-survey adjustment with expanded raking, GREG and CEM under Scenario 3 made nonprobability online panels almost as accurate as a probability-based online panel overall (AAE combined). For the three remaining substantive items, the most accurate nonprobability online panel (Panel 3) was even more accurate after advanced adjustments than the probability online panel after basic raking.

Scenario 4: Availability of Census Aggregated Statistics and a Smaller-Scale Probability-Based Survey Data with Matching Variables from Nonprobability-Based Survey Data

To illustrate potential added value of having access to a smaller-scale external survey data source (i.e., OPBS 2017 replication sample from a probability online panel) with non-demographic matching covariates, we are presenting results for CEM, MDM, and PSW[22] in Figure 4.

The results present mixed evidence on the efficiency of post-survey adjustment methods using smaller-scale external survey data with no demographics or health-associated items. First, there was a fairly moderate and inconsistent effect of post-survey adjustments on the total accuracy of nonprobability samples. In most cases, the decrease of AAE combined was less than 0.5, and no method seemed to have a clear advantage. The only exception to the rule was MDM with the data from five nonprobability-based panels combined (AAE: unweighted 6.2, MDM 5.2, probability panel 4.6). Overall, basic raking with primary demographics from Australian Census seems to be a more reliable method than any other method for improving the combined accuracy of secondary demographic and non-demographic estimates with webographics.

Comparing AAE for substantive items, we can observe as many instances of post-survey adjustment deteriorating estimates as instances of improving estimates. The least accurate nonprobability-based panel stands out as the sample with no decrease in AAE before or after adjustment, and CEM as the method with limited efficiency for only one sample (the most accurate). The best result overall can again be attributed to MDM (AAE: unweighted 7.6, MDM 6.5, probability panel 5.4), and we can also see a positive effect of PSW on the accuracy of Panel 3 (AAE: unweighted 7.1, PSW 6.0, probability panel 5.4).

[22] A variety of other methods and their combinations would be possible under this scenario with auxiliary microdata, including calibration such as raking, GREG and MRP. However, calibration is normally carried out with benchmarks from the highest-quality censuses or large-scale surveys, and smaller-scale probability-based survey tend to introduce more error (see Table 5, probability samples).

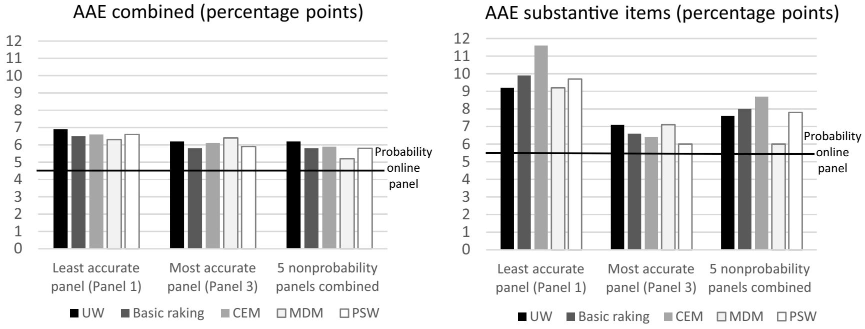


Figure 4 Accuracy of post-survey adjusted nonprobability panel samples for Scenario 4 - average absolute error (AAE) for all sample estimates (see Table 9), unweighted and adjusted (raking and matching methods)

Summary of Post-Survey Adjustment Efficiency

To sum up, we are presenting a review of all post-survey adjustment results by four data availability scenarios. All AAE combined values from Scenarios 1-4 and associated AAE reduction % (as a proportion of unadjusted/unweighted AAE) are now combined.

Based on the results from Table 4 (as well as Figures 1-4), we are offering the following main findings of our study:

- the best post-survey adjustment results can be expected under Scenario 3, i.e., by using a combination of primary, secondary, and non-demographic covariates from nationally representative data sources;
- expanded calibration with additional secondary demographic covariates further improves accuracy, in comparison to basic calibration with primary demographic covariates (and to a similar extent);
- secondary demographic covariates seem to have a better potential to improve the accuracy of secondary demographic estimates, and non-demographic covariates seem to have a better potential to improve the accuracy of non-demographic estimates (in this particular study, those were health-related items);
- webographics from probability-based online panel survey data did not consistently improve the accuracy of nonprobability samples (see Scenario 4 results);
- there are some observable differences between the analyzed methods, albeit they are little in this study, and MDM was the method with the least consistent results;
- while we could not reduce error by more than 23% no matter the chosen auxiliary data, covariates or methods, we have to note that the probability samples

Table 4 Post-survey adjustments efficiency for all methods and samples under four scenarios

Scenario	Post-survey adjustment method	Least accurate nonprobability panel (1), AAE combined			Most accurate nonprobability panel (3), AAE combined			5 nonprobability panels combined, AAE combined		
		AAE (UW)	AAE (adjusted)	AAE reduction (%)	AAE (UW)	AAE (adjusted)	AAE reduction (%)	AAE (UW)	AAE (adjusted)	AAE reduction (%)
Scenario 1	Basic raking	6.9	6.5	6%	6.2	5.8	6%	6.2	5.8	6%
	Basic GREG		6.5	6%		5.8	6%		5.8	6%
	MRP		6.2	10%		5.7	8%		5.7	8%
Scenario 2*	Basic raking	6.6	6.3	4%	5.7	5.4	5%	5.9	5.5	7%
	Expanded raking 1		5.9	11%		5.2	9%		5.3	10%
	Expanded GREG 1		5.9	11%		5.2	9%		5.4	8%
Scenario 3*	Basic raking	7.7	7.0	9%	6.9	6.9	0%	7.1	6.5	8%
	Expanded raking 2		6.1	21%		5.9	14%		6.0	15%
	Expanded GREG 2		6.1	21%		5.9	14%		6.1	14%
	CEM		5.9	23%		6.2	10%		5.5	23%
	MDM		6.8	12%		6.5	6%		6.5	8%
	PSW		6.2	19%		6.0	13%		6.0	15%
Scenario 4	Basic raking	6.9	6.5	6%	6.2	5.8	6%	6.2	5.8	6%
	CEM		6.6	4%		6.1	2%		5.9	5%
	MDM		6.3	9%		6.4	-3%		5.2	16%
	PSW		6.6	4%		5.9	5%		5.8	6%

*AAE were calculated for all items excluding the covariates in an expanded post-survey adjustment scheme, UW – unweighted estimates from OPBS 2015 and the OPBS 2017 Replication sample were about 20-30% more accurate than the studied nonprobability samples[23].

Discussion and Conclusion

This investigation into improving inference in nonprobability sample surveys supports the conclusion that the issue of improving inference in nonprobability sample surveys is a three-dimensional problem. First, the quality of post-survey adjustments is dependent on the availability of relevant high-quality covariates which are associated with either representation bias in nonprobability samples or outcome variables. Second, as the covariates in nonprobability samples should have matching covariates in external representative data sources, the availability and ability to access auxiliary data is a key aspect in mitigating bias. Third, the efficiency of post-survey adjustments is also dependent on the selection and combination of post-survey adjustment methods, albeit to a lesser extent.

In this study, we presented evidence that post-survey adjustment can reduce representation bias in nonprobability online samples to some extent, but cannot consistently eliminate it. These findings are in line with evidence from Tourangeau et al. (2014) and Kalton and Flores-Cervantes (2003). However, we demonstrated a greater potential to mitigate representation bias in nonprobability panels if having access to more external data sources and more covariates matching in nonprobability samples and auxiliary data. Ideally, we would have access to large-scale survey microdata, since smaller-scale surveys come with some nonignorable error. While those probability surveys mostly remain more accurate than nonprobability surveys even after post-survey adjustments, they are more susceptible to coverage, sampling, and nonresponse error (or even measurement mode effect) than most high-quality government surveys, and the total representation error can be carried over to post-survey adjustment results (e.g., after matching or PSW). For that reason, improving inference in nonprobability samples should be planned in the survey design stage, and relevant external data sources reviewed before data collection, if possible.

Moreover, identification of covariates from external data sources which are associated with representation bias or target outcome variables can lead to a more efficient mitigation of bias. While post-survey adjustments using primary demographics have little positive effect on the quality of nonprobability estimates, we have shown how including secondary demographics can improve the quality of other demographics and including non-demographics can decrease the error from

[23] This research did not take into account that the accuracy of probability samples could be further improved with the same post-survey adjustment methods including secondary demographic and non-demographic items.

associated non-demographics. This is consistent with findings from Bethlehem (2002). Similarly, Mercer et al. (2018) reported that including political attitude covariates in adjustment improved the quality of political engagement estimates. However, we found inconsistent evidence on the suitability of ‘webographics’ and other internet-associated covariates for mitigating bias in nonprobability samples. Unfortunately, we could not distinguish between the effect of those covariates and the effect of the data source on the post-survey adjustment efficiency. While auxiliary variables like early adopter items (traditionally used to mitigate bias in nonprobability samples, e.g., DiSogra et al. 2011) did not distinguish our probability online sample and nonprobability online panel samples well, we identified new covariates for post-survey adjustment that could be considered as ‘webographics’, such as the number of surveys participated in. Therefore, we believe it is crucial to carry out more investigation into ‘good’ webographic variables for post-survey adjustment, as previously suggested by Dutwin and Buskirk (2017). Our study also highlights the importance of selection bias and representativeness, and how this varies between different nonprobability samples (Lehdonvirta et al. 2021).

The investigation into the suitability of post-survey adjustment methods did not highlight a particular method or a combination of them which consistently performed better parameter estimates. This supports the finding from Mercer et al. (2018). While a detailed technical investigation into calibration methods was not the focus of this study, we found little differences in efficiency between the investigated methods: raking and the model-based methods (such as GREG or MRP), which was consistent with findings from Kalton and Flores-Cervantes (2003). Therefore, we suggest the selection of calibration methods to be instead based on the availability of joint distributions of covariates weighed against the computational intensity of methods. While matching methods and PSW under limited scenarios might have a better potential for efficient post-survey adjustment, we observed less consistency in bias reduction between different samples and scenarios. We also observed an increase of design effect for CEM and, consequently, confidence intervals for estimates (see Kolenikov 2014).

This study has several limitations, including the availability of external data and covariates both in nonprobability surveys and high-quality government surveys. Having access to additional data sources could improve post-survey adjustments and help distinguish better between the efficiency of covariates, the effect of quality of external data sources, and the efficiency of methods. Moreover, since estimates for only 18 items were compared to benchmarks and the majority of substantive items were more or less associated with one topic (i.e., health status), the findings would be more robust if survey items with corresponding benchmarks would be associated with other aspects of respondent’s lives, not only health. In addition, the total survey error framework (Biemer 2010; Groves et al. 2009; Groves & Lyberg 2010) has been proposed to provide a comprehensive overview of all possible sources of sampling and non-sampling errors and give a systematic measure of survey quality that encompasses not just accuracy but also bias. The framework attempts to account for, and assess, many sources of error that arise through the survey process (which we could not study separately, e.g., measurement mode effects versus representation bias). This framework lends itself to the Bayesian paradigm through incorporating prior information (Shirani-Mehr et al. 2018) or using expert opinion (Toepoel & Emerson 2017) in assessing the survey quality of surveys not based on probability schemes. We would suggest future research on improving inference in nonprobability samples to be more targeted, planned and properly designed in advance. Nonetheless, the approaches discussed in this chapter have distinct long-term benefits in improving the inferences from surveys conducted using nonprobability samples.

References

Australian Bureau of Statistics. (n.d.-a). *TableBuilder.* Retrieved November 1, 2020, from https://www.abs.gov.au/websitedbs/d3310114.nsf/home/about+tablebuilder

Australian Bureau of Statistics. (n.d.-b). *Compare data services.* Retrieved January 16, 2021, from https://www.abs.gov.au/websitedbs/D3310114.nsf/4a256353001af3ed4b2562bb00121564/c00ee824af1f033bca257208007c3bd5!OpenDocument

Australian Bureau of Statistics. (2016). *2016 Census of Population and Housing* [Census TableBuilder], accessed 1 November, 2020.

Australian Bureau of Statistics. (2018a, March 28). *Household use of information technology.* https://www.abs.gov.au/statistics/industry/technology-and-innovation/household-use-information-technology/latest-release

Australian Bureau of Statistics. (2018b, December 12). *National Health Survey: First results.* https://www.abs.gov.au/statistics/health/health-conditions-and-risks/national-health-survey-first-results/latest-release

Baker, R., Blumberg, S. J., Brick, J. M., Couper, M. P., Courtright, M., Dennis, J. M., Dillman, D., Frankel, M. R., Garland, P., Groves, R. M., Kennedy, C., Krosnick, J., Lavrakas, P. J., Lee, S., Link, M., Piekarski, L., Rao, K., Thomas, R. K., & Zahs, D. (2010). Research Synthesis: AAPOR Report on Online Panels. *Public Opinion Quarterly, 74*(4), 711–781. https://doi.org/10.1093/poq/nfq048

Baker, R., Brick, J. M., Bates, N. A., Battaglia, M., Couper, M. P., Dever, J. A., Gile, K. J., & Tourangeau, R. (2013). Summary Report of the AAPOR Task Force on Non-probability Sampling. *Journal of Survey Statistics and Methodology, 1*(2), 90–143. https://doi.org/10.1093/jssam/smt008

Barnier, J., Briatte, F., & Larmarange, J. (2020). *Questionr: Functions to make surveys processing easier* (R package version 0.7.1) [Computer software]. The Comprehensive R Archive Network. Available from https://CRAN.R-project.org/package=questionr

Battaglia, M. P., Hoaglin, D. C., & Frankel, M. R. (2009). Practical Considerations in Raking Survey Data. *Survey Practice, 2* (5). https://doi.org/10.29115/SP-2009-0019.

Bethlehem, J. G. (2002). *Weighting nonresponse adjustments based on auxiliary information.* Wiley.

Bethlehem, J. G. (2008). *Weighting.* In P. J. Lavrakas (Ed.), *Encyclopedia of survey research methods* (pp. 957-960). Sage.

Bethlehem, J. G. (2016). Solving the nonresponse problem with sample matching?. *Social Science Computer Review, 34*, 59-77.

Bethlehem, J. G., & Biffignandi, S. (2012). *Handbook of Web Surveys.* Wiley & Sons.

Biemer, P. P. (2010). Total survey error: Design, implementation, and evaluation. *Public Opinion Quarterly*, *74*(5), 817-848.

Bon, J. J., Ballard, T., & Baffour, B. (2019). Polling bias and undecided voter allocations: US presidential elections, 2004–2016. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *182*(2), 467-493.

Brick, J. M. (2013). Unit nonresponse and weighting adjustments: A critical review. *Journal of Official Statistics*, *29*(3), 329.

Budescu, D. V. (1993). Dominance analysis: a new approach to the problem of relative importance of predictors in multiple regression. *Psychological bulletin*, *114*(3), 542.

Callegaro, M., & DiSogra, C. (2008a). Computing response metrics for online panels. *Public opinion quarterly*, *72*(5), 1008-1032.

Callegaro, M., & DiSogra, C. (2008b). Probability of Selection. In P. J. Lavrakas (Ed.), *Encyclopedia of survey research methods* (pp. 617-618). Sage.

Cochran, W. G. (1977). *Sampling Techniques.* John Wiley & Sons.

Cornesse, C., Blom, A. G., Dutwin, D., Krosnick, J. A., de Leeuw, E. D., Legleye, S., Pasek, J., Pennay, D., Phillips, B., Sakshaug, J. W., Struminskaya, B., & Wenz, A. (2020). A Review of Conceptual Approaches and Empirical Evidence on Probability and Nonprobability Sample Survey Research. *Journal of Survey Statistics and Methodology*, 8(1), 4–36. https://doi.org/10.1093/jssam/smz041

Couper, M. P. (2000). Web surveys: A review of issues and approaches. *The Public Opinion Quarterly*, 64(4), 464-494.

Daikeler, J., Bošnjak, M., & Lozar Manfreda, K. (2020). Web versus other survey modes: an updated and extended meta-analysis comparing response rates. *Journal of Survey Statistics and Methodology,* 8(3), 513-539.

DataReportal. (2022, February 9). *Digital 2022: Australia.* https://datareportal.com/reports/digital-2022-australia

Deville, J. C., & Särndal, C. E. (1992). Calibration estimators in survey sampling. *Journal of the American statistical Association*, *87*(418), 376-382.

Deville, J. C., Särndal, C. E., & Sautory, O. (1993). Generalized raking procedures in survey sampling. *Journal of the American statistical Association*, *88*(423), 1013-1020.

DiSogra, C., Cobb, C., Chan, E., & Dennis, J. M. (2011). Calibrating non-probability internet samples with probability samples using early adopter characteristics. *Joint Statistical Meetings, Survey Research Methods*, 4501-4515.

Dutwin, D., & Buskirk, T. D. (2017). Apples to oranges or gala versus golden delicious? Comparing data quality of nonprobability internet samples to low response rate probability samples. *Public Opinion Quarterly*, *81*(S1), 213-239.

Dutwin, D., & Buskirk, T. D. (2021). Telephone sample surveys: dearly beloved or nearly departed? Trends in survey errors in the era of declining response rates. *Journal of Survey Statistics and Methodology*, *9*(3), 353-380.

Elliott, M. R. (2009). Combining data from probability and non-probability samples using pseudo-weights. *Survey Practice, 2,* 1-7.

Elliott, M. R., & Valliant, R. (2017). Inference for nonprobability samples. *Statistical Science*, *32*(2), 249-264.

ESOMAR (2021, March). *Questions to help buyers of online samples.* https://esomar.org/uploads/attachments/ckqqecpst00gw9dtrl32xetli-questions-to-help-buyers-of-online-samples-2021.pdf

Gelman, A. (2007). Struggles with survey weighting and regression modeling. *Statistical Science*, *22*(2), 153-164.

Gelman, A., & Little, T. C. (1997). Poststratification into many categories using hierarchical logistic regression. *Survey Methodology, 23,* 127–135.

Goodrich, B., Gabry, J., Ali, I., & Brilleman, S. (2020). *rstanarm: Bayesian applied regression modeling via Stan.* R package version 2.21.1, https://mc-stan.org/rstanarm.

Goot, M. (2021). How good are the polls? Australian election predictions, 1993–2019. *Australian Journal of Political Science*, 56(1), 35-55.

Groves, R. M. (2006). Nonresponse rates and nonresponse bias in household surveys. *Public opinion quarterly*, *70*(5), 646-675.

Groves, R. M., Fowler, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2009). *Survey Methodology.* John Wiley & Sons.

Groves, R. M., & Lyberg, L. (2010). Total survey error: Past, present, and future. *Public opinion quarterly*, *74*(5), 849-879.

Hade, E. N., & Lemeshow, S. (2011). In P. J. Lavrakas (Ed.), *Encyclopedia of survey research methods* (pp. 621-623). Sage.

Harrell, F. E. Jr., Dupont, C., & others (2020). *Hmisc: Harrell miscellaneous* (R package version 4.4-1) [Computer software]. The Comprehensive R Archive Network. Available from https://CRAN.R-project.org/package=Hmisc.

Heckman, J. (1979). Sample selection bias as a specification error. *Econometrica, 47*(1), 153–62.

Hewitt, M. (2017). *National Drug Strategy Household Survey 2016.* (ADA Dataverse, Version 7) [Data set]. ADA. https://doi.org/10.4225/87/JUDY2Y

Hug, S. (2003). Selection Bias in Comparative Research: The Case of Incomplete Data Sets. *Political Analysis, 11*(3), 255-274. https://doi.org/10.1093/pan/mpg014

Iacus, S. M., King, G., & Porro, G. (2009). CEM: Coarsened Exact Matching Software. *Journal of Statistical Software, 30.* http://gking.harvard.edu/cem

Iacus, S. M., King, G., & Porro, G. (2011). Multivariate matching methods that are monotonic imbalance bounding. *Journal of the American Statistical Association*, *106*(493), 345-361.

Iacus, S. M., King, G., & Porro, G. (2020). *Cem: Coarsened exact matching (R package version 1.1.20)* [Computer software]. The Comprehensive R Archive Network. Available from https://CRAN.R-project.org/package=cem

International Organisation for Standardisation (2022). *ISO 26362:2009.* https://www.iso.org/standard/43521.html

Kaczmirek, L., Phillips, B., Pennay, D. W., Lavrakas, P. J., & Neiger, D. (2019). Building a probability-based online panel: Life in Australia™. *CSRM and SRC Methods Paper, 2019* (2).

Kalton, G., & Flores-Cervantes, I. (2003). Weighting methods. *Journal of Official Statistics*, *19*(2), 81-97.

Kaplan, J. (2020). *fastDummies: Fast creation of dummy (binary) columns and rows from Categorical Variables* (R package version 1.6.1) [Computer software]. The

Comprehensive R Archive Network. Available from https://CRAN.R-project.org/package=fastDummies

Keiding, N., & Louis, T. A. (2016). Perils and potentials of self-selected entry to epidemiological studies and surveys. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 319-376.

Kennedy, C., & Hartig, H. (2019). *Response rates in telephone surveys have resumed their decline*. Pew Research Center.

Kennedy, C., M. Blumenthal, S. Clement, J. D. Clinton, C. Durand, C. Franklin, K. McGeeney, L. Miringoff, K. Olson, D. Rivers, L. Saad, G. E. Witt, & Wlezien, C. (2018). An evaluation of the 2016 election polls in the United States. *Public Opinion Quarterly*, 82(1), 1-33.

Kennedy, C., Mercer, A., Keeter, S., Hatley, N., McGeeney, K., & Gimenez, A. (2016). *Evaluating online nonprobability surveys.* Pew Research Center.

King, G., Lucas, C., & Nielsen, R. A. (2015). *{MatchingFrontier}: {R} Package for Computing the Matching Frontier ()* [Computer software]. The Comprehensive R Archive Network. Available from http://projects.iq.harvard.edu/frontier

Kolenikov, S. (2014). Calibrating survey data using iterative proportional fitting (raking). *The Stata Journal*, *14*(1), 22-59.

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.

Lehdonvirta, V., Oksanen, A., Räsänen, P., & Blank, G. (2021). Social media, web, and panel surveys: using non-probability samples in social and policy research. *Policy & internet*, *13*(1), 134-155.

Little, R. J. A., & Rubin, D. B. (2002). Single imputation methods. In R. J. A. Little, & D. B. Rubin (Eds.), *Statistical analysis with missing data* (pp. 59-74). Wiley.

Lüdecke, D. (2020). *Sjstats: Statistical functions for regression models* (version 0.18.0) () [Computer software]. The Comprehensive R Archive Network. Available from https://CRAN.R-project.org/package=sjstats

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.

Malhotra, N., & Krosnick, J. A. (2007). The Effect of Survey Mode and Sampling on Inferences about Political Attitudes and Behavior: Comparing the 2000 and 2004 ANES to Internet Surveys with Nonprobability Samples. *Political Analysis, 15*, 286–323.

Matei, A. (2018). On Some Reweighting Schemes for Nonignorable Unit Nonresponse. *Survey Statistician, 77*, 21–33.

Mercer, A., Lau, A., & Kennedy, C. (2018). *For weighting online opt-in samples, what matters most.* Pew Research Center.

Mercer, A. W., Kreuter, F., Keeter, S., & Stuart, E. A. (2017). Theory and practice in nonprobability surveys: parallels between causal inference and survey inference. *Public Opinion Quarterly*, *81*(S1), 250-271.

Park, D. K., Gelman, A., & Bafumi, J. (2004). Bayesian multilevel estimation with poststratification: State-level estimates from national polls. *Political Analysis*, *12*(4), 375-385.

Park, D. K., Gelman, A., & Bafumi, J. (2006). State-level opinions from national surveys: Poststratification using multilevel logistic regression. In J. E. Cohen (Ed.), *Public opinion in state politics* (pp. 209-228). Stanford University Press.

Pasek, J. (2018). *Anesrake: Anes raking implementation* (R package version 0.80) [Computer software]. The Comprehensive R Archive Network. Available from https://CRAN.R-project.org/package=anesrake

Pennay, D. W., Borg, K., Neiger, D., Misson, S., Honey, N., & Lavrakas, P. (2016). *Online Panels Benchmarking Study, 2015* (ADA Dataverse, Version V1) [Data set]. ADA. https://doi.org/10.4225/87/FSOYQI

Pennay, D. W., & Neiger, D. (2020). *Health, Wellbeing and Technology Survey (OPBS replication), 2017* (ADA Dataverse, Version V1) [Data set]. ADA. https://doi.org/10.26193/YF8AF1

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 nonprobability online panel surveys in Australia. *CSRM and SRC Methods Paper, 2018* (2).

Pfeffermann, D., Eltinge, J. L., & Brown, L. D. (2015). Methodological issues and challenges in the production of official statistics: 24th annual Morris Hansen lecture. *Journal of Survey Statistics and Methodology, 3*(4), 425-483.R Core Team. (2020). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing. https://www.R-project.org/

Rivers, D. (2007, July 29–August 2). *Sampling for Web Surveys* [Conference presentation]. 2007 Joint Statistical Meetings, Salt Lake City, United States of America.

Rivers, D. (2013). Comment on task force report. *Journal of Survey Statistics and Methodology, 1*(2), 111-117.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70*(1), 41-55.

Rothman, K. J., Greenland, S., & Lash, T. L. (2008). Validity in epidemiologic studies. In K. J. Rothman, S. Greenland, T. L. Lash (Eds.), *Modern epidemiology* (3rd ed.) (pp. 128-147). Lippincott Williams & Wilkins.

Schonlau, M., & Couper, M. P. (2017). Options for conducting web surveys. *Statistical Science*, *32*(2), 279-292.

Schonlau, M., Soest, V. A., & Kapteyn, A. (2007). Are "Webographic" or Attitudinal Questions Useful for Adjusting Estimates from Web Surveys Using Propensity Scoring?. *Survey Research Methods*, *1*, 155–163.

Shirani-Mehr, H., Rothschild, D., Goel, S., & Gelman, A. (2018). Disentangling bias and variance in election polls. *Journal of the American Statistical Association*, *113*(522), 607-614.

Sizemore, S., & Alkurdi, R. (2019). *Matching Methods for Causal Inference: A Machine Learning Update*. Available from https://humboldt-wi.github.io/blog/research/applied_predictive_modeling_19/matching_methods/.

Stekhoven, D. J. (2013). *missForest: Nonparametric missing value imputation using random forest* (R package version 1.4) [Computer software]. The Comprehensive R Archive Network. Available from https://github.com/stekhoven/missForest

Stuart, E. A. (2010). The Use of Propensity Scores to Assess Generalizability. *Journal of the Royal Statistical Society, Series A: Statistics in Society, 174*(2), 369–386.

Stuart, E. A., & Rubin, D. B. (2008). Best practices in quasi-experimental designs. In J. W. Osborne (Ed.), *Best practices in quantitative methods* (pp. 155-176). Sage.

Toepoel, V., & Emerson, H. (2017). Using experts' consensus (the Delphi method) to evaluate weighting techniques in web surveys not based on probability schemes. *Mathematical Population Studies*, *24*(3), 161-171, DOI: 10.1080/08898480.2017.1330012.

Tourangeau, R., Edwards, B., Johnson, T. P., Wolter, K. M., & Bates, N. (2014). *Hard-to-survey populations.* Cambridge University Press.

Tourangeau, R., & Smith, W. (1985). Finding subgroups for surveys. *Public Opinion Quarterly, 49*(3), 351-365.

Valliant, R. (2020). Comparing alternatives for estimation from nonprobability samples. *Journal of Survey Statistics and Methodology*, *8*(2), 231-263.

Valliant, R., & Dever, J. A. (2011). Estimating propensity adjustments for volunteer web surveys. *Sociological Methods & Research*, *40*(1), 105-137.

Wang, W., Rothschild, D., Goel, S., & Gelman, A. (2015). Forecasting elections with non-representative polls. *International Journal of Forecasting, 31*(3), 980-991.

Wang, Y., Dai, Y., Li, H., & Song, L. (2021). Social Media and Attitude Change: Information Booming Promote or Resist Persuasion?. *Frontiers in Psychology*, 12, 2433.

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.

methods, data, analyses | Vol. 17(2), 2023, pp. 207-248 DOI: 10.12758/mda.2022.15

Social Media Recruitment in Online Survey Research: A Systematic Literature Review

Zaza Zindel

Bielefeld University

Abstract

The growing percentage of the population on social media creates new and expanded opportunities for survey researchers. Recently, a growing number of studies have been using social media to recruit survey respondents. Many social media platforms have powerful targeting capabilities that can be used to recruit even rare or hard-to-reach populations. However, thus far, the survey research literature lacks a comprehensive overview of potentials and limitations. This literature review aims 1) to provide an overview of the current literature on the use of social media as a recruitment tool, 2) to highlight the potential advantages and disadvantages for survey research, 3) to identify current research gaps, and finally, 4) to provide practical guidance for researchers interested in integrating social media recruitment into their research.

Keywords: Literature Review; Social Media; Participant Recruitment; Research Design; Online Survey Research

The Internet has changed the social sciences dramatically by opening up new forms and fields of research, including the study of human behavior in online social networks (e.g., Ferg et al., 2021; Orehek & Human, 2017) and investigations of the Internet's impact on human (co-)existence (e.g., Erhardt & Freitag, 2021; Lu & Yu, 2019). The Internet also offers new forms of readily available data that can complement or, in some cases, replace primary data collection (e.g., Bach et al., 2021; Stier et al., 2020). Moreover, the Internet is itself a valuable tool for social research. Today, online research methods are used in most of the social sciences.

Given the potential of Internet technology and the unique features of online human behavior, social media (SM) sites offer a promising approach for recruiting survey participants. Platforms such as Facebook, Instagram, and Twitter connect hundreds of millions of users, all of whom represent potential respondents. Over the past decade, numerous studies have shown that it is possible to reach and recruit large numbers of participants for scientific surveys through SM (e.g., Grow et al., 2020; Kühne & Zindel, 2020; Pötzschke & Braun, 2017). The growing percentage of the population on SM creates new and expanded opportunities for the recruitment of participants in social research. Many SM platforms, in particular, Facebook, Instagram, and Twitter, have powerful targeting capabilities that can be used to recruit hard-to-reach populations. SM targeting tools allow researchers to track and reach users with specific demographic characteristics and interests based on their behavior both on the SM sites themselves and on other third-party websites that users interact with through their SM accounts. These features reduce the time and resources required to recruit rare and hard-to-reach populations. In light of the low effectiveness of traditional recruitment methods in reaching these groups, SM recruitment tools may prove to be an effective and efficient means of recruiting otherwise overlooked populations.

To decide whether SM recruitment tools could be useful for their own surveys, researchers need a better understanding of 1) which participants are likely to be reached through online surveys, 2) how other researchers have recruited similar samples via SM, and 3) what advantages and disadvantages SM recruitment strategies have compared to other recruitment strategies. To date, the lack of a comprehensive literature review on the role of SM in recruiting participants for social surveys makes it difficult for researchers to determine whether SM could be a viable method for their purposes.

To enable more informed decisions about the use of SM in survey recruitment, this research synthesis provides a broad overview of existing publications using SM recruitment. In reviewing the existing literature, my aim was to evalu-

Direct correspondence to
Zaza Zindel, Bielefeld University, Universitätsstraße 25, 33615 Bielefeld
E-mail: zaza.zindel@uni-bielefeld.de

ate the methodology and effectiveness of survey recruitment, highlight advantages and disadvantages for survey research, identify current research gaps, and provide practical guidance for further research. I examined: (1) the effectiveness of SM sampling strategies for the targeted populations, (2) the cost-effectiveness of these approaches, and (3) the comparability and quality of the various approaches based on the demographic distribution of the SM samples.

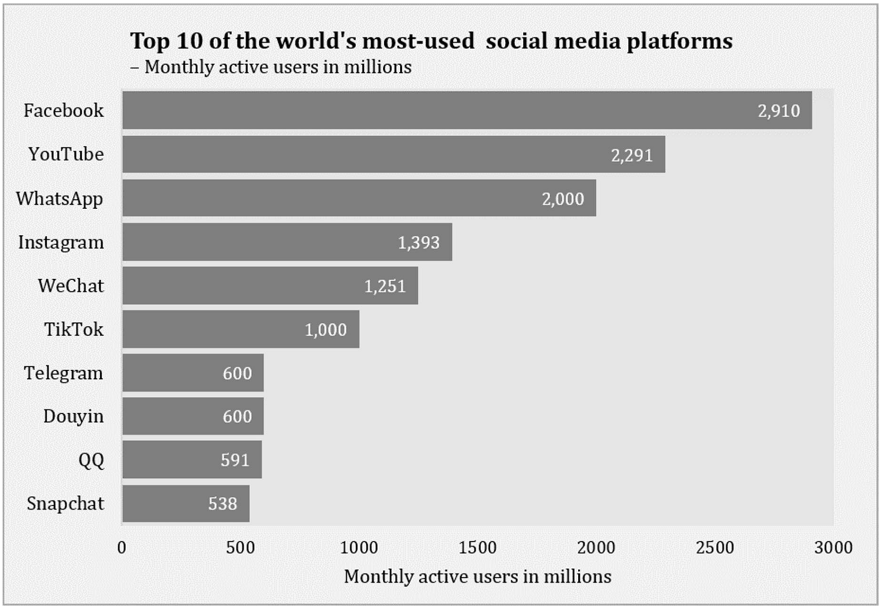
Background

Social Media Platforms

The rise and spread of SM platforms is a complex and important social and cultural phenomenon of the twenty-first century (for a detailed overview of the history and development of SM, see, e.g., Dijck, 2013; Boyd & Ellison, 2008). SM platforms employ a variety of interactive and computer-based technologies that enable users to share information, personal messages, and other content, such as videos and images, in communities and networks that they create themselves. Obar and Wildman (2015) describe the four main features of SM platforms. They are (1) interactive web-based Web 2.0 applications that consist of (2) user-generated content that is shared via (3) service-specific profiles created by the users themselves and through which they (4) connect with others, thereby developing social networks.

The various SM platforms offer users different forms of expression, including news feeds on Facebook and Twitter, discussion forums on Reddit, live streams on Instagram and YouTube, private messages on WeChat and WhatsApp, and videos on TikTok. Figure 1 presents an extrapolation of the top 10 most widely used SM platforms by people all over the world. Based on the latest figures reported by Kepios et al. (2021),[1] Internet users worldwide spend an average of 2 hours and 27 minutes per day using SM. Furthermore, approximately 57.6% of the global population is represented on at least one SM platform. This creates abundant opportunities to connect with the members of a population and recruit them to participate in online surveys.

[1] Since almost none of the SM platforms regularly publish figures on their active user base, most insights come from projections based on the platforms' self-serving advertising systems. The data reported here are based on the metadata report by Kepios et al. (2021). Furthermore, please note that users do not necessarily represent unique individuals, as it cannot be ruled out that multiple or fake accounts are included in the extrapolation.



Sources: Facebook, as of September 30, 2021 (Facebook, October 25, 2021); Telegram, as of November 8, 2021 (Telegram, November 8, 2021); all other metrics, as of October 17, 2021 (Kepios Pte. Ltd. et al., 2021).

Figure 1 The World's Top 10 Most-Used Social Media Platforms

In addition, some studies have found differences in the composition of users of different SM. For example, Hargittai (2020) found that younger populations (ages 18-34) were more likely to be active on platforms such as Facebook, Reddit, Twitter, and Tumblr. Women were slightly more likely to use Facebook, whereas men were significantly more likely to use Reddit. More highly educated populations tended to use Twitter more frequently. Hellemans et al. (2020) found the same age effects and gender differences, reporting more male users on Twitter and more female users on Instagram and Facebook. Understanding patterns of SM use on different platforms is essential for recruiting survey participants. Before deciding to use a particular SM platform, it is crucial first to understand which population groups are more strongly represented on which SM and which are unlikely to be reached.

Recruiting Population Members via Social Media

SM serves as a recruitment tool for researchers by allowing the creation and placement of content or advertising designed to reach target audiences. This can be accomplished through various approaches, which can be broadly divided into unpaid and paid strategies.[2]

Unpaid recruitment strategies encompass a variety of approaches. One of these is to reach potential participants and share survey invitations through groups. These may be existing groups that are thematically suited to the planned survey (e.g., Zimmer & Imhoff, 2020) or new groups or communities created explicitly to recruit target populations (e.g., Brickman Bhutta, 2012). Groups can also serve as a starting point for private messaging: Group members who are identified as potential survey participants can be contacted via private message and sent a survey invitation (e.g., Pagoto et al., 2014).

Another approach is to use profile pages in SM networks. Here as well, one can either use existing content or create new content. When working with other institutions to conduct a survey, the partner institutions can share an invitation link on their profile pages (e.g., Al-Shaqsi et al., 2020). The invitation can also be shared on a page created specifically for the survey project and maintained by the project team. When using SM platforms that are primarily based on visual content, such as Instagram and TikTok, it is often convenient to publish videos inviting people to participate. These videos may be posted for potential participants to find in the "explore" section of SM platforms, alongside an array of other videos that have been shared publicly. Such videos often introduce the survey and invite SM users to participate. Rather than including the survey link, they usually include a note that the link for participation can be found in the profile description of the account that posted the video.

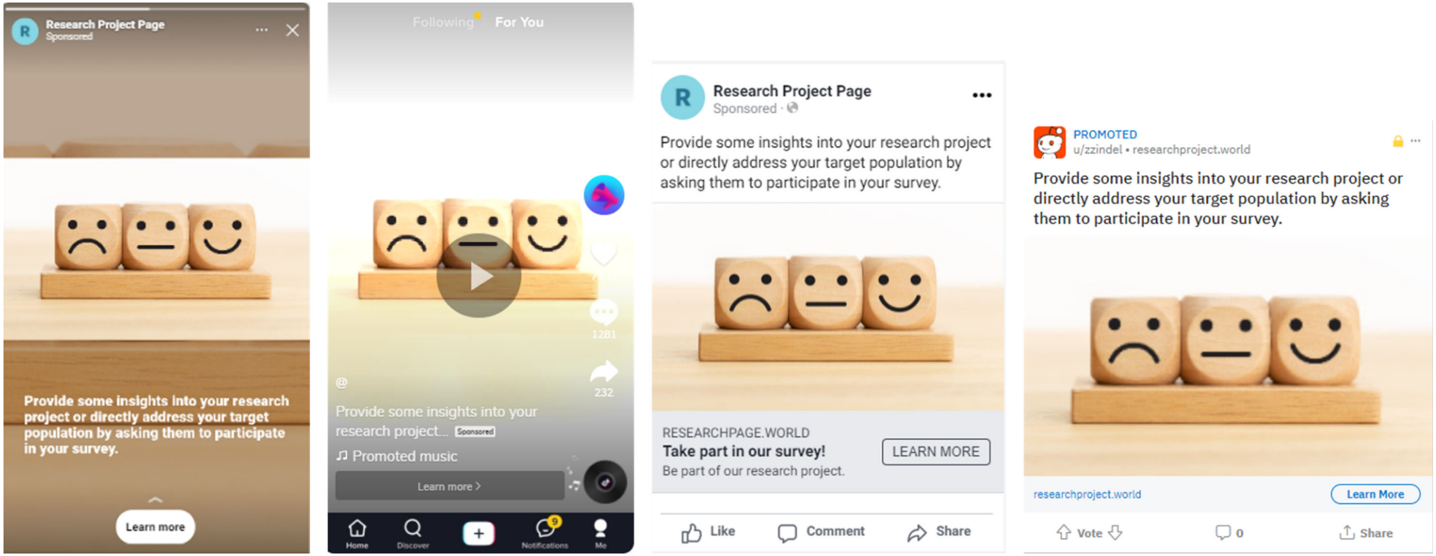
Paid strategies make use of promotion options and SM advertising. Most SM platforms currently offer their services to users free of charge and rely on an advertising revenue model. Researchers can purchase advertising on the platforms for a limited period and either promote existing content (e.g., Barnes et al., 2021) or place new ads for the research project at hand. Most SM platforms provide a sophisticated advertising targeting system that allows specific audiences to be identified based on multiple parameters, such as demographic characteristics, interests, or behaviors (i.e., digital activities, device usage, purchase behavior, etc.) (e.g., Meta Inc.; Twitter Inc.). These targeting options are the result of both the data entered by users on their own profiles as well as the behavior of the users on the platforms. The targeting parameters can be used to customize ads to reach very specific or rare populations. In addition, ads can be placed in different positions on a site depending

[2] It is worth noting that the distinction between paid and unpaid advertising is not clear-cut. The paid approaches are based on purchasing advertising space on the platforms. Nevertheless, in most cases, users have the option of sharing ads and promoted content in their own networks or, for example, on their profiles. Promoted content is also not necessarily created for this purpose but may already exist prior to the use of a paid strategy and thus already have reached SM users.

on the platform and end device of the target group – in the newsfeed, at the edge of the screen, or between "stories" (i.e., user-generated videos or images only visible for a limited period of time, usually 24 hours; see Figure 2). For a more detailed description of ad design on SM, see, Pötzschke & Braun (2017). Finally, ads can either link directly to an external survey website or point the user to an SM profile page that contains a link to the survey.

SM sites differ in several respects that strongly influence the conditions under which they might be suitable. For example, whereas Facebook allows for all the advertising options mentioned above, platforms like Instagram and TikTok do not have topic-specific groups that could be used for recruitment. Moreover, while Facebook, Instagram, and WeChat have very detailed demographic targeting options, Reddit and Twitter, for example, provide only a minimum amount of demographic information. Additionally, registration standards vary widely between platforms. Whereas platforms like Facebook, TikTok, and WeChat require detailed verification of new accounts, others like Twitter or Reddit do not, leading to a potential disparity between the number of accounts and the number of actual users. Furthermore, behavioral norms, site rules, and opportunities for different types of targeting vary with the current state of algorithmic updates, both across SM sites and over time. For an extensive overview of the different paid and unpaid strategies as well as the targeting options available on a selection of SM platforms, see Table 1.

Besides the varying characteristics of specific platforms that influence the use of SM platforms as recruitment tools, several other factors should also be considered when using SM strategies. In general, SM platforms offer both advantages and disadvantages in recruiting survey respondents, especially in comparison to more established offline or online methods. Table 2 provides an overview of the regularly cited advantages and disadvantages of SM recruitment. Where available, empirical evidence for the respective statements is given.



Research Project Page
Sponsored

Provide some insights into your research project or directly address your target population by asking them to participate in your survey.

Provide some insights into your research project... Sponsored

Promoted music

Research Project Page
Sponsored ·

Provide some insights into your research project or directly address your target population by asking them to participate in your survey.

RESEARCHPAGE.WORLD
Take part in our survey!
Be part of our research project.

PROMOTED
u/zzindel • researchproject.world

Provide some insights into your research project or directly address your target population by asking them to participate in your survey.

researchproject.world

Note. From left to right: story ad on Instagram, video ad on TikTok, ad in the Facebook news feed, promoted post on Reddit. Source: Own creation.

Figure 2 Examples of Ads on Instagram, TikTok, Facebook, and Reddit

Table 1 Overview of the Different Paid and Unpaid Recruitment Strategies as well as Targeting Options for Selected SM Platforms

Social media platform	Unpaid strategies	Paid strategies	Targeting options
Facebook Meta Platforms February 4, 2004	▪ Via posts in own or other groups ▪ Via profile post ▪ Via private message	▪ Paid ads placed at various positions on the platform ▪ Boosted post ▪ Boosted page	▪ Location (very detailed, to within one mile of specific coordinates) ▪ Demographics (e.g., age, gender, language, education, work, marital status) ▪ Interests ▪ Platform behaviors ▪ Connections (through own Facebook page or events) ▪ Devices
Instagram Meta Platforms October 6, 2010	▪ Via link in profile bio ▪ Via private message	▪ Paid ads at various positions within the platform ▪ Boosted post	▪ Location (very detailed, to within one mile of specific coordinates) ▪ Demographics (e.g., age, gender, language, education, work, marital status) ▪ Interests ▪ Platform behaviors ▪ Connections (through own Facebook page or events) ▪ Devices
Twitter Twitter Inc. July 15, 2006	▪ Via individual tweet ▪ Via private message to followers	▪ Paid Tweets at various positions within the platform (Tweet ads) ▪ Promoted Twitter accounts (Follower ads) ▪ Promoted trends (Trend Takeover)	▪ Location (detailed down to city and postal or zip code level) ▪ Demographics (age, gender, language) ▪ Devices ▪ Keywords ▪ Interests ▪ Platform behaviors ▪ Conversion topics

Social media platform	Unpaid strategies	Paid strategies	Targeting options
Reddit Advance Publications June 23, 2005	▪ Via post in subreddit ▪ Via link in profile bio ▪ Via private message	▪ New created user posts or videos (promoted post) ▪ Boosted post (organic post) ▪ Reddit takeover ads (promoted posts, banner ads)	▪ Location at country level (also county level in the USA) ▪ Interests ▪ Specific communities (= subreddits) ▪ Devices
WeChat Tencent Holdings Limited January 21, 2011	▪ Via groups ▪ Via private message	▪ Sponsored posts on WeChat Moments ▪ Banner ads within WeChat articles ▪ Mini Program ads ▪ WeChat influencer collaboration	▪ Location ▪ Demographics (e.g., gender, age, marital status, education level) ▪ Interests ▪ Platform behaviors ▪ Devices
TikTok ByteDance September 2016	▪ Via private message to accounts the user follows ▪ Via link in profile bio	▪ Paid ads on TikTok's "For You" page ▪ Promoted videos	▪ Location (down to level of states or large metropolitan areas depending on the country) ▪ Demographics (age, gender, language) ▪ Interests ▪ Platform behaviors ▪ Devices

Table 2 Advantages and Disadvantages of the Use of Social Media for Survey Recruitment

Advantages	Statement	Empirical evidence (examples)
Costs	SM strategies are inexpensive.	Ali et al., 2020; Webler et al., 2020; Batterham, 2014.
Reach	SM enables recruitment of a larger number of participants.	Admon et al., 2016; Bennetts et al., 2019; Samuels & Zucco, 2014.
Variety of users	SM allows collection of data from a broad range of participants.	Chard et al., 2018; Perrotta et al., 2021; Pötzschke & Braun, 2017.
Fast turnaround	SM strategies recruit survey participants quickly.	Guillory et al., 2018; Reuter et al., 2019; Zhang et al., 2020.
Targeting options	Targeted SM strategies allow very specific audiences to be reached.	Guillory et al., 2018; Harfield et al., 2021; Pötzschke & Braun, 2017.
Follow-up	SM strategies may provide the option to easily (re-) contact participants for follow-up studies.	Bolanos et al., 2012; Ersanilli & van der Gaag, 2022.
Disadvantages	Statement	Publication
Under-coverage bias	SM strategies have no chance of reaching all target population members.	Bennetts et al., 2019; Lehdonvirta et al., 2021; Rosenzweig & Zhou, 2021.
Over-coverage bias	SM strategies have a (high) risk of reaching large numbers of invalid accounts and lead to a high number of duplicate responses and fraudulent enrollments.	Pozzar et al., 2020; Quach et al., 2013; Yuan et al., 2014.
Self-selection bias	SM strategies may reach participants who differ systematically from non-participants.	Canan et al., 2021; Lehdonvirta et al., 2021; Williamson & Malik, 2021.
Selection bias	Targeted SM strategies are influenced by unknown mechanisms of advertising algorithms that allocate the advertised content.	No empirical evidence.

Review Methodology

A systematic search was applied to identify the relevant literature. The overall goal was to identify research articles that used SM platforms to recruit respondents for social-science-related online surveys. For this, the Web of Science database was used to access the Social Sciences Citation Index (SSCI), a multidisciplinary citation database specifically focusing on journals in various disciplines of the social sciences. The search was conducted on October 5, 2021, and used a combination of the following search terms with the Boolean operator "OR" and then combined with the Boolean operator "AND":

> (("recruit*") OR ("participant recruit*") OR ("recruit* strategies") OR ("social media recruit*") OR ("online sampling") OR ("survey sampling")) AND (survey) AND (("social media") OR ("social network*") OR ("social networking") OR (Facebook) OR (Instagram) OR (YouTube) OR (WhatsApp) OR (Tumblr) OR (Twitter) OR (Myspace) OR (Snapchat) OR (TikTok) OR (Vimeo) OR (Flickr) OR (Clubhouse) OR (Reddit) OR (4chan) OR (8chan) OR (8kun) OR (Telegram) OR (LinkedIn) OR (Pinterest) OR (Badoo) OR (QZone) OR ("Sina Weibo") OR (WeChat) OR ("Tencent Weibo") OR (Youku) OR (Vkontakte) OR (Twitch) OR (Xing) OR (Kuaishou) OR (Douyin) OR (WEIXIN))

The search field was limited to the topics category, meaning the search terms could only appear within the title, abstract, authors' keywords, and the databases' "keywords plus" category. Furthermore, only papers published in scientific journals and written in English were considered relevant. The publication period was defined to begin January 1, 2002, one year before SM hit the mainstream (Boyd & Ellison, 2008), and to end on October 5, 2021, to encompass a wide range of applications. The resulting records (N=1,199) were imported into the Citavi literature management software for further data screening. Subsequently, I performed a two-stage screening procedure. The first step involved exclusion based on the information contained in the abstracts. The following exclusion criteria were applied: (1) the abstract did not mention a reference to social media recruitment at all, (2) the abstract did not mention survey recruitment, (3) the abstract mentioned an overall sample size n≤100, and (4) the abstract mentioned that the authors of the paper did not do their own data collection. Overall, 624 articles were excluded during the abstract screening, leaving 575 full texts for review.

In a second step, the full texts of the remaining articles were screened based on the same exclusion criteria as in the abstract screening, but here on a full-text level as well as based on three additional criteria: (5) the article did not specify the SM platform, (6) the article either did not distinguish participants recruited via SM from participants recruited via other strategies, or multiple SM platforms were grouped together into the same category, and (7) the article did not include enough relevant information to be included in at least one of the analyses in the literature review. In the second step, a further 481 articles were excluded. Additionally, 11 articles were excluded because of duplication of study results. Further, 10 articles were excluded due to a lack of data access.

A total of 73 journal articles covering a total of 83 separate studies remained for inclusion in the literature review (Online Appendix 1).[3] Finally, the articles were systematically searched for data such as the number of individuals recruited, recruitment performance metrics, and cost. Online Appendix 2 provides a flow chart showing the exclusion process (Online Appendix 2, Figure 1), an extensive summary of the studies included (Online Appendix 2, Table 1), as well as the URL to replicate the search.

Recruitment Effectiveness

The effective recruitment of participants and, consequently, a large analysis sample is essential for quantitative research. At the same time, the overall effectiveness of the recruitment strategy must be considered in the context of the target population and the study objective. Therefore, to assess the effectiveness of SM recruitment strategies, I reviewed the evaluations of effectiveness by the articles' authors and recorded the size of the samples recruited. Additionally, where feasible, I compared the effectiveness of the strategies used with other recruitment strategies. I considered a recruitment approach to be effective if the authors had found it to be sufficient for the purpose of their study. In addition, I considered a method to be more effective if it reached a larger percentage of respondents than another method.

Recruitment Costs

The effectiveness of a recruitment strategy is always influenced by its cost. In survey practice, many designs must be modified within cost constraints. There are usually limited resources available to conduct a survey, which inevitably affects the choice of recruitment method. By formally evaluating and comparing the costs of different recruitment methods, one can determine their overall effectiveness (Groves, 2004). I therefore assessed cost-effectiveness in terms of cost per participant, and compared this, where possible, to the costs of other recruitment methods.

[3] Please note that in the remaining sections of this paper, the articles by Batterham (2014), Brodovsky et al. (2018), Ford et al. (2019), Lee et al. (2020), and Sunderland et al. (2017) are each counted as a single article or as multiple studies, according to the conclusions drawn, as they present results from multiple studies. This brings the number of studies included in this literature review to 83 studies in 73 journal articles.

Representativeness

The effectiveness and costs of sample recruitment must be balanced against the ability of samples to represent the intended target population. In line with the concept of total survey error (Groves & Lyberg, 2010), various sources of representation error such as coverage error, (self-)selection error, and non-response error are to be expected in surveys. The same applies to surveys recruited through SM platforms. Due to the reduction of the sampling frame to SM users only and the selective nature of convenience sampling approaches, severe limitations on representativeness are to be expected with SM recruitment. Nonetheless, SM recruitment is used frequently with the aim of producing a representative sample. To clarify whether the SM samples matched population estimations, I compared demographic characteristics of the recruited participants to national data included in the articles.

Findings

The articles included in this literature review were published between 2011 and 2021, with the number of publications increasing steadily over the period. This trend highlights the growing scientific relevance of the topic and the urgent need to systematically investigate its potential for survey research.

The majority (n=52) of the included articles used the social networking site Facebook as their only recruitment tool. Fifteen articles used a combination of Facebook and other SM platforms, for example, Reddit (e.g., Cahill et al., 2019; Côté-Léger & Rowland, 2020), Instagram (e.g., Garey et al., 2020; Guillory et al., 2018), and Twitter (e.g., Cavallo et al., 2020; Yuan et al., 2014). Other articles relied solely on other platforms or used a combination of them. The large number of articles that used Facebook for recruitment indicates that this was the most popular SM site for recruiting participants, certainly due to the high prevalence of usage among the world population.

Most studies used at least one paid recruiting approach (n=65). Targeted ads were used in 60 cases, and three studies used untargeted ad space on SM platforms (Dean et al., 2012; Sullivan et al., 2011; Wagenaar et al., 2012a). One study each used the option of promoting a Facebook page (Ellis et al., 2018) and a Facebook post (Barnes et al., 2021), both of which were created specifically for the study purpose.

The remaining articles used unpaid strategies, such as posting in specific groups or communities (e.g., Arentz et al., 2021; Avery-Desmarais et al., 2021), publishing multiple posts or tweets on private or institutional profile pages (McRobert et al., 2018), and sending private messages to specific users (Barratt et al., 2015; McRobert et al., 2018).

Nineteen of the 73 articles combined SM recruitment with other recruitment approaches. Overall, 11 combined SM recruitment exclusively with other online recruitment methods, for example, the use of e-mail lists (e.g., Arentz et al., 2021; Harfield et al., 2021), online panels (Zhang et al., 2020; Guillory et al., 2016), or the crowdsourcing data acquisition platform Amazon Mechanical Turk (Reuter et al., 2019; Côté-Léger & Rowland, 2020). Three studies used a combination of online and offline approaches (Baxter et al., 2017; Barrat et al., 2015; McRobert et al., 2018). Another two used venue-based approaches (Admon et al., 2016; Guillory et al., 2018). Finally, one study used newspaper ads (Carter-Harris et al., 2016).

Most studies were conducted in the United States (n=37), followed by Australia (n=17). Three studies were conducted in Canada (Chu & Snider, 2013; Shaver et al., 2019; Archer-Kuhn et al., 2021), and one study each was conducted in Brazil (Samuels & Zucco, 2014), Egypt (Wiliamson et al., 2021), Jordan (Suliman et al., 2018), Malaysia (Shakir et al., 2019), Norway (Robstad et al., 2019), and Thailand (Khumsaen & Stephenson, 2017). A total of seven studies took a cross-national approach (e.g., Barratt et al., 2015; Chard et al., 2018), while another three recruited respondents across national boundaries (e.g., Ellis et al., 2018; Dean et al., 2012). Overall, 57 of the 73 selected articles focused on cross-regional populations within countries, and 16 on specific regions (e.g., Russomanno & Tree, 2020; Wilson et al., 2019).

The majority of studies focused on adult-aged participants (n=48). The rest targeted very specific age groups (e.g., 13-20 years, Ford et al., 2019; 55-77 years, Carter-Harris et al., 2016). In addition, most of the studies covered all genders (n=66). Seven targeted female participants only (e.g., Archer-Kuhn et al., 2021; Arentz et al., 2021), and eight focused on male respondents only (e.g., Seidler et al., 2021; Wagenaar et al., 2012a). A single study targeted transgender and gender nonconforming people (Russomanno & Tree, 2020).

Apart from basic demographic characteristics, most studies focused on specific target groups, for example, (ex-)smokers (e.g., Carter-Harris et al., 2016; Guillory et al., 2016), users of (illegal) drugs (e.g., Borodovsky et al., 2018; Daniulaityte et al., 2018), parents (e.g., Akard et al., 2015; Arcia, 2014), or certain ethnic groups (e.g., Admon et al., 2016; Harfield et al., 2021). A total of 24 studies focused on specific rare populations. Eleven of these studies focused on members of the LGBTQI* community (e.g., Mitchell & Petroll, 2012; Sharma et al., 2018), eight on patients with rare diseases (e.g., Chung et al., 2019; Woodward et al., 2016), two on specific professional groups (Robstad et al., 2019; Suliman et al., 2018), and one study each on indigenous populations (Harfield et al., 2021), victims of sextortion (Wolak et al., 2018), and parents of children with cancer (Akard et al., 2015). A further nine studies addressed hard-to-reach populations. Four of these studies targeted young smokers (e.g., Garey et al., 2020; Pepper et al., 2019), and one each focused on cannabis cultivators (Barratt et al., 2015), heavy-drinking smokers (Bold et al., 2016), mothers who had experienced domestic violence (Archer-Kuhn et al., 2021), men who seek help from mental health services (Seidler et al., 2021), and Polish migrants (Pötzschke & Braun, 2017).

Recruitment Effectiveness

The total number of participants recruited via SM ranged from one participant recruited using a single post on a LinkedIn profile (McRobert et al., 2018) to 71,612 participants recruited using Facebook ads for a cross-national survey (Perrotta et al., 2021). The wide variation can be attributed to the recruitment strategies used as well as the different target groups.

Overall, the majority of unpaid SM strategies resulted in sample sizes n≤100. This includes all strategies that involved posting the survey invitation on a (profile) page. This was the case for posts on personal pages (e.g., Facebook profile page: n=21, Côté-Léger & Rowland, 2020; LinkedIn: n=1, Google +: n=41, McRobert et al., 2018), as well as on pages created specifically for the study (Facebook page: n=100, McRobert et al., 2018). Other unpaid strategies, such as posting a home-made video on YouTube (n=7; Barratt et al., 2015) and direct messaging on Twitter (n=67; Barratt et al., 2015), also resulted in a comparatively small number of cases. In contrast, unpaid strategies that relied on the group structure of SM platforms performed better. Of a total of twelve articles reporting results for group strategies, eight achieved a case count above 100 participants.

Table 3 Efficiency of the Studies Identified

Publication			N recruited via SM	% via SM	Other recruitment methods
Paid strategies					
Admon et al., 2016		Facebook	1,178	84.32	*Venue-based* – clinic-based recruitment
Carter-Harris et al., 2016		Facebook	331[a]	91.69[a]	*Newspaper ads*
Guillory et al., 2016		Twitter	568	26.36	*Online-Panel* – Qualtrics' panel aggregator
Guillory et al., 2018		Facebook & Instagram	6,611	47.27	*Venue-based* – LGBT social venues via in-person intercept interviews
Harfield et al., 2021		Facebook & Instagram	2,003	73.53	*E-mail*
Reuter et al., 2019		Twitter	704	48.82	*Internet-mediated recruitment method* – Amazon Mechanical Turk
Samuels & Zucco, 2014		Facebook	3,212[b]	72.46[b]	*Not specified* – National Probability Sample / face-to-face survey
Thornton et al., 2016		Facebook	553	56.03	*Internet-mediated recruitment methods* – community research database, first-year psychology courses at the University of Newcastle, New South Wales, Australia
Wolak et al., 2018		Facebook	2,148[a]	91.90[a]	*Internet-mediated recruitment methods* – website, shared by members of an advisory panel, ads on Google searches
Zhang et al., 2020		Facebook	2,432[a]	64.37[a]	*Online-Panel* – Gfk Panel Provider
Paid and unpaid strategies					
Bennetts et al, 2019	paid	Facebook	3,440[a]	73.74[a]	*Internet-mediated recruitment methods* – ads in a popular online single-parent community, via e-mail
	unpaid		1,146[a]	24.57[a]	

Publication			N recruited via SM	% via SM	Other recruitment methods
Côté-Léger & Rowland, 2020	unpaid	Facebook	21[a]	1.95[a]	*Internet-mediated recruitment methods* – giveaway websites, Amazon Mechanical Turk, other not specified
		Reddit	98[a]	9.09[a]	
	paid	Facebook	626[a]	58.07[a]	
		Reddit	80[a]	7.42[a]	
Unpaid strategies					
Arentz et al., 2021		Facebook	311[a]	63.08[a]	*E-mail* – Polycystic Ovary Association of Australia (organization) – e-mail
Barratt et al., 2015		Facebook	1,087[b]	12.91[b]	*Internet-mediated recruitment methods* – e-mail/e-newsletter, user website/forum, online chat *Offline methods* – news article, referrals by friends, family & associates, flyers/posts, grower magazine, radio
		Twitter	67[b]	0.80[b]	
		YouTube	7[b]	0.08[b]	
Baxter et al., 2017		Facebook	17	5.31	*Internet-mediated recruitment methods* – direct mailing, e-mail, website, *Offline methods* – family, friends, clinic-based
McRobert et al., 2018		Twitter	552	28.83	*Internet-mediated recruitment methods* – e-mail, website placement, newsletter *Offline methods* – flyer ads, in-person survey invitations, postal research flyers
		Facebook	100	5.22	
		Google+	41	2.14	
		LinkedIn	1	0.05	
Robstad et al., 2019		Facebook	21	15.22	*E-mail* – *e-mail* list of nurses
Welton et al., 2020		Facebook	339	26.76	*Internet-mediated recruitment methods* – post on MedAdvisor

Note: a = completed interviews, b = eligible interviews.

Table 3 shows the eighteen studies that evaluated the effectiveness of SM recruitment using an additional method. Overall, the percentage of participants recruited via SM ranged from 5.31% (Baxter et al., 2017) to 91.90% (Wolak et al., 2018). The median percentage was 59.56%. Eleven of the 18 articles reported more than 50% recruitment via SM.

The comparison with other strategies highlights that most unpaid SM approaches were less effective. Only Arentz et al. (2021), comparing an invitation to participate in a pre-existing Facebook group to direct e-mail to members of an association, achieved a larger sample with SM (n=91; 90.10%). The other articles suggested that alternative recruitment approaches, such as ads on collaborating websites (Baxter et al., 2019), ads on mobile apps (Welton et al., 2020), or venue-based approaches (Robstad et al., 2019), were more effective in achieving a sufficient analytic sample for their study purpose.

Nevertheless, none of the articles concluded that recruitment via SM was not advisable overall. Some studies with small sample sizes using unpaid SM approaches combined these with more effective paid SM approaches (Bennetts et al., 2019; Côté-Léger & Rowland, 2020). Here, it is important to keep in mind that whereas unpaid recruitment strategies rely on sporadic releases of content on SM, paid recruitment strategies usually entail continuous promotion of content over a period of several days. It is therefore inevitable that the paid strategies perform better in terms of recruitment rates, as more people are exposed to the content overall. Other articles argue that the small SM samples nonetheless provide greater diversity to their study population (e.g., Baxter et al., 2017; Robstadt et al., 2019). Based on these findings, it can be concluded that unpaid SM approaches are less effective than paid approaches and other recruitment strategies overall. Still, they can serve as a complementary sampling method to expand the sample population in a cost-effective way.

The paid SM strategies reached a higher number of recruited individuals than the unpaid approaches. The median number of individuals reached through paid approaches was 2,003, with absolute numbers ranging from 154 to 144,034. The large difference in performance was due primarily to the duration of each recruitment strategy. Taking recruitment duration into account, the median number of individuals recruited per day was 35.51, again with wide variation in the number per day (range: 1.26 – 685.90).

The comparison with other strategies provides evidence that paid SM strategies may be advantageous over offline approaches. Three studies used targeted ads in combination with offline recruitment methods (Admon et al., 2016; Guillory et al., 2018; Carter-Harris et al., 2016). In all cases, the authors concluded that recruitment via SM was more effective than the offline approach.[4] Compared with other Internet-mediated approaches, results for paid strategies were more diverse but generally showed a positive trend toward SM recruitment. Out of eight studies, six reported higher rates of recruitment via social media. Of these, one study combined Facebook ads with an online panel (Zhang et al., 2020), and one combined Facebook and Instagram ads with recruitment via an e-mail list (Harfield et al., 2021). Four studies used a combination of various other Internet-mediated recruitment methods (Bennetts et al., 2019; Côté-Léger & Rowland, 2020; Thornton et al., 2016; Wolak et al., 2018). The two studies that showed lower recruitment rates for paid SM strategies combined Twitter ads with an online panel (Guillory et al., 2016) and Twitter ads with Amazon Mechanical Turk (Reuter et al., 2019). Guillory et al. (2016) aimed to recruit 190 participants using each of the applied recruitment approaches and therefore concluded that the SM sample was effective for their study purpose.

Only two of the 24 studies focusing on rare populations combined an SM strategy with another approach. Welton et al. (2020) used a mix of unpaid posts in three Facebook groups and posts in a health app to reach individuals with seizure disorders and epilepsy. Overall, 26.76% (n=339) of participants enrolled via Facebook. The authors concluded that the combination of the two strategies was effective in obtaining a more diverse sample of the target population. Guillory et al. (2018) compared targeted ads on Facebook and Instagram with in-person intercept recruitment in LGBT bars and nightclubs to reach 18-24-year-old LGBT individuals. They concluded that both virtual (n=6,611; 47.27%) and local venues (n=7,375; 52.73%) were highly effective in recruiting a sufficient number of participants. Although more respondents were recruited through social venues, the researchers argued that SM was more efficient. Time spent recruiting in venues was much higher, as it included training, travel time to and from recruitment venues, and time to recruit at locations. In contrast, SM recruitment only required ad placement before the self-selection of participants into the survey could begin. Thus, much less time was needed to generate a large sample. Finally, the outstanding success in reaching LGBT* individuals might be because these rare population groups are particularly active in social venues and on SM in connecting with other community members. In conclusion, it is reasonable to assume that highly connected and active subgroups can be reached effectively via SM.

[4] Although Guillory et al. (2018) recorded more participants with the venue-based approach, they concluded that ads on Facebook and Instagram were more effective because of the time savings.

Recruitment Costs

Of all studies that used at least partially paid strategies, 73 reported at least some information about recruitment costs. Table 4 contains an extensive overview of all reported financial and performance metrics. Given the variation in sample sizes, recruitment length, and SM strategies, it is not surprising that the overall cost varied widely. Total SM recruitment expenditures were given in 43 studies and ranged from \$50.20 (n=404; Dean et al., 2012) to \$10,388.17 (n=4,010; Lee et al., 2020 – Study 1).[5] The median total spent on recruitment via SM was \$812.03. The cost per click (CPC), which applies to any paid advertising or promotion, was determined by daily fluctuating bid prices, as SM platforms offer advertising slots based on a competitive bidding system. Advertisers can bid on limited slots, and thus, demand determines the performance of the ads. Thus, while researchers can set a budget for ad campaigns, they have no control over the number of clicks generated by an ad. The information on the average CPC, available in 25 studies, varied between \$0.02 (n=1,562; Shakir et al., 2019) to \$2.16 (n=2,432; Zhang et al., 2020), with a median CPC of \$0.36. Additionally, 39 studies reported the average cost per participant (CPP). The amount ranged from \$0.18 (n=6,602; Ali et al., 2020) to \$43.41 (n=661; Cavallo et al., 2020 – Twitter), and the median of CPP was \$4.33.

Moreover, four studies compared the costs of different recruitment methods. Batterham (2014) found the cost of recruiting by postal and telephone recruitment (CPP: \$13.56) to be significantly higher than for targeted ads on Facebook (CPP: \$1.07 in Study 1; \$7.09 in Study 2). Two studies came to a similar conclusion when comparing targeted ads on Facebook with venue-based recruitment (Admon et al., 2016: CPP: \$14.63 vs. \$23.51) and newspaper advertising (Carter-Harris et al., 2016: CPP: \$1.51 vs. \$40.8). The reasons were very high personnel and processing costs when recruiting offline. Here, SM approaches provide a clear advantage. Reuter et al. (2019) compared paid Tweet ads on Twitter with recruitment via Amazon Mechanical Turk to reach 500 participants with each approach. Again, the SM-based approach was more cost-effective (Overall cost: \$980 vs. \$3,500).

[5] In the figures reported in the following, the dollar sign refers to U.S. dollars.

Table 4 Recruitment Costs of the Identified Studies

Publication		Cost of SM recruitment in USD			Cost of other recruitment in USD		Platform performance				Recruit. length in days
		Overall	CPC	CPP	Overall	CPP	Impressions	Users reached	Clicks	Unique clicks	
Paid strategies											
Admon et al., 2016	Facebook	11,103.25*	NR	14.63*	4,466.68*	23.51*	NR	364,035	9,972	NR	27
Ahmed et al., 2013	Facebook	NR	0.67	20.14	NA	NA	36,154,610	469,678	8,339	7,940	134
Akard et al., 2015	Facebook	1,129.88	1.08	< 17.00	NA	NA	3,897,981	NR	1,050	NR	74
Ali et al., 2020	Facebook	906.00	0.09	0.18	NA	NA	NR	236,017	9,609	NR	10
Altshuler et al., 2015	Facebook	3,970.00	NR	3.00	NA	NA	NR	NR	8,673	NR	109
Arcia, 2014	Facebook	3,821.81	0.63	16.52	NA	NA	10,577,381	7,248,985	6,094	5,963	129
Batterham & Calear, 2021	Facebook	NR	NR	NR	NA	NA	NR	NR	7,174	NR	NC
Batterham, 2014	study 1 – Facebook	8,946.00[a]	NR	7.09[a]	215,982.00[a]	13,56[a]	NR	NR	12,773	NR	NC
	study 2 – Facebook	653.20[a]	NR	1.07[a]	215,982.00[a]	13,56[a]	NR	NR	NR	NR	NC
Bold et al., 2016	Facebook	480.89	0.27	4.37	NA	NA	102,697	NR	1,781	NR	14
Borodovsky et al., 2018	study 1 – Facebook	800.00	NR	NR	NA	NA	NR	168,894	3,708	NR	43
	study 2 – Facebook	809.00	NR	NR	NA	NA	NR	231,400	3,932	NR	28
	study 3 – Facebook	350.00	NR	NR	NA	NA	NR	126,945	5,480	NR	20
	study 4 – Facebook	293.00	NR	NR	NA	NA	NR	78,974	3,135	NR	6

		Cost of SM recruitment in USD			Cost of other recruitment in USD		Platform performance				Recruit. length in days
Publication		Overall	CPC	CPP	Overall	CPP	Impressions	Users reached	Clicks	Unique clicks	
	study 5 – Facebook	402.00	NR	NR	NA	NA	NR	68,525	2,599	NR	9
	study 6 – Facebook	377.00	NR	NR	NA	NA	NR	96,096	5,612	NR	7
Calear & Batterham, 2019	Facebook	NR	NR	NR	NA	NA	NR	NR	NR	7,174	NC
Carter-Harris et al., 2016	Facebook	500.00	NR	1.51	1,224.00	40.80	56,621	NR	NR	1,121	18
Cavallo et al, 2020	Facebook & Instagram	NR	0.81/ 1.32	33.82	NR	NR	1,027,738	NR	8,507	NR	273‡
	Twitter	NR	NR	43.41	NR	NR	2,998,715	NR	1,198	NR	273‡
Chard et al., 2018	Facebook	NR	NR	NR	NA	NA	NR	NR	NR	11,850	5-16
Chu & Snider., 2013	Facebook	1,053.91^c	0.30c	11.97^c	NA	NA	17,527,703	NR	3,440	NR	NC
Crosier et al., 2016	Facebook	2,150.00	0.17-0.36	8.14	NA	NA	186,430	199,928	NR	NR	39
Daniulaityte et al., 2018	Twitter	2,100.00	NR	NR	NA	NA	NR	NR	NR	NR	18
Ellis et al., 2018	Facebook	NR	NR	NR	NA	NA	NR	NR	243	NR	NC
Folk et al., 2020	Facebook	1,802.72	0.53	10.73	NA	NA	NR	500,208	NR	3,394	23
Ford et al., 2019	study 1 – Facebook	274.56	0.30	4.76$^+$	NA	NA	38,108	NR	915	NR	21
	study 1 – Instagram	267.26	0.33	4.76$^+$	NA	NA	2,222	NR	803	NR	20
	study 1 – Snapchat	400.00	0.25	4.76$^+$	NA	NA	114,200	NR	1,600	NR	10

		Cost of SM recruitment in USD			Cost of other recruitment in USD		Platform performance				Recruit. length in days
Publication		Overall	CPC	CPP	Overall	CPP	Impres-sions	Users reached	Clicks	Unique clicks	
	study 2 – Facebook	25.44	0.28	4.76[+]	NA	NA	5,401	NR	89	NR	26
	study 2 – Instagram	300.00	0.34	4.76[+]	NA	NA	98,982	NR	864	NR	26
	study 2 – Snapchat	674.00	0.37	4.76[+]	NA	NA	504,700	NR	1,818	NR	NC
Garey et al., 2020	Facebook	522.39	0.50	NR	NA	NA	345,223	56,459	1,054	4,902	64
	Instagram	2,084.82	0.33	NR	NA	NA	1,507,887	388,813	6,234	4,902	64
Guillory et al., 2016	Twitter	6,848.25	NR	NR	NR	NR	NR	590,954	2,691	NR	NR
Guillory et al., 2018	Facebook	NR	NR	NR	NR	NR	NR	324,959	7,249	NR	12
Harfield et al., 2021	Facebook & Instagram	631.90[a]	0.20[a]	NR	NA	NA	173,452	98,445	3,190	NR	‡
Khumsaen & Stephenson, 2017	Facebook	NR	NR	NR	NA	NA	154,210	NR	16,391	NR	14
Knapp et al., 2019	Facebook & Instagram	NR	NR	NR	NA	NA	NR	126,945	5,480	NR	20
Leach et al., 2019	Facebook	5,170.00	NR	0.68-4.86	NA	NA	NR	NR	NR	NR	NC
Lee et al., 2020	study 1 – Facebook	10,388.17[a]	NR	3.23	NA	NA	NR	413,742	15,291	NR	NC
	study 2 – Facebook	4,600.63	NR	1.10	NA	NA	NR	261,457	10,702	NR	NC
Manski & Kottke, 2015	Facebook	NR	0.52	5.98	NA	NA	NR	NR	3,720	NR	NC
Mitchell & Petroll, 2012	Facebook	NR	NR	NR	NA	NA	8,500,000	NR	NR	7,994	70[‡]

Publication		Cost of SM recruitment in USD			Cost of other recruitment in USD		Platform performance				Recruit. length in days
		Overall	CPC	CPP	Overall	CPP	Impressions	Users reached	Clicks	Unique clicks	
Nelson et al., 2014	Facebook	NR	NR	1.36	NA	NA	3,254,666	NR	2,440	NR	72
Obamiro et al., 2020	Facebook	NR	0.04°	1.40	NA	NA	590,757	136,640	9,627	NR	91.25‡
Pepper et al., 2019	Facebook	NR	NR	NR	NA	NA	NR	NR	25,730	NR	NC
	Instagram	NR	NR	NR	NA	NA	NR	NR	16,300	NR	NC
Perrotta et al., 2021	Facebook	NR	0.17	1.25	NA	NA	19,300^{+}	NR	NR	NR	15-37
Pötzschke & Braun, 2017	Facebook & Instagram	557	0.13/ 0.25	0.52	NA	NR	173,084	90,436	5,080	3,721	30
Ramo & Prochaska, 2012	Facebook	6,628.24	0.45	4.28	NA	NA	28,683,151	NR	14,808	NR	395‡
Reuter et al., 2019	Twitter	980.00	NR	NR	3,500.00	NR	NR	NR	NR	NR	17
Rosenzweig & Zhou, 2021	Facebook	1,959.84	NR	NR	NA	NA	2,730,047	1,337,866	31,263	NR	15
Rosso & Sharma, 2020	Facebook & Instagram	NR	NR	NR	NA	NA	680,290	NR	NR	3,849	NC
Russomanno & Tree, 2020	Facebook	NR	NR	NR	NA	NA	NR	NR	NR	742	NC
Salk et al., 2020	Facebook & Instagram	1,536.00	NR	NR	NA	NA	377,469	NR	8,747	NR	NC
Samuels & Zucco, 2014	Facebook	4,972.79	0.22	1.74*	NR	NR	47,100,000	4,600,000	NR	22,181	29
Shakir et al., 2019	Facebook	NR	0.02b	NR	NA	NA	5,282,661	1,652,361	114,054	NR	70
Sharma et al., 2018	Facebook	NR	NR	NR	NA	NA	352,997	NR	14,968	NR	NC
Shaver et al., 2019	Facebook	1,365.00^{c}	NR	1.30c	NA	NA	132,021	34,012	2,316	2,067	40
Sullivan et al., 2011	MySpace	NR	NR	NR	NA	NA	8,257,271	NR	30,559	NR	29

Publication		Cost of SM recruitment in USD			Cost of other recruitment in USD		Platform performance				Recruit. length in days
		Overall	CPC	CPP	Overall	CPP	Impressions	Users reached	Clicks	Unique clicks	
Sunderland et al., 2017	study 1 – Facebook	NR	NR	NR	NA	NA	NR	NR	39,945	NR	NC
	study 2 – Facebook	NR	NR	NR	NA	NA	NR	NR	7,174	NR	NC
Thornton et al., 2016	Facebook	975.83	0.44	1.86	NA	NA	4,106,729	NR	2,220	NR	34
Wagenaar et al., 2012[a]	MySpace	NR	NR	NR	NA	NA	8,257,271	NR	30,559	NR	29
Webler et al., 2020	Facebook	200.00	NR	0.53	NA	NA	NR	NR	NR	NR	7[‡]
Williamson & Malik, 2021	Facebook	NR	NR	NR	NA	NA	NR	4,057,249	NR	10,237	NC
Wilson et al., 2019	Facebook	NR	NR	0.41[a]	NA	NA	NR	NR	NR	NR	NC
Wolak et al., 2018	Facebook	NR	NR	NR	NR	NR	NR	1,370,802	NR	NR	NC
Woodward et al., 2016	Facebook	NR	NR	4.44[a]	NR	NR	NR	NR	NR	1,668	96
Zhang et al., 2020	Facebook	NR	2.16	4.05	NR	NR	NR	NR	NR	7,642	14[‡]
Paid and unpaid strategies											
Archer-Kuhn et al., 2021	Facebook – targeted ads	432.98	0.31	4.76	NA	NA	NR	42,488	1,375	521	122
Barnes et al, 2021	Facebook – boosted post	815.06[a]	NR	1.02[a]	NA	NA	71,787	88,650	1,739	NR	74
	Facebook – post & direct message	NA	NA	NA	NA	NA	22,482	14,542	156	NR	74

Publication		Cost of SM recruitment in USD			Cost of other recruitment in USD		Platform performance				Recruit. length in days
		Overall	CPC	CPP	Overall	CPP	Impressions	Users reached	Clicks	Unique clicks	
Bennetts et al, 2019	Facebook	5,658.17[a]	1.07[a]	1.65[a]	142.00[a]	NR	NR	446,787	8,364	NR	135
Cahill et al., 2019	Facebook	500.00	1.91[†]	3.16/3.13	NA	NA	NR	NR	NR	NR	14
	Twitter	500.00	9.90[†]	10.00/1.76	NA	NA	NR	NR	NR	NR	10
Chung et al., 2019	Facebook – promoted page	206.00	NR	2.92	NA	NA	NR	1,933	15	NR	35[‡]
	Facebook – targeted ads	176.00	NR	2.19	NA	NA	NR	NR	NR	699	35[‡]
Dean et al., 2012	Second Life	50.20	NR	NR	NA	NA	NA	NA	NA	NA	
Yuan et al., 2014	Facebook, Twitter, LinkedIn, Tumblr	5,021.00	0.64	3.56	NA	NA	NR	NR	10,006	NR	NC

Note. Results from Samules & Zucco (2014) were supplemented with data from Samules & Zucco (2013); results from Perrotta et al. (2021) were supplemented with data from Grow et al. (2020), and results from Ahmed et al. (2013) were supplemented with data from Fenner et al. (2012). CPC = cost per click; CPP = cost per completed; NR = not reported; NA = not applicable; NC = not computable; * = incl. other costs (e.g., incentives, salaries, pilot study costs, etc.); ° = cost per impression; † = cost per 1,000 impressions; + = over all platforms/studies; ‡ = own calculations (1 week = 7 days; 1 month = 30.42 days); a = AUD converted to USD at a rate of 0.71; b = MYR converted to USD at a rate of 0.24; c = CD converted to USD at a rate of 0.78; AUD converted to USD at a rate of 0.71.

Overall, few of the studies on rare or hard-to-reach populations provided information on costs. Of the 33 total studies, 25 used at least one paid SM strategy. However, only nine of the 25 studies included information on the exact cost of completed interviews (Akard et al., 2015; Archer-Kuhn et al., 2021; Bold et al., 2016; Chung et al., 2019; Crosier et al., 2016; Pötzschke & Braun, 2017; Ramo & Prochaska, 2012; Woodward et al., 2016; Yuan et al., 2014). All studies used targeted ads on Facebook as a paid recruitment strategy. The most cost-effective completed interviews were reported by Pötzschke and Braun (2017), who used a combination of Facebook and Instagram ads to recruit Polish migrants in four European countries ($0.52 per interview). The most expensive completed interviews were recorded by Akard et al. (2015), who used targeted advertising on Facebook to reach parents of children with cancer (just under $17.00 per interview). All nine studies concluded that the SM approaches were highly cost-effective. These findings suggest great potential for the cost-effective recruitment of rare or hard-to-reach population groups via SM. This is especially relevant in terms of cost planning for data collection, particularly since probabilistic sampling strategies require a very high number of attempts to contact these populations, which in turn greatly increases costs. On SM platforms, on the other hand, group structures and targeting options can be used to reach specific individuals in a targeted manner, making recruitment more cost-effective.

Social Media's Representation of the Population

Recruitment effectiveness and cost-effectiveness must be balanced against a sample's ability to represent an intended target population. Since samples recruited via SM are non-probability-based, target group members have unequal chances of being included. One of the biggest challenges in correcting this selectivity is the lack of available information about who actively decides not to participate in the survey. In most cases, data about the total population active on each SM platform is unavailable. Without this information, there is almost no way to make probabilistic inferences about the population. As a result, the conclusions of most SM samples cannot be readily extrapolated (Lehdonvirta et al., 2020).

However, the goal of such survey designs is not always to obtain a representative sample of respondents. All reviewed articles discussed the issue of the scope of the recruited sample, at least in terms of study limitations. Most concluded that their studies could not be generalized to the entire target population. None of the eight articles that used exclusively unpaid strategies included a comparison of the SM sample with known distributions of the target population. However, some of these studies aimed not to create a representative sample but rather to gain insight into an area of research (e.g., Avery-Desmarais et al., 2021).

More than two-thirds of the articles that used at least one paid strategy (n=44; 67.69%) concluded that the sample was not representative due to the SM population's unknown composition or the "black box" of advertising algorithms. The remaining 21 articles evaluated the representativeness of characteristics of the population of interest. Their findings were mixed overall. Most comparisons concluded that the samples were only partially representative. The characteristics most often described as imbalanced included age, gender, education level, and ethnicity/race. Table 5 provides a detailed listing of biased and unbiased demographics.

Age bias was reported in twelve articles. A total of seven articles reported the overrepresentation of young people or adolescents. Additionally, Batterham and Calear (2021) reported an underrepresentation of elderly populations. The biased estimates may be due to the comparatively younger SM population on most platforms. Furthermore, younger people generally spend more time on SM, which increases the chances of reaching this group. However, the findings of Ali et al. (2020) and Perrotta et al. (2021) differ. Both had a comparatively high proportion of older individuals in their samples. This could be due to the specific topic of the surveys: Ali et al. (2020) and Perrotta et al. (2021) surveyed beliefs and behaviors in the context of the COVID-19 pandemic. Because an infection poses a higher risk of severe complications, particularly for older adults, it is reasonable to assume that this group would have a higher interest in study participation.

Regarding the distribution of gender and education level, there were further limitations on representativeness. Five studies using Facebook for recruitment found an overrepresentation of female participants (Batterham & Calear, 2021; Batterham, 2014; Carter-Harris et al., 2016; Chung et al., 2019; Harfield et al., 2021). This may be because females tend to be more active on SM (Pew Research Center, 2015). Dean et al. (2012) found the opposite gender effect, that is, a higher number of males, in a sample recruited via Second Life. In addition, eight studies found a trend toward participants with higher levels of education (Ahmed et al., 2013; Ali et al., 2020; Bennetts et al., 2019; Carter-Harris et al., 2016; Nelson et al., 2014; Perrotta et al., 2021; Rosenzweig & Zhou, 2021; Zhang et al., 2020). All these studies used targeted ads on Facebook to recruit their participants.

Table 5 Check for Representativeness

Publication	Comparison	Sociodemographics	SM
Paid strategies			
Ahmed et al., 2013	Australian Census Data 2006	- age: younger age group (16-17 years) underrepresented - education: higher educational level overrepresented - geographic area: representative to comparison - socioeconomic status: representative to comparison	Facebook
Ali et al., 2020	U.S. Census 2018-2019	- gender: representative to comparison - age: younger adults underrepresented - education: higher education overrepresented - ethnicity/race: Non-Hispanic whites overrepresented	Facebook
Altshuler et al., 2015	U.S. Census	- gender: representative to comparison - age: younger age group (13-18 years) overrepresented - ethnicity/race: Hispanics/Latinos underrepresented; Blacks/African Americans overrepresented; individuals with two or more races overrepresented - geographic area: Southerners underrepresented; Westerners overrepresented	Facebook
Arcia, 2014	National Vital Statistics Reports 2010	- age: younger mothers overrepresented - ethnicity/race: Hispanics underrepresented	Facebook
Batterham & Calear, 2021	Australian population (not defined further)	- gender: females overrepresented - age: older age-group (65+ years) underrepresented	Facebook
Batterham, 2014	Australian Census Data & National Survey Data	- gender: females overrepresented - age: younger adults overrepresented; older adults underrepresented	Facebook
Borodovsky et al., 2018	U.S. Census 2015	- geographic area: representative to comparison	Facebook

Publication	Comparison	Sociodemographics	SM
Carter-Harris et al., 2016	parallel Newspaper recruitment (not representative for the U.S. population)	- gender: females overrepresented - education: high school graduates or higher overrepresented - ethnicity/race: non-Hispanic Caucasians overrepresented - geographic area: representative to zip code areas	Facebook
Daniulaityte et al., 2018	U.S. Census 2017	- ethnicity/race: representative to comparison	Twitter
Harfield et al., 2021	Australian Bureau of Statistics information 2016	- gender: females overrepresented - age: younger age group (16-19) overrepresented - ethnicity/race: Aboriginal and Torres Strait Islanders overrepresented - geographic area: urban young people overrepresented	Facebook & Instagram
Nelson et al., 2014	Census data Minneapolis/St. Paul	- education: higher education overrepresented	Facebook
Perrotta et al., 2021	Eurostat 2019, U.S. census 2018	- age: older people overrepresented in Italy and the United Kingdom - education: higher education overrepresented in Belgium, France, Spain, the United Kingdom, and the United States	Facebook
Rosenzweig & Zhou, 2021	Afrobarometer	- age: younger people overrepresented - education: higher educational level overrepresented - geographic area: urban areas overrepresented - socioeconomic status: wealthy adults overrepresented	Facebook
Seidler et al., 2021	Ten to men study cohort	- age: representative to comparison - ethnicity/race: Aboriginal or Torres Strait Islander men underrepresented	Facebook
Shaver et al., 2019	Census Data for Newfoundland and Labrador 2016	- age: older age group (60-64 years) overrepresented - geographic area: representative to comparison - socioeconomic status: lower income groups underrepresented	Facebook
Wagenaar et al., 2012b	U.S. Bureau & CIA information	- ethnicity/race: whites overrepresented in South Africa; Blacks overrepresented in the United States	Facebook

Publication	Comparison	Sociodemographics	SM
Zhang et al., 2020	American Community Survey 2016	- age: younger age-group (18-24 years) overrepresented - education: higher education levels overrepresented - ethnicity/race: whites underrepresented	Facebook
Paid and unpaid strategies			
Bennetts et al, 2019	Longitudinal Study of Australian Children (LSAC)	- education: higher educational level overrepresented - migration background: migration background underrepresented	Facebook
Chung et al., 2019	National Data of Kidney transplant recipients 2016	- gender: females overrepresented - ethnicity/race: African Americans underrepresented	Facebook
Dean et al., 2012	American Community Survey	- age: younger people overrepresented - gender: males overrepresented - ethnicity/race: Blacks underrepresented; Asians overrepresented	Second Life
Yuan et al., 2014	HIV Population in the U.S. (Centers for Disease Control and Prevention information)	- ethnicity/race: African Americans / Latinos underrepresented	Facebook, Twitter, LinkedIn, Tumblr

Studies reported mixed findings regarding the representation of ethnic groups. Ali et al. (2020), Arcia (2014), and Altshuler et al. (2015) found an underrepresentation of Hispanics and overrepresentation of non-Hispanics, respectively. Chung et al. (2019) and Yuan et al. (2014) found an underrepresentation of African Americans. Since all these studies used Facebook as their primary sampling frame, the results indicate a distinct limitation of this platform as a recruitment tool.

To increase the national representativeness of their surveys, Perrotta et al. (2021) and Zhang et al. (2020) applied weights to compare their samples to population data. Perrotta et al. (2021) used a post-stratification weighting approach, and Zhang et al. (2020) used an inverse probability weighting approach. These procedures led, at least partially, to corrected results comparable to population data. Appropriate weighting strategies might thus increase the quality of SM samples. Nevertheless, these two examples should not be taken as irrefutable proof of the effectiveness of weighting methods in obtaining representative results. Despite weighting strategies, bias was still found in both studies.

Discussion

This literature review synthesized the available evidence on the strengths and weaknesses of SM as a recruitment tool for online surveys. The majority of studies included in this review concluded that recruitment via SM was an effective method for their study purposes. In particular, studies comparing SM and offline strategies showed SM to have the advantages of a wide reach and the ability to reach audiences. In addition, studies comparing SM and other online strategies showed that the options of targeting and promoting ads in SM were beneficial. Unpaid SM strategies, on the other hand, tended to be less effective than other approaches. Nevertheless, unpaid strategies should not be considered generally ineffective. Indeed, the review showed that these approaches could be used effectively to complement other recruitment strategies to reach specific subgroups of a target population.

Beyond that, studies showed that SM recruitment was effective for reaching rare populations. This is one of the most significant advantages of these recruitment strategies. Due to the many SM platforms and numerous daily online user interactions, researchers can reach even very rare populations at a scale sufficient for their study purposes. The group structures, as well as the extensive targeting options, enable targeting of even very precisely defined populations. Thus, SM strategies offer a recruitment option for cases in which probabilistic sampling methods meet their limits due to financial, personnel, or time constraints, as well as a lack of sampling frames. The benefits were particularly evident for the recruitment of LGBT people, who tend to be highly connected through SM.

The costs of SM approaches varied widely across studies. Most studies reported SM strategies to be very cost-effective, but the costs depended heavily on the target audience. Comparisons of SM recruitment with other recruitment approaches highlighted the advantages of SM in terms of recruitment costs. Likewise, studies focusing on rare or hard-to-reach populations illustrated the argument for high cost-effectiveness of SM recruitment, as probabilistic recruitment strategies to reach these groups would involve a high number of contact attempts and would thus inevitably increase costs.

Finally, most studies that aimed for representative results showed bias in some sociodemographic variables. Study results showed, for example, disproportionate percentages of women and highly educated individuals in samples recruited via Facebook. A remaining problem is the lack of control when relying on paid SM strategies and allocation algorithms. The SM platforms covered in this review are not transparent as to the sum of all underlying decision-making mechanisms that influence the placement of ads or promoted content. Potential selectivity bias cannot be ruled out without further insight into the allocation mechanism. However, many of the studies reviewed did not aspire to generalize their results, arguing that the results were not transferable or valid outside a narrow framework.

Unpaid strategies such as posts in groups or communities can be used in the form of an online venues-based approach to reach certain subgroups. Direct messages offer a digital version of an outreach event. Finally, paid SM recruitment strategies allow monitoring of participant demographic characteristics and can be used to target population members accordingly.

Additionally, this literature review produced some more general findings. Several scientific papers lacked sufficient documentation. Only if the recruitment process is transparent can results be interpreted in a real context, making follow-up research or reproducible studies possible. Furthermore, many studies lacked reflection on the quality of the data obtained. While most studies described the lack of representativeness, few commented on the impacts of, for example, the devices used or the risk of falsified or faked responses.

The Internet and SM have a significant influence on survey research. The ongoing growth of the Internet and the increasing number of SM users offer great potential for future participant recruitment. It is essential to continue research in this area and (critically) reflect on new developments to ensure and update scientific standards accordingly.

Directions for Future Research

The literature review highlighted several areas for future research. Only a fraction of the studies included using paid strategies explicitly reported performance metrics for the individual advertisements. Therefore, the question of what types of ad design were more appealing to potential participants could not be answered in most cases. Future studies should explicitly address the performance of individual ads and the effect of design differences between ads.

The results of many of the articles included in this review cannot be generalized beyond specific populations. Only in a few cases, when controlling for several indicators and using probability-based survey data and a known population base, can the results be reasonably generalized. All of the studies considered reported at least a low level of bias. Further research is needed to systematically address whether SM can be used at all to recruit representative samples and, if so, which SM platforms and strategies are the most suitable for this purpose. Furthermore, to comprehensively describe the representativeness of samples recruited via SM, future studies should explicitly include parameters matching census data or national surveys in the questionnaire to allow for comparability.

When using SM for survey recruitment, users need to see and read the invitation to join the online survey. Without information on the group that has been exposed to the invitation, it is impossible to determine whether representativeness problems were due to algorithm allocation processes that caused underrepresented groups not to see the ads, or whether underrepresented groups simply did not want to participate. To date, little research exists on the perception of ads and promoted content on SM platforms. A future approach could be to study attention to ads on SM through, for example, eye-tracking tests.

In general, only a few articles covered in this review addressed the possibility of fraudulent enrolment when recruiting survey participants via SM. Since many of the SM platforms use limited account validation measures, there is always a risk of multiple participation and intentional falsification of survey data. Furthermore, it cannot be ensured that participants originated from the platforms. Since recruitment is uncontrolled, survey links can be shared and distributed outside the platforms. Further work is needed to evaluate methods to ensure data authenticity, such as tracking IP addresses or referral URLs, to investigate participant conversion patterns further.

Few studies mentioned the use of incentives to recruit respondents via SM. There is a tension between the use of incentives and the simultaneous risk of generating a high proportion of fraudulent interviews. Future research should test incentivization methods for SM surveys, taking the resulting data quality into consideration. In addition, only one study incorporated incentive costs into expenditures. However, the use of incentives could have an impact on the evaluation of cost-effectiveness. This is where further research could come in and examine whether the argument for cost-effective recruitment remains valid when incentive costs are considered.

References

Admon, L., Haefner, J. K., Kolenic, G. E., Chang, T., Davis, M. M., & Moniz, M. H. (2016). Recruiting Pregnant Patients for Survey Research: A Head to Head Comparison of Social Media-Based Versus Clinic-Based Approaches. *Journal of Medical Internet Research*, *18(12)*. doi:10.2196/jmir.6593

Ahmed, N., Jayasinghe, Y., Wark, J. D., Fenner, Y., Moore, E. E., Tabrizi, S. N., Fletcher, A., & Garland, S. M. (2013). Attitudes to chlamydia screening elicited using the social networking site Facebook for subject recruitment. *Sexual Health*, *10*(3), 224–228. doi:10.1071/SH12198

Akard, T. F., Wray, S., & Gilmer, M. J. (2015). Facebook Advertisements Recruit Parents of Children With Cancer for an Online Survey of Web-Based Research Preferences. *Cancer Nursing*, *38*(2), 155–161. doi:10.1097/NCC.0000000000000146

Ali, S. H., Foreman, J., Capasso, A., Jones, A. M., Tozan, Y. & DiClemente, R. J. (2020). Social media as a recruitment platform for a nationwide online survey on COVID-19 knowledge, beliefs, and practices in the United States: methodology and feasibility analysis. *BMC Medical Research Methodology, 116*. doi: 10.1186/s12874-020-01011-0

Al-Shaqsi, S. Z., Rai, A., Forrest, C., & Phillips, J. (2020). Public Perception of a Normal Head Shape in Children With Sagittal Craniosynostosis. *Journal of Craniofacial Surgery*, *31*(4), 940–944. doi:10.1097/SCS.0000000000006260

Altshuler, A. L., Storey, H. L., & Prager, S. W. (2015). Exploring abortion attitudes of US adolescents and young adults using social media. *Contraception*, *91*(3), 226–233. doi:10.1016/j.contraception.2014.11.009

Archer-Kuhn, B., Beltrano, N. R., Hughes, J., Saini, M., & Tam, D. (2021). Recruitment in response to a pandemic: pivoting a community-based recruitment strategy to facebook for hard-to-reach populations during COVID-19. *International Journal of Social Research Methodology*, pp. 1–12. doi:10.1080/13645579.2021.1941647

Arcia, A. (2014). Facebook Advertisements for Inexpensive Participant Recruitment Among Women in Early Pregnancy. *Health Education & Behavior*, *41*(3), 237–241. doi:10.1177/1090198113504414

Arentz, S., Smith, C. A., Abbott, J., & Bensoussan, A. (2021). Perceptions and experiences of lifestyle interventions in women with polycystic ovary syndrome (PCOS), as a management strategy for symptoms of PCOS. *BMC Women's Health*, *21*(1). doi:10.1186/s12905-021-01252-1

Avery-Desmarais, S. L., McCurry, M. K., Sethares, K. A., Batchelder, A., & Stover, C. (2021). Internet Recruitment of a Diverse Population of Lesbian, Gay, and Bisexual Nurses in a Study of Substance Use and Minority Stress. *Journal of Transcultural Nursing*, 10436596211042071. doi:10.1177/10436596211042071

Bach, R. L., Kern, C., Amaya, A., Keusch, F., Kreuter, F., Hecht, J., & Heinemann, J. (2021). Predicting Voting Behavior Using Digital Trace Data. *Social Science Computer Review*, *39*(5), 862–883. doi:10.1177/0894439319882896

Barnes, L. A., Barclay, L., McCaffery, K., Rolfe, M. I., & Aslani, P. (2021). Using Facebook to recruit to a national online survey investigating complementary medicine product use in pregnancy and lactation: A case study of method. *Research in Social & Administrative Pharmacy*, *17*(5), 864–874. doi:10.1016/j.sapharm.2020.07.011

Barratt, M. J., Potter, G. R., Wouters, M., Wilkins, C., Werse, B., Perala, J., Pedersen, M. M., Nguyen, H., Malm, A., Lenton, S., Korf, D., Klein, A., Heyde, J., Hakkarainen, P., Frank, V. A., Decorte, T., Bouchard, M., & Blok, T. (2015). Lessons from conducting trans-national Internet-mediated participatory research with hidden populations of cannabis cultivators. *International Journal of Drug Policy*, *26*(3), 238–249. doi:10.1016/j.drugpo.2014.12.004

Batterham, P. J. (2014). Recruitment of mental health survey participants using Internet advertising: content, characteristics and cost effectiveness. *International Journal of Methods in Psychiatric Research*, *23*(2), 184–191. doi:10.1002/mpr.1421

Batterham, P. J., & Calear, A. L. (2021). Incorporating psychopathology into the interpersonal-psychological theory of suicidal behavior (IPTS). *Suicide and Life-Threatening Behavior*, *51*(3), 482–491. doi:10.1111/sltb.12727

Baxter, M., Erby, L., Roter, D., Bernhardt, B. A., Terry, P., & Guttmacher, A. (2017). Health screening behaviors among adults with hereditary hemorrhagic telangiectasia in North America. *Genetics in Medicine*, *19*(6), 659–666. doi:10.1038/gim.2016.161

Bennetts, S. K., Hokke, S., Crawford, S., Hackworth, N. J., Leach, L. S., Nguyen, C., Nicholson, J. M., & Cooklin, A. R. (2019). Using Paid and Free Facebook Methods to Recruit Australian Parents to an Online Survey: An Evaluation. *Journal of Medical Internet Research*, *21*(3), e11206. doi:10.2196/11206

Bolanos, F., Herbeck, D., Christou, D., Lovinger, K., Pham, A., Raihan, A., Rodriguez, L., Sheaff, P., & Brecht, M. L. (2012). Using facebook to maximize follow-up response rates in a longitudinal study of adults who use methamphetamine. *Substance abuse: research and treatment*, *6*, pp. 1–11. doi:10.4137/SART.S8485

Bold, K. W., Hanrahan, T. H., O'Malley, S. S., & Fucito, L. M. (2016). Exploring the Utility of Web-Based Social Media Advertising to Recruit Adult Heavy-Drinking Smokers for Treatment. *Journal of Medical Internet Research*, *18*(5). doi:10.2196/jmir.5360

Borodovsky, J. T., La Marsch, & Budney, A. J. (2018). Studying Cannabis Use Behaviors With Facebook and Web Surveys: Methods and Insights. *JMIR Public Health and Surveillance*, *4*(2), 370–383. doi:10.2196/publichealth.9408

Boyd, D. M., & Ellison, N. B. (2008). Social Network Sites: Definition, History, and Scholarship. *Journal of Computer-Mediated Communication*, *13*, pp. 210–230. doi:10.1111/j.1083-6101.2007.00393.x

Brickman Bhutta, C. (2012). Not by the Book: Facebook as a Sampling Frame. *Sociological Methods & Research*, *41*(1), 57–88. doi:10.1177/0049124112440795

Cahill, T., Wertz, B., Zhong, Q. K., Parlato, A., Donegan, J., Forman, R., Manot, S., Wu, T. Y., Xu, Y. Z., Cummings, J. J., Cunningham, T. N., & Wang, C. (2019). The Search for Consumers of Web-Based Raw DNA Interpretation Services: Using Social Media to Target Hard-to-Reach Populations. *Journal of Medical Internet Research*, *21*(7). doi:10.2196/12980

Calear, A. L., & Batterham, P. J. (2019). Suicidal ideation disclosure: Patterns, correlates and outcome. *Psychiatry Research*, *278*, 1–6. doi:10.1016/j.psychres.2019.05.024

Canan, S. N., Kaplan, A. M. & Jozkowski, K. N. (2021). Comparing Rates of Sexual Assault Between Panel Quota and Social Media Samples: Findings Across Sexual Orientation Categories. *Journal of Interpersonal Violence*. doi:10.1177/0886260521105627

Carter-Harris, L., Ellis, R. B., Warrick, A., & Rawl, S. (2016). Beyond Traditional Newspaper Advertisement: Leveraging Facebook-Targeted Advertisement to Recruit Long-Term Smokers for Research. *Journal of Medical Internet Research*, *18*(6). doi:10.2196/jmir.5502

Cavallo, D., Lim, R., Ishler, K., Pagano, M., Perovsek, R., Albert, E., Gonzalez, S. K., Trapl, E., & Flocke, S. (2020). Effectiveness of Social Media Approaches to Recruiting

Young Adult Cigarillo Smokers: Cross-Sectional Study. *Journal of Medical Internet Research*, *22*(7). doi:10.2196/12619

Chard, A., Metheny, N. S., Sullivan, P. S., & Stephenson, R. (2018). Social Stressors and Intoxicated Sex Among an Online Sample of Men who have Sex with Men (MSM) Drawn from Seven Countries. *Substance Use & Misuse*, *53*(1), 42–50. doi:10.1080/10826084.2017.1322985

Chu, J. L., & Snider, C. E. (2013). Use of a Social Networking Web Site for Recruiting Canadian Youth for Medical Research. *Journal of Adolescent Health*, *52*(6), 792–794. doi:10.1016/j.jadohealth.2012.12.002

Chung, S. Y., Hacker, E. D., Rawl, S., Ellis, R., Bakas, T., Jones, J., & Welch, J. (2019). Using Facebook in Recruiting Kidney Transplant Recipients for a REDCap Study. *Western Journal of Nursing Research*, *41*(12), 1790–1812. doi:10.1177/0193945919832600

Côté-Léger, P., & Rowland, D. L. (2020). Estimations of Typical, Ideal, Premature Ejaculation, and Actual Latencies by Men and Female Sexual Partners of Men During Partnered Sex. *Journal of Sexual Medicine*, *17*(8), 1448–1456. doi:10.1016/j.jsxm.2020.04.317

Crosier, B. S., Brian, R. M., & Ben-Zeev, D. (2016). Using Facebook to Reach People Who Experience Auditory Hallucinations. *Journal of Medical Internet Research*, *18*(6). doi:10.2196/jmir.5420

Daniulaityte, R., Zatreh, M. Y., Lamy, F. R., Nahhas, R. W., Martins, S. S., Sheth, A., & Carlson, R. G. (2018). A Twitter-based survey on marijuana concentrate use. *Drug and Alcohol Dependence*, *187*, pp. 155–159. doi:10.1016/j.drugalcdep.2018.02.033

Dean, E., Cook, S., Murphy, J., & Keating, M. (2012). The Effectiveness of Survey Recruitment Methods in Second Life. *Social Science Computer Review*, *30*(3), 324–338. doi:10.1177/0894439311410024

Dijck, J. (2013). The Culture of Connectivity: A Critical History of Social Media. Oxford: Oxford University Press.

Ellis, R. J., Ganci, A., Head, K. J., & Ofner, S. (2018). Characteristics of Adults Managing Vitamins/Supplements and Prescribed Medications – Who Is Using, Not Using, and Abandoning Use of Pillboxes? A Descriptive Study. *Clinical Nurse Specialist*, *32*(5), 231–239. doi:10.1097/NUR.0000000000000395

Erhardt, J., & Freitag, M. (2021). The Janus-Face of Digitalization: The Relation Between Internet Use and Civic Engagement Reconsidered. *Social Science Computer Review*, *39*(3), 315–334. doi:10.1177/0894439319861966

Ersanilli, E. & van der Gaag, M. (2022). Data report: online surveys, Wave 2 & Polish 2021 online survey. *MOBILISE working papers*. doi: 10.31235/osf.io/twrzq

Facebook (October 25, 2021). Facebook Q3 2021 Earnings. Facebook Reports Third Quarter 2021 Results. Retrieved February 18, 2022, from: www.s21.q4cdn.com/399680738/files/doc_financials/2021/q3/FB-09.30.2021-Exhibit-99.1.pdf

Fenner, Y., Garland, S. M., Moore, E. E., Jayasinghe, Y., Fletcher, A., Tabrizi, S. N., Gunasekaran, B., & Wark, J. D. (2012). Web-Based Recruiting for Health Research Using a Social Networking Site: An Exploratory Study. *Journal of Medical Internet Research*, *14*(1). doi:10.2196/jmir.1978

Ferg, R., Conrad, F., & Gagnon-Bartsch, J. (2021). A Critical Evaluation of Tracking Public Opinion with Social Media: A Case Study in Presidential Approval. *methods, data, analyses, 15(2)*, 26. doi:10.12758/mda.2021.04

Folk, J. B., Harrison, A., Rodriguez, C., Wallace, A., & Tolou-Shams, M. (2020). Feasibility of Social Media-Based Recruitment and Perceived Acceptability of Digital Health

Interventions for Caregivers of Justice-Involved Youth: Mixed Methods Study. *Journal of Medical Internet Research*, *22*(4). doi:10.2196/16370

Ford, K. L., Albritton, T., Dunn, T. A., Crawford, K., Neuwirth, J., & Bull, S. (2019). Youth Study Recruitment Using Paid Advertising on Instagram, Snapchat, and Facebook: Cross-Sectional Survey Study. *JMIR Public Health and Surveillance*, *5*(4), 111–119. doi:10.2196/14080

Garey, L., Japuntich, S. J., Nelson, K. M., & Scott-Sheldon, L. A. (2020). Using Social Media to Recruit Youth Who Use Electronic Cigarettes. *American Journal of Health Behavior*, *44*(4), 488–498. doi:10.5993/AJHB.44.4.10

Groves, R. M. (2004). *Survey errors and survey costs* (Repr). *Wiley-Interscience paperback series*. Hoboken, New Jersey: Wiley-Interscience.

Groves, R. M., & Lyberg, L. (2010). Total Survey Error: Past, Present, and Future. *Public Opinion Quarterly*, *74*(5), 849–879. doi:10.1093/poq/nfq065

Grow, A., Perrotta, D., Del Fava, E., Cimentada, J., Rampazzo, F., Gil-Clavel, S., & Zagheni, E. (2020). Addressing Public Health Emergencies via Facebook Surveys: Advantages, Challenges, and Practical Considerations. *Journal of Medical Internet Research*, *22*(12), e20653. doi:10.2196/20653

Guillory, J., Kim, A., Murphy, J., Bradfield, B., Nonnemaker, J., & Hsieh, Y. (2016). Comparing Twitter and Online Panels for Survey Recruitment of E-Cigarette Users and Smokers. *Journal of Medical Internet Research*, *18*(11). doi:10.2196/jmir.6326

Guillory, J., Wiant, K. F., Farrelly, M., Fiacco, L., Alam, I., Hoffman, L., Crankshaw, E., Delahanty, J., & Alexander, T. N. (2018). Recruiting Hard-to-Reach Populations for Survey Research: Using Facebook and Instagram Advertisements and In-Person Intercept in LGBT Bars and Nightclubs to Recruit LGBT Young Adults. *Journal of Medical Internet Research*, *20*(6). doi:10.2196/jmir.9461

Harfield, S., Elliott, S., Ramsey, L., Housen, T., & Ward, J. (2021). Using social networking sites to recruit participants: methods of an online survey of sexual health, knowledge and behaviour of young South Australians. *Australian and New Zealand Journal of Public Health*, *45*(4), 348–354. doi:10.1111/1753-6405.13117

Hargittai, E. (2020). Potential Biases in Big Data: Omitted Voices on Social Media. *Social Science Computer Review*, *38*(1), 10-24. doi:10.1177/0894439318788322

Hellemans, J., Willems, K., & Brengman, M. (2020). Daily Active Users of Social Network Sites: Facebook, Twitter, and Instagram-Use Compared to General Social Network Site Use. In F. J. Martínez-López & S. D'Alessandro (Eds.), *Springer Proceedings in Business and Economics. Advances in Digital Marketing and eCommerce* (pp. 194–202). Springer International Publishing. doi: 10.1007/978-3-030-47595-6_24

Kepios Pte. Ltd., We Are Social Ltd., Hootsuite Inc. (October 21, 2021). Digital 2021. October Global Statshot Report. Retrieved February 18, 2022, from: www.datareportal.com/reports/digital-2021-october-global-statshot

Khumsaen, N., & Stephenson, R. (2017). Beliefs and Perception about HIV/AIDS, Self-Efficacy, and HIV Sexual Risk Behaviors Among Youth Thai Men Who Have Sex With Men. *AIDS Education and Prevention*, *29*(2), 175–190. doi:10.1521/aeap.2017.29.2.175

Knapp, A. A., Lee, D. C., Borodovsky, J. T., Auty, S. G., Gabrielli, J., & Budney, A. J. (2019). Emerging Trends in Cannabis Administration Among Adolescent Cannabis Users. *Journal of Adolescent Health*, *64*(4), 487–493. doi:10.1016/j.jadohealth.2018.07.012

Kühne, S., & Zindel, Z. (2020). *Using Facebook and Instagram to Recruit Web Survey Participants: A Step-by-Step Guide and Application*. doi:10.13094/SMIF-2020-00017

Leach, L. S., Bennetts, S. K., Giallo, R., & Cooklin, A. R. (2019). Recruiting fathers for parenting research using online advertising campaigns: Evidence from an Australian study. *Child Care Health and Development*, *45*(6), 871–876. doi:10.1111/cch.12698

Lee, S., Torok, M., Shand, F., Chen, N., McGillivray, L., Burnett, A., Larsen, M. E., & Mok, K. (2020). Performance, Cost-Effectiveness, and Representativeness of Facebook Recruitment to Suicide Prevention Research: Online Survey Study. *JMIR Mental Health*, *7*(10). doi:10.2196/18762

Lehdonvirta, V., Oksanen, A., Räsänen, P., & Blank, G. (2021). Social Media, Web, and Panel Surveys: Using Non-Probability Samples in Social and Policy Research. *Policy & Internet*, *13*(1), 134–155. doi:10.1002/poi3.238

Lu, J., & Yu, X. (2019). The Internet as a Context: Exploring Its Impacts on National Identity in 36 Countries. *Social Science Computer Review*, *37*(6), 705–722. doi:10.1177/0894439318797058

Manski, R., & Kottke, M. (2015). A Survey of Teenagers' Attitudes Toward Moving Oral Contraceptives Over the Counter. *Perspectives on Sexual and Reproductive Health*, *47*(3), 122–128. doi:10.1363/47e3215

McRobert, C. J., Hill, J. C., Smale, T., Hay, E. M., & van der Windt (2018). A multi-modal recruitment strategy using social media and internet-mediated methods to recruit a multidisciplinary, international sample of clinicians to an online research study. *PLOS ONE*, *13*(7). doi:10.1371/journal.pone.0200184

Meta Inc. *Reach everyone, or just a few.* Facebook. Retrieved February 18, 2022, from: www.facebook.com/business/ads/ad-targeting

Mitchell, J. W., & Petroll, A. E. (2012). Patterns of HIV and Sexually Transmitted Infection Testing Among Men Who Have Sex With Men Couples in the United States. *Sexually Transmitted Diseases*, *39*(11), 871–876. doi:10.1097/OLQ.0b013e3182649135

Nelson, E. J., Hughes, J., Oakes, J. M., Pankow, J. S., & Kulasingam, S. L. (2014). Estimation of Geographic Variation in Human Papillomavirus Vaccine Uptake in Men and Women: An Online Survey Using Facebook Recruitment. *Journal of Medical Internet Research*, *16*(9). doi:10.2196/jmir.3506

Obamiro, K., West, S., & Lee, S. (2020). Like, comment, tag, share: Facebook interactions in health research. *International Journal of Medical Informatics*, *137.* doi:10.1016/j.ijmedinf.2020.104097

Orehek, E., & Human, L. J. (2017). Self-Expression on Social Media: Do Tweets Present Accurate and Positive Portraits of Impulsivity, Self-Esteem, and Attachment Style? *Personality & Social Psychology Bulletin*, *43*(1), 60–70. doi:10.1177/0146167216675332

Pagoto, S. L., Schneider, K. L., Oleski, J., Smith, B., & Bauman, M. (2014). The Adoption and Spread of a Core-Strengthening Exercise Through an Online Social Network. *Journal of Physical Activity & Health 11*(3), 648–653. doi:10.1123/jpah.2012-0040

Pepper, J. K., Coats, E. M., Nonnemaker, J. M., & Loomis, B. R. (2019). How Do Adolescents Get Their E-Cigarettes and Other Electronic Vaping Devices? *American Journal of Health Promotion*, *33*(3), 420–429. doi:10.1177/0890117118790366

Perrotta, D., Grow, A., Rampazzo, F., Cimentada, J., Del Fava, E., Gil-Clavel, S., & Zagheni, E. (2021). Behaviours and attitudes in response to the COVID-19 pandemic: insights from a cross-national Facebook survey. *EPJ Data Science*, *10*(1). doi:10.1140/epjds/s13688-021-00270-1

Pew Research Center. (2015, August 28). Men catch up with women on overall social media use. *Pew Research Center*. Retrieved September 25, 2022, from https://www.pewresearch.org/fact-tank/2015/08/28/men-catch-up-with-women-on-overall-social-media-use/

Pötzschke, S., & Braun, M. (2017). Migrant Sampling Using Facebook Advertisements: A Case Study of Polish Migrants in Four European Countries. *Social Science Computer Review*, *35*(5), 633–653. doi:10.1177/0894439316666262

Pozzar, R., Hammer, M. J., Underhill-Blazey, M., Wright, A. A., Tulsky, J. A., Hong, F., Gundersen, D. A., & Berry, D. L. (2020). Threats of Bots and Other Bad Actors to Data Quality Following Research Participant Recruitment Through Social Media: Cross-Sectional Questionnaire. *Journal of Medical Internet Research*, *22*(10), e23021. doi:10.2196/23021

Quach, S., Pereira, J. A., Russell, M. L., Wormsbecker, A. E., Ramsay, H., Crowe, L., Quan, S. D., Kwong, J. (2013). The Good, Bad, and Ugly of Online Recruitment of Parents for Health-Related Focus Groups: Lessons Learned. *Journal of Medical Internet Research, 15*(11), e250. doi:10.2196/jmir.2829

Ramo, D. E., & Prochaska, J. J. (2012). Broad Reach and Targeted Recruitment Using Facebook for an Online Survey of Young Adult Substance Use. *Journal of Medical Internet Research*, *14*(1). doi:10.2196/jmir.1878

Reuter, K., Zhu, Y. F., Angyan, P., Le, N., Merchant, A. A., & Zimmer, M. (2019). Public Concern About Monitoring Twitter Users and Their Conversations to Recruit for Clinical Trials: Survey Study. *Journal of Medical Internet Research*, *21*(10). doi:10.2196/15455

Robstad, N., Westergren, T., Siebler, F., Soderhamn, U., & Fegran, L. (2019). Intensive care nurses' implicit and explicit attitudes and their behavioural intentions towards obese intensive care patients. *Journal of Advanced Nursing*, *75*(12), 3631–3642. doi:10.1111/jan.14205

Rosenzweig, L. R., & Zhou, Y.-Y. (2021). Team and Nation: Sports, Nationalism, and Attitudes Toward Refugees. *Comparative Political Studies*, 001041402199749. doi:10.1177/0010414021997498

Rosso, M. T., & Sharma, A. (2020). Willingness of Adults in the United States to Receive HIV Testing in Dental Care Settings: Cross-Sectional Web-Based Study. *JMIR Public Health and Surveillance*, *6*(3), 99–110. doi:10.2196/17677

Russomanno, J., & Tree, J. M. (2020). Food insecurity and food pantry use among transgender and gender non-conforming people in the Southeast United States. *BMC Public Health*, *20*(1). doi:10.1186/s12889-020-08684-8

Salk, R. H., Thoma, B. C., & Choukas-Bradley, S. (2020). The Gender Minority Youth Study: Overview of Methods and Social Media Recruitment of a Nationwide Sample of US Cisgender and Transgender Adolescents. *Archives of Sexual Behavior*, *49*(7), 2601–2610. doi:10.1007/s10508-020-01695-x

Samuels, D., & Zucco, C. (2014). The Power of Partisanship in Brazil: Evidence from Survey Experiments. *American Journal of Political Science*, *58*(1), 212–225. doi:10.1111/ajps.12050

Samuels, D. J., & Zucco, C. (2013). Using Facebook as a Subject Recruitment Tool for Survey-Experimental Research. *SSRN Electronic Journal*. Advance online publication. doi:10.2139/ssrn.2101458

Seidler, Z. E., Wilson, M. J., Walton, C. C., Fisher, K., Oliffe, J. L., Kealy, D., Ogrodniczuk, J. S., & Rice, S. M. (2021). Australian men's initial pathways into mental health services. *Health Promotion Journal of Australia: Official Journal of Australian Association of Health Promotion Professionals.* Advance online publication. doi:10.1002/hpja.524

Shakir, S. M. M., Wong, L. P., Abdullah, K. L., & Adam, P. (2019). Factors associated with online sexually transmissible infection information seeking among young people in Malaysia: an observational study. *Sexual Health*, *16*(2), 158–171. doi:10.1071/SH17198

Sharma, A., Kahle, E. M., Sullivan, S. P., & Stephenson, R. (2018). Birth Cohort Variations Across Functional Knowledge of HIV Prevention Strategies, Perceived Risk, and HIV-Associated Behaviors Among Gay, Bisexual, and Other Men Who Have Sex With Men in the United States. *American Journal of Mens Health*, *12*(6), 1824–1834. doi:10.1177/1557988318790875

Shaver, L. G., Khawer, A., Yi, Y. Q., Aubrey-Bassler, K., Etchegary, H., Roebothan, B., Asghari, S., & Wang, P. P. (2019). Using Facebook Advertising to Recruit Representative Samples: Feasibility Assessment of a Cross-Sectional Survey. *Journal of Medical Internet Research*, *21*(8). doi:10.2196/14021

Stier, S., Breuer, J., Siegers, P., & Thorson, K. (2020). Integrating Survey Data and Digital Trace Data: Key Issues in Developing an Emerging Field. *Social Science Computer Review*, *38*(5), 503–516. doi:10.1177/0894439319843669

Suliman, M., Al Qadire, M., Alazzam, M., Aloush, S., Alsaraireh, A., & Alsaraireh, F. A. (2018). Students nurses' knowledge and prevalence of Needle Stick Injury in Jordan. *Nurse Education Today*, *60*, 23–27. doi:10.1016/j.nedt.2017.09.015

Sullivan, P. S., Khosropour, C. M., Luisi, N., Amsden, M., Coggia, T., Wingood, G. M., & DiClemente, R. J. (2011). Bias in Online Recruitment and Retention of Racial and Ethnic Minority Men Who Have Sex With Men. *Journal of Medical Internet Research*, *13*(2). doi:10.2196/jmir.1797

Sunderland, M., Batterham, P. J., Calear, A. L., & Carragher, N. (2017). The development and validation of static and adaptive screeners to measure the severity of panic disorder, social anxiety disorder, and obsessive compulsive disorder. *International Journal of Methods in Psychiatric Research*, *26*(4). doi:10.1002/mpr.1561

Telegram (November 8th, 2021). *Telegram Info.* Retrieved February 18, 2022, from: www.t.me/tginfo/3147

Thornton, L. K., Harris, K., Baker, A. L., Johnson, M., & Kay-Lambkin, F. J. (2016). Recruiting for addiction research via Facebook. *Drug and Alcohol Review*, *35*(4), 494–502. doi:10.1111/dar.12305

Twitter Inc. (n.d.) *Reach the right people at the right time with Twitter's targeting tools.* Twitter. Retrieved February 18, 2022, from: www.business.twitter.com/en/advertising/targeting.html

Wagenaar, B. H., Christiansen-Lindquist, L., Khosropour, C., Salazar, L. F., Benbow, N., Prachand, N., Sineath, R. C., Stephenson, R., & Sullivan, P. S. (2012a). Willingness of US Men Who Have Sex with Men (MSM) to Participate in Couples HIV Voluntary Counseling and Testing (CVCT). *PLOS ONE*, *7*(8). doi:10.1371/journal.pone.0042953

Wagenaar, B. H., Sullivan, P. S., & Stephenson, R. (2012b). HIV Knowledge and Associated Factors among Internet-Using Men Who Have Sex with Men (MSM) in South Africa and the United States. *PLOS ONE*, *7*(3). doi:10.1371/journal.pone.0032915

Webler, T., Holewinski, M., Orrick, B., & Kaur, R. (2020). Toward a method for the rapid collection of public concerns and benefits of emerging energy technologies. *Journal of Risk Research*, *23*(1), 35–46. doi:10.1080/13669877.2018.1485174

Welton, J. M., Walker, C., Riney, K., Ng, A., Todd, L., & D'Souza, W. J. (2020). Quality of life and its association with comorbidities and adverse events from antiepileptic medications: Online survey of patients with epilepsy in Australia. *Epilepsy & Behavior*, *104*. doi:10.1016/j.yebeh.2019.106856

Williamson, S., & Malik, M. (2021). Contesting narratives of repression: Experimental evidence from Sisi's Egypt. *Journal of Peace Research*, *58*(5), 1018–1033. doi:10.1177/0022343320961835

Wilson, K. M., Beggs, S. A., Zosky, G. R., Bereznicki, L. R., & Bereznicki, B. J. (2019). Parental knowledge, beliefs and management of childhood fever in Australia: A nationwide survey. *Journal of Clinical Pharmacy and Therapeutics*, *44*(5), 768–774. doi:10.1111/jcpt.13000

Wolak, J. J. D., Finkelhor, D., Walsh, W., & Treitman, L. (2018). Sextortion of Minors: Characteristics and Dynamics. *Journal of Adolescent Health*, *62*(1), 72–79. doi:10.1016/j.jadohealth.2017.08.014

Woodward, S. C., Bereznicki, B. J., Westbury, J. L., & Bereznicki, L. R. (2016). The effect of knowledge and expectations on adherence to and persistence with antidepressants. *Patient Preference and Adherence*, *10*. doi:10.2147/PPA.S99803

Yuan, P., Bare, M. G., Johnson, M. O., & Saberi, P. (2014). Using Online Social Media for Recruitment of Human Immunodeficiency Virus-Positive Participants: A Cross-Sectional Survey. *Journal of Medical Internet Research*, *16*(5), 101–109. doi:10.2196/jmir.3229

Zhang, B. B., Mildenberger, M., Howe, P. D., Marlon, J., Rosenthal, S. A., & Leiserowitz, A. (2020). Quota sampling using Facebook advertisements. *Political Science Research and Methods*, *8*(3), 558–564. doi:10.1017/psrm.2018.49

The Role of Public Opinion Research in the Democratic Process: Insights from Politicians, Journalists, and the General Public

Henning Silber[1], *Allyson L. Holbrook*[2] & *Timothy P. Johnson*[2]

[1] *GESIS – Leibniz Institute for the Social Sciences*

[2] *University of Illinois at Chicago*

Abstract

This study reveals the existence of a paradox in how the public views polling within the democratic process. Specifically, even though the public believes that it can influence policymaking, it considers public opinion polls not as useful as other, less representative forms of public input, such as comments at town hall meetings. Analyzing data from multiple surveys conducted in the United States of America, we find no evidence for the democratic representation hypothesis with respect to polling. Comparisons across stakeholders (public, journalists, and politicians) demonstrate that general perceptions of inputs into the democratic process are similar, which confirms the citizen-elite congruence hypothesis. However, unlike members of the public, experts are more likely to believe that public opinion polls are the optimal method by which the public can successfully inform policymaking, a finding consistent with the legitimization hypothesis. With respect to perceptions of politicians, we found substantial differences regarding party registration with Democrats and Independents favoring public opinion polling and Republicans preferring alternative methods (e.g., town hall meetings) of informing policymakers.

Keywords: public opinion research, public policy, democratic representation, policymaking, survey value, preferences, media, politicians

There has long been a connection between public opinion polling and policymaking (Burstein, 2003, 2010; Page & Shapiro, 1983, 2010; Sobel, 2001; Wlezien & Soroka, 2012). With respect to democratic representation, polling has an important democratic function by informing politicians about beliefs of the electorate, which may guide their policy decisions (Bowler, Donovan, & Karp, 2007). Compared with other public policy input sources such as town hall meetings, campaign events, demonstrations, phone calls, letters, or emails from members of the public to a politician or policymaker, public opinion surveys remain the most systematic and representative aggregations of public opinion (Verba, 1996).

Policy leaders have for many years used public opinion polls both to understand what the public thinks and to actively shape public opinion (Jacobs & Shapiro, 1995, 2000). Public opinion is also used by the public as a source of information regarding what other people think (Moy & Rinke, 2012). While political elites regularly conduct their own public opinion polls, the public relies on others to sponsor them. This role is often taken by the media, which at the same time summarizes public opinion data derived from other sources such as think tanks, town hall meetings, attendance at political events, and person-on-the-street interviews (Herbst, 1993; Jacobs & Shapiro, 2005; Rosenstiel, 2005; Strömbäck, 2012).

This research investigates the perceived role of public opinion research in the democratic process by contrasting perceptions of members of the public with elite perceptions of journalists and politicians. Public opinion research is compared with other policy input sources (e.g., interest groups) and other means by which the pub-

Acknowledgements

We thank Liz Hamel from the Kaiser Family Foundation for providing additional information about the data collection. We also thank the Survey Research Group of the Department of Public Administration at the University of Illinois at Chicago and the participants of the Research Colloquium of the Survey Operations Team of the Department of Survey Design and Methodology at GESIS – Leibniz Institute for the Social Sciences for helpful comments on a previous version of the manuscript.

The research for this manuscript was conducted while the first author was a Fulbright Research Fellow at the Department of Public Administration at The University of Illinois at Chicago and a Visiting Researcher at the Center of Excellence in Survey Researcher at NORC at The University of Chicago.

This research uses data from the Roper Archive (USAIPOGNS1999-9904022; USPSRA2001-POL015A; USPSRA2001-POL015B). The data sets were collected by the Kaiser Family Foundation and by the Gallup organization.

We refer to data reports by Kantar (https://us.kantar.com/public-affairs/politics/2013/kantars-path-to-public-opinion/) and McClatchy-Marist (http://maristpoll.marist.edu/wp-content/misc/usapolls/us170323_MCC_HBO/Complete%20Survey%20Findings_McClatchy-Marist%20Poll_March%202017.pdf).

Direct correspondence to

Henning Silber, GESIS – Leibniz Institute for the Social Sciences, B6 4-5,
68072 Mannheim, Germany
E-mail: henning.silber@gesis.org

lic can interact with politicians (e.g., town hall meetings) and influence their decision making. We also investigate which political party is less or more supportive of public opinion research (e.g., Democrats or Republicans). While the study is mainly exploratory, three specific research hypotheses are examined. First, the *democratic representation hypothesis*, which assumes that in a democracy, citizens prefer that politicians base their decisions on the views of the public. Second, the *elite-citizen congruence hypothesis* that postulates similarity between the perceptions of both groups. Third, the *legitimization hypothesis*, which suggests that elites perceive polling as more influential than do members of the public because they use polling professionally. For the empirical analyses, we use data of four studies from the United States of America, which measured perceptions regarding the societal role of public opinion research across members and the public, journalists, and politicians.

The paper continues with an overview of previous research on democratic representation through polling from the perspective of citizens and elites. Afterward, we describe our data and methods, present the empirical results, and discuss our findings.

Democratic Representation and Polling

While certainly not without problems such as nonattitudes, information levels, and multiple conflicting preferences (Burstein, 2010; Converse, 1964; Zaller, 1992), findings from public opinion polls have many important functions within a democracy (Delli Carpini & Keeter, 1996; Page, 1994; Shapiro, 1998; 2011). No function is more important, though, than its role informing and providing policy decision-makers and the public with reliable information regarding general public sentiment and preferences regarding contentious policy issues. Of course, policy decisions are also based on input from sources other than public opinion (Burstein, 2003; Gray, 2004; Verba, 1996). MacInnis, Anderson, and Krosnick (2018, 9) identify six different information sources which policymakers in Congress often consider: the general public, the issue public, economic elites, donors and sponsors, political parties, and the president. Two of these sources relate to the public (general and issue), three to special interest groups (economic elites, donors, and sponsors), and two to political elites (political parties and the president). Other information sources that may also receive attention include the media and both policy and political experts (Jacobs & Shapiro, 2000).

Our data allows us to examine the *democratic representation hypothesis* with respect to public opinion polls, which suggests that in a democracy public opinion influences governmental decisions (Newport et al., 2013). Following this argument, polling - as the most representative aggregations of public opinion (Verba, 1996) – should be perceived as the ideal way of affecting political decision-making.

There are only a few studies that have compared the various information sources and interest groups that may supply inputs to public officials for policymaking purposes and those comparisons have mostly focused on the opinions of the general public (Doherty, 2013; Doherty et al., 2019; MacInnis, Anderson, & Krosnick, 2018; Soroka, 2002). MacInnis, Anderson, and Krosnick (2018) found that the public believes members of Congress should pay the most attention to the general public and to people who feel strongly about an issue, while believing that their representatives actually pay more attention to the preferences of their supporters, campaign donors, and economic elites. Doherty et al. (2019) examined the relative differences between three different groups of representations in the policy formation process (campaign promises, voters, the general public) and found that the public believes all three should be considered equally. Soroka (2002) used data collected from several different sources, including the public, the media, and elected officials, concluding that both the public and the media play important roles in policymaking and agenda-setting. In summary, existing studies have not compared public opinion polling to other policy inputs in the eyes of the public, nor have the views of the public with respect to the role of polling been compared to the beliefs of elites.

Politicians and Polling

Politicians receive policy input from many different sources including the public, interest groups, lobbyists, the media, experts, their party and other politicians, from which they have to select and prioritize, especially when considering important political questions (Walgrave et al., 2018). Starting at least as far back as Kennedy, U.S. presidents have used public opinion polls to understand what the public thinks about various issues (Beal & Hinckley, 1984; Heith, 1998). Thus, Presidents and political candidates are believed to consider and consult polls for elections and when making important political decisions. Public opinion research has likewise informed policymakers on the state level almost since its inception (e.g., Erikson, 1976; Percival, Johnson, & Neiman, 2009).

When comparing the decision-making process of politicians with that of the electorate, Sheffer et al. (2018) showed that the reasoning characteristics of the two groups are quite comparable. This may also be applicable to polling so that attitudes and values toward polling possibly will have the same effect for the public and for politicians on their perception of polling within the democratic process (*citizen-elite congruence hypothesis*; Hibbing & Theiss-Morse, 2001; André & Depauw, 2017). In addition, studies by Cayton (2017) and Joly, Hofmans, and Loewen (2018) showed partisanship-based differences within political elites. Joly, Hofmans, and Loewen (2018), for example, reported higher levels of openness to experience among progressive political parties. Thus, we may observe likewise party differences regarding beliefs about public opinion polling.

Media and Polling

Public opinion polling is also inextricably linked to mass media coverage (Jacobs & Shapiro, 2005; Rosenstiel, 2005; Strömbäck, 2012). Many media organizations conduct their own public opinion polls and also gather and summarize polling data from multiple sources. Journalists use the results of public opinion surveys to inform the public about political issues such as opinion trends and politician ratings (Rosenstiel, 2005). Through that active and highly visible role, the news media may be considered the "leading actor" in the public opinion polling business (Gollin, 1987, 87). Frequent polling updates became possible through the introduction of telephone interviewing in the late 1970s (Curtin, Presser, & Singer 2005), which allowed for the fast and inexpensive gathering of nation-wide public opinion data. The introduction of web surveys over the past two decades (Couper & Miller, 2008) served to accelerate this process. While media reports of public opinion data regarding countless societal questions have become an integral part of everyday life, those reports receive even greater attention during election periods, when findings are reported daily (Hillygus, 2011; Patterson, 2005). While public opinion polls clearly receive considerable attention from the news media, there is little existing evidence as to the relative value that media actors place on public opinion as a public policy input source. In this context, the *legitimization hypothesis* suggests that media actors are likely to assume a relatively high impact of polling within the democratic process, since this would legitimize their professional efforts in this area.

Public Perceptions of Polling

On a societal level, a variety of factors, including misuse and misinterpretation of polling data, over-surveying, and both marketing and fundraising under the guise of public opinion research, have converged to undermine the legitimacy of public opinion polling (Johnson, 2018). In response, researchers have begun to investigate the public's perceptions of public opinion research by studying the "survey climate" in various nations (e.g., de Leeuw et al., 2019; Gengler et al., 2019; Looseveldt & Storms, 2008; Lyberg & Lyberg, 1991; Stocké & Langfeldt, 2004). Measures of survey climate capture societal factors such as trust in public institutions, civic and social engagement, and satisfaction with democracy, as well as individual factors such as knowledge of, trust in, and beliefs regarding the value and reliability of surveys, and the degree to which citizens pay attention to and discuss them (de Leeuw et al., 2019; Looseveldt & Joye, 2016). While intuition suggests that positive beliefs regarding the efficacy of public opinion surveys should be associated with support for their use in policy decision-making, there is currently no evidence that addresses this question.

We turn now to original analyses of multiple data sets that provide the opportunity to investigate these questions regarding public perceptions of the value of opinion polling as a policy input and how they compare with the beliefs of other actors in the policy process: politicians and journalists.

Methods

Data

This article examines survey data from four data sets in which the general public (Studies 1 and 2), journalists (Study 3), and politicians (Study 4) were each interviewed (see Table 1). All data sets relied on telephone survey methodology and were collected between 1999 and 2001 by Gallup Research and the Kaiser Family Foundation in the United States of America. All data were obtained from the Roper Center Public Opinion Data Archive.[1]

The two surveys of the public (Studies 1 and 2) both used probability sampling approaches and included more than 1,000 respondents (see Table 1). A response rate was not available for Study 1. For Study 2, the response rate was 62.3% (Princeton Survey Research Associates 2001a). In addition, two expert surveys were available, one with journalists and one with politicians (Brodie et al. 2001). The survey of journalists (Study 3) included professionals from top newspapers (180), TV and radio networks (70), and news services and magazines (51). The politician survey (Study 4) included 96 senior executive branch officials, 2 members of Congress, 40 senior Congressional staff, 70 think tank scholars, 54 lobbyists, and 38 trade association executives. The response rate of Study 3 was 44.9%, and the response

[1] **Study 1**: Gallup Organization. Gallup Poll: Baseline Study on Polls and Polling Organization Awareness, 1999 [Dataset]. Roper #31088772, Version 2. Gallup Organization [producer]. Cornell University, Ithaca, NY: Roper Center for Public Opinion Research [distributor]. doi:10.25940/ROPER-31088772.
Study 2: Henry J. Kaiser Family Foundation in collaboration with Public Perspective magazine. Kaiser Family Foundation/Public Perspective Magazine Poll: Polling & Democracy, 2001 [Dataset]. Roper #31096753, Version 2. Princeton Survey Research Associates [producer]. Cornell University, Ithaca, NY: Roper Center for Public Opinion Research [distributor]. doi:10.25940/ROPER-31096753.
Studies 3 and 4: Henry J. Kaiser Family Foundation in collaboration with Public Perspective magazine. Kaiser Family Foundation/Public Perspective Magazine Poll: Polling & Democracy: Policy Makers and Media, 2000 [Dataset]. Roper #31096754, Version 1. Princeton Survey Research Associates [producer]. Cornell University, Ithaca, NY: Roper Center for Public Opinion Research [distributor]. doi:10.25940/ROPER-31096754.

rate of Study 4 was 27.9% (Princeton Survey Research Associates, 2001b). Detailed information regarding each study is available from the reports cited here and from the Roper Center Archive. Table 2 summarizes the demographics and political orientations of respondents from all four surveys. Considering sample composition, the sample of journalists was especially unbalanced regarding party registration and political ideology since it was composed of only 4.7% Republicans and 6.3% Conservatives. Therefore, the results regarding political orientations should be treated with caution for this group, and the overall results reflect primarily the

Table 1 Overview of the Survey Data Sources (all USA)

Study	Organization	Year	Population	Sample Size	Survey Mode	Sampling Method	Response Rate
1	Gallup	1999	Public	1,011	Telephone	Probability	NA[a]
2	Kaiser	2001	Public	1,206	Telephone	Probability	62.3%
3	Kaiser	2001	Journalists	301	Telephone	Nonprobability	44.9%
4	Kaiser	2001	Politicians	300	Telephone	Nonprobability	27.9%

[a] A response rate was requested, but not available (NA) for Study 1.

Table 2 Description of the Survey Data Sources

	Study 1	Study 2	Study 3	Study 4
	Public	Public	Journalists	Politicians
Education[a]				
Low	40.3%	43.3%	NA	NA
Medium	28.3%	21.7%	NA	NA
High	31.4%	34.9%	NA	NA
Age (mean)	44.7 years	44.6 years	45.5 years	49.3 years
Female	52.2%	53.2%	36.2%	25.3%
Party Registration				
Republican	40.5%	35.0%	4.7%	24.3%
Independent	15.3%	25.7%	31.1%	27.6%
Democrat	44.2%	39.4%	64.2%	46.2%
Political Ideology				
Conservative	38.7%	36.5%	6.3%	19.2%
Moderate	41.2%	41.2%	66.0%	54.9%
Liberal	20.1%	22.3%	27.6%	25.9%
(n)	1206	1000	301	300

[a] Education was not available (NA) for Study 3 and 4.

perspectives of journalists with other political orientations. The sample size of the journalist survey (Study 3) was 301, and the politician survey (Study 4) included 300 respondents.

Measures

The precise question wordings of all questions examined in this paper are provided in the Online Appendix.

Attention to polling. All four data sets included a measure on how much attention respondents felt should be paid to polling when policy decisions are being made. Specifically, the Gallup data set included three questions as to whether or not policymakers, the U.S. president, or the public as a whole would be better off if more or less attention would be given to polling. The three Kaiser data sets included a rating question in which respondents were asked how much attention governmental officials are currently paying to several policy input sources (their own knowledge, their conscience, lobbyists, campaign contributors, journalists, policy experts, members of the public, and public opinion polls). The public dataset included an experimental design in which a random half of respondents were assigned to a different version of the before-mentioned question, which used the exact same wording but asked respondents how much attention “should be” (rather than “is”) paid to each of the several policy-input sources, allowing a comparison of those responses.

Preferences of the public. Studies 2-4 included comparative measures of respondents’ opinions regarding different ways that public preferences could influence political decision-making. Specially, those studies included questions that asked how important respondents felt that (1) town hall meetings, (2) conducting public opinion polls, (3) talking to people at shopping malls and on the street, and (4) talking to people who call, write or e-mail the official’s office, were as ways of learning what the majority of the public believes. The question was asked in two ways, first, all respondents received the question in a rating format in which they were asked to rate each of the response alternatives as a very good, somewhat good, not too good, or not at all good way to learn what the majority believes (rating). Afterward, all respondents were asked to identify which of these methods of obtaining policy input they rated as the most valuable for political officials (ranking).

Survey value. All four surveys included measures of the perceived value of surveys. In Study 1 and 2, survey value was measured using 2 (alpha = .570) or 3 items (alpha = .671), respectively. These measures included questions such as “polls on social and political issues serve a useful purpose” and “do you feel polls give you a better understanding of the news of the day, or not.” In Study 3 and 4, one item was available to assess survey value: “Public opinion polling is far from perfect, but it is one of the best means we have for communicating what the public is thinking."

Background variables. The four studies included several background questions. Those questions asked respondents to report their gender, age, and education. Respondents were also asked to report their political affiliation (Republican, Democrat, or Independent) and political ideology (conservative, moderate, or liberal). For Study 3 and 4, education was not available.

Analyses

Three sets of analyses are presented. First, means and mean differences between the experimental groups (Study 2) and between the different samples are tested using t-tests for single comparisons and One-way ANOVAs with Bonferroni adhoc tests of multiple means. These assess the extent to which the public thinks that policymakers should pay more or less attention to public opinion polls and to other potential sources of information. They also enable us to compare public opinion about the use of public opinion polls in policymaking to the opinions of policymakers and journalists. Second, logistic and linear regression models are calculated to test for possible predictors of the measures of "attention to polling" and "preferences of the public." In those models, we use logistic regressions when the dependent variable was dichotomous and linear regressions when the dependent variable was measured with a rating scale. Independent variables include survey value, political orientations, and demographics. Third, correlation coefficients are calculated to test the congruency between the beliefs of the public and the two expert groups. For example, the means of the answers to the eight rating questions on the policy input sources of the public are correlated with the mean answers of politicians and journalists. All analyses are unweighted.

Results

RQ1: Public Perceptions of the Role of Public Opinion Research

Using Study 1, we first examined whether members of the public think that policymakers should use information from surveys more or less than they currently do (see Table 3). The three questions that directly ask respondents whether more or less attention should be paid to polling by policymakers, the president, or the general public each show that the public believes there should be less attention given to polling (differences ranged from -17.0% to -23.0%, $p < .001$). While 51% or more of the respondents thought that polling should receive less attention, not more than 37% believed that there should be more attention paid to polling.

Table 3 Attention that policymakers, the president, or the country as a whole *should* pay to polling

	Policymakers[a]	President[a]	General public[a]
Less/too much attention[b]	56.6	51.2	55.6
More/not enough attention[b]	36.6	34.2	32.2
Right amount[c]	1.8	4.7	5.6
Don't know/refusal[c]	5.0	9.7	6.5
(n)	1011	1011	1011

[a] in percent

[b] The policymaker and president questions asked respondents whether policymakers pay "too much" or "not enough" attention to polls, and the general public question asked respondents whether "less" or "more" attention than now should be paid to polls by policymakers (see Online Appendix for the question wording).

[c] Those response options were not stated in the question and only volunteered.

Data source: Study 1, Gallup 1999 (General Population)

Next, we examined predictors of support for the use of surveys by policymakers, by the president, and by the general public. In general, the logistic regression models presented in Table 4 indicate that respondents who perceived surveys as more valuable were more likely to believe that more attention should be given to polling in the policymaking process. Also, Democrats, Liberals, respondents of younger age, those with less education, and females had a higher probability of believing that surveys should receive more consideration regarding policy decisions. Notably, the results in Table 4 are very similar for each information recipient (policymakers, the president or the general public), suggesting that public beliefs regarding the importance of polling in the policymaking process are fairly stable.

Using Study 2, we also compared public beliefs about the extent to which surveys should be used more/less in policymaking with public beliefs about whether other sources of information should be (and are) used more or less. When comparing the various policy input sources, the public believes that policymakers pay the most attention to campaign contributors and lobbyists, while polls are rated as sixth out of the eight policy input sources examined (see Table 5). In contrast, the public believes that policymakers should pay much less attention to campaign contributors and lobbyists and more attention to the public, represented through members who contact them and via public opinion polls. The great discrepancy between beliefs of the public on how much attention politicians pay and should pay to various policy input sources can be illustrated by the correlation between the mean values for each, which was -.445. When considering how policymakers should be informed by the public, input from members of the public who directly contact them was preferred compared to mediated input through polling. Notably, when the question

Table 4 Predictors of whether more or less attention *should* be paid to polling by policymakers, the president, or the general public

	Policymakers	President	General Public
Education	.617***	.665***	.508***
Age	.986**	.988*	.991
Female	1.426*	1.545*	1.101
Party registration (ref. Republican)			
Independent	1.866*	1.367	.848
Democrat	1.967***	1.991***	1.779**
Political Ideology (ref. conservative)			
Moderate	1.019	1.206	1.403
Liberal	1.797*	1.760*	1.809*
Survey value	2.342***	2.095***	4.406***
R^2	.285	.250	.452
(n)	772	706	746

****p*<.001 ***p*<.01 **p*<.05

Note. Analyses are based on logistic regression models connected to Table 3 (1=more/not enough attention, 0=less/too much attention). Odds ratios and Nagelkerke's Pseudo-R^2 are displayed.

Data source*:* Study 1, Gallup 1999 (General Population)

of whether polling should receive more attention than it receives now was asked in two separate questions, and in context with multiple other policy input sources, the public believed that polling should receive more attention than it does now (difference = .265, $t(1172) = 5.543$, $p < .001$), which somewhat contradicts the results of Table 3 and illustrates that univariate distributions should always be interpreted with consideration of the question context.

The question of how much attention 'polls' and 'members of the public' should receive by policymakers compared to the other policy input sources (see Column 2 "Should pay attention" in Table 5) allowed us to examine the democratic representation hypothesis. While this hypothesis is supported for 'members of the public' (ranked first), 'polling' is only ranked fifth out of eight and even 'policy experts' (ranked fourth) are assessed as a preferable policy input source compared to polling (difference = .295, $t(570) = 6.771$, $p < .001$). Hence, the democratic representation hypothesis is not supported for public opinion polling.

In Study 2, several of the inputs that respondents were asked about focused on other sources of information about public opinion (e.g., town hall meetings, etc.), facilitating comparisons between beliefs about these sources of information with

Table 5 Comparison of the public, journalists, and politicians beliefs regarding sources public officials pay attention to (and should pay attention to)

	Public		Journalists	Politicians
	Pay attention	Should pay attention	Pay attention	Pay attention
Own knowledge	3.20^{b}	3.38^{a}	3.36^{a}	3.46^{a}
Their conscience	2.79^{bd}	3.35^{acd}	2.90^{bd}	3.15^{abc}
Lobbyists	3.31^{bc}	2.35^{acd}	3.66^{abc}	3.41^{bd}
Campaign contributors	3.52^{bcd}	2.33^{acd}	3.68^{abd}	3.32^{abc}
Journalists	2.77^{b}	2.30^{acd}	2.66^{b}	2.83^{b}
Policy experts	3.13^{b}	3.34^{acd}	3.01^{b}	3.08^{b}
Members of the public	2.62^{bd}	3.46^{acd}	2.73^{bd}	2.93^{abc}
Polls	2.78^{bcd}	3.05^{ac}	3.28^{abd}	3.10^{ac}
(n)	600	600	301	300

Note. The table displays means. Response categories: 1 “not at all” 2 “not too much” 3 “a fair amount” 4 “a great deal”

Differences between the means are tested with post-hoc-tests of multiple means using Bonferroni correction. “a” refers to significant differences ($p < .05$) between public “pay attention” and the three other measurements; “b” refers to significant differences ($p < .05$) between public should pay attention and the other three measurements; “c” refers to significant differences ($p < .05$) between journalists and the three other measurements; “d” refers to significant differences ($p < .05$) between politicians and the other three measurements.

Data sources: Studies 2 to 4, Kaiser 2001 (General Population, Expert Samples Journalists and Politicians)

beliefs about public opinion polls. The results show that holding a town hall meeting was the approach most favored by the public for influencing policy decisions (see Table 6). In comparison, conducting a public opinion poll was rated third when the question was asked in the rating format, and second when the question was asked in the ranking format.

Altogether, with respect to our first research question about the preferences of the public, the results suggest that the public does not prefer public opinion polling compared to other, more direct means of having policy input. In fact, when asked specifically, they believe that polling should have less impact on policymaking. In line with that, the public believes that direct ways of communicating with politicians, for instance, through town hall meetings, are a better way to influence policy decisions. Most supportive of public opinion polling as a policy input source were those members of the public who place greater value in them. Political orientation, in contrast, did not have an impact on perceptions about the policy relevance of opinion polling.

Table 6 Different ways policymakers can learn what the public wants

	Public		Journalists		Politicians	
	Rating	Ranking	Rating	Ranking	Rating	Ranking
Holding a town meeting	3.38	43.0	3.17	25.2	3.20	33.6
Conducting a public opinion poll	3.10	25.4	3.29	51.8	3.13	49.8
Talking to people at shopping malls or on the street	2.98	13.0	2.84	7.6	2.75	10.5
Talking to people who call, write, or e-mail	3.18	15.3	2.76	3.3	2.73	6.1
(n)	1187	1165	299	265	299	277

Note. Response categories: Rating: 1 "not at all good" 2 "not too good" 3 "somewhat good" 4 "very good"; Ranking: Best way in %

Data sources: Studies 2 to 4, Kaiser 2001 (General Population, Expert Samples Journalists and Politicians)

RQ2: Perceptions of Politicians and Journalists of the Role of Public Opinion Research

We next turned our attention to comparing public beliefs about how much surveys are used in policymaking with the beliefs of two important groups of experts – politicians and journalists. Confirming the citizen-elite congruence hypothesis, compared to the public and to each other, journalists and politicians have a very similar view of how much attention policymakers pay to the various policy input sources available (see Table 5, Study 2 to 4). The correlation between the beliefs of the public and journalists was .851, the correlation between the beliefs of the public and politicians was .779, and the correlation between the beliefs of journalists and politicians was .870. Supplementary analyses of politicians with respect to party registration show that the congruence between beliefs of the public and politicians is driven by Democrats (r = .774) and Independents (r = .856), whereas for Republicans (r = .335) the hypothesis is not supported (see Table A.1). With respect to polls, Table 5 shows that the public actually somewhat underestimates how much attention politicians pay to polling when considering policy decisions. In contrast, journalists slightly overestimate the impact of polling on policymaking.

With respect to who believes that polling influences policymaking, the model for journalists did not a reveal a significant effect of any of the explanatory variables, while for politicians, gender was the only impactful variable (Table 7). Spe-

Table 7 Predictors of whether more or less attention is (or should be) paid to polling

	Public (pay attention)	Public (should pay attention)	Journalists (pay attention)	Politicians (pay attention)
Education[a]	.032	-.224***	NA	NA
Age	.051*	-.024	-.004	.003
Female	-.050	.065	-.086	-.256**
Party registration (ref. Republican)				
Independent	.028	.128	-.072	.082
Democrat	.123	.313**	-.140	.217
Political Ideology (ref. conservative)				
Moderate	.110	-.094	.067	.038
Liberal	.041	-.136	.096	-.006
Survey value	.024	.132***	.071	.051
R^2	.025	.186	.018	.067
(n)	469	445	240	259

****p*<.001 ***p*<.01 **p*<.05

Note. Analyses are based on OLS regression models and connected to Table 5.

[a] Education was not available (NA) for Study 3 and 4.

Data sources: Studies 2 to 4, Kaiser 2001 (General Population, Expert Samples Journalists and Politicians)

cifically, politicians who were male perceived surveys as more important for policymaking than female politicians.

When comparing the different ways, the public can influence political decision-making, the comparison of the three groups of respondents' shows that both expert samples, journalists and politicians, rate conducting public opinion polls more favorably than does the public (see Table 6). Specifically, the ranking questions show that about 50% of both groups of experts rated public opinion polls as the most important source of input from the public. For these two groups, town hall meetings ranked second with a difference of at least 20 percentage points. The differential perceptions of the three groups are also reflected by the correlations, which show that the answers of journalists and politicians correlate at .948 for the rating and .992 for the ranking items, and the correlation between the public and journalists was only .342 for the rating and .469 for the ranking items, and between the public and politicians .608 for the rating and .573 for the ranking items. This result suggests that there seems to be a lack of citizen-elite congruence with respect

Table 8 Predictors of whether conducting a poll is a good way to inform the public

	Public		Journalists		Politicians	
	Rating	Ranking	Rating	Ranking	Rating	Ranking
Education[a]	-.090***	.973	NA	NA	NA	NA
Age	-.001	1.016	.002	.987	.003	1.005
Female	.116**	.949	.105	1.375	-.193*	.456*
Party registration (ref. Republican)						
Independent	.047	.698	.158	3.776	.236*	3.177*
Democrat	.046	.784	.143	2.218	.311**	2.792*
Political Ideology (ref. conservative)						
Moderate	-.047	1.111	-.164	.588	-.096	.427
Liberal	.012	1.563	-.217	.662	-.039	.511
Survey value	.170***	1.433***	.509***	6.863***	.299***	4.511***
R^2	.272	.109	.393	.324	.287	.348
(n)	921	914	241	221	261	246

***p<.001 **p<.01 *p<.05

Note. OLS regression models are based on the ratings in Table 6; logistic regression models are based on the rankings in Table 6. For the logistic regression models, odds ratios and Nagelkerke's Pseudo-R^2 are displayed.

[a] Education was not available (NA) for Study 3 and 4.

Data sources: Studies 2 to 4, Kaiser 2001 (General Population, Expert Samples Journalists and Politicians)

to how the public can best affect political decisions. However, supplementary analyses of party registration for politicians (see Table A.2) show that the rankings and ratings of Republicans are quite similar to the public (r = .908 for the rating, r = .889 for the ranking), whereas there is a lack of congruence for Democrats (r = .553 for the rating, r = .415 for the ranking) and independent politicians (r = .628 for the rating, r = .642 for the ranking).

When considering for whom polling is believed to be the optimal approach for the public to inform policy decision-making, for all three groups-the public, journalists, and politicians-perceived survey value was an important variable (see Table 8). In addition, we observed a negative effect of education and a positive effect for female respondents in the public data set (Study 2), and positive effects for male respondents, Independents, as well as Democratic party registration for politicians (Study 4). Strikingly, the explained variance amounted to 27.2% for the

Table 9 Predictors of whether polling is favored compared to talking to politicians at shopping malls or on the street

	Public	Journalists	Politicians
Education[a]	-.005	NA	NA
Age	.013	-.002	-.014*
Female	.049	.108	-.310*
Party registration (ref. Republican)			
Independent	.002	.399	.490**
Democrat	.074	.465	.488**
Political Ideology (ref. conservative)			
Moderate	-.057	-.344	-.538**
Liberal	.080	-.465	-.499*
Survey value	.136***	.467***	.227**
R^2	.077	.139	.114
(n)	914	241	261

****p*<.001 ***p*<.01 **p*<.05

Note. OLS regression models are based on the *difference* in ratings (see Table 6) between "conducting a public opinion poll" and "talking to people at shopping malls or on the street." For instance, if a respondent answered the first question about polling with 4 "very good" and the second question about talking to politicians at shopping malls with 2 "not too good," the resulting value on the dependent variable would be 4 – 2 = +2. In contrast, if a respondent would rate polls as 1 "not at all good" and talking to politicians at shopping malls as 3 "good," a value of 1 – 3 = -2 would be assigned.

[a] Education was not available (NA) for Study 3 and 4.

Data sources: Studies 2 to 4, Kaiser 2001 (General Population, Expert Samples Journalists and Politicians)

public, 39.3% for journalists, and 34.8% for politicians, suggesting strong explanatory models.

Our data also allowed us to compare the relative attention the various groups believe that policymakers should pay to polling, compared to more informally talking with people in shopping malls and on streets, a policy input method that can be considered to be less comprehensive and scientific, although which some will argue to be more direct and more authentic (see Table 9). That comparison showed that for all three groups the perceived value of surveys had a significant effect on the difference between the answers, meaning that members of the public, journalists, and politicians who perceived surveys as more valuable thought also that they were a preferred method compared to more direct discussions with people at shopping malls or on the street. Solely for politicians, demographics, party registration and political ideology had a significant impact, in addition to their perceptions of polling. Specifically, only Independents and Democrats preferred polling over direct interactions at shopping malls and on the street.

With respect to our second research question about preferences of the two expert groups, the results suggest that the views of the public, journalists, and politicians are relatively similar. Yet, with respect to perceptions of public opinion polling, we found that both expert groups-journalists and politicians-believed that polling has more impact within the democratic process than does the public, which provides support for the legitimization hypothesis.

Further analyses of why town hall meetings are often preferred over public opinion polls by the public but not by elites (see ranking in Table 6) showed that all three groups (public, journalists, and politicians) believe that "polls don't give people the opportunity to say what they really think on an issue" (see Row 1 in Table A.3). Town hall meetings, on the other hand, provide the opportunity of in-depth expression of political positions. Yet, the more negative perception of polls by members of the public compared to elites is likely also grounded in the belief that polling is not always "based on sound scientific evidence" (see Row 2 in Table A.3).

Discussion

Summary of Results

Extending previous research on public representation (e.g., Burstein, 2003; Doherty et al., 2019; MacInnis, Anderson, & Krosnick, 2019), this study examines public preferences regarding the policy information process in the United States while emphasizing the role of public opinion polling as a policy input source for political decision-making. The preferences of the public appear to be contradictory since members of the public aspire to, on the one hand, having more political influence for the electorate, but prefer, on the other hand, direct contact with policymakers, which is less useful for providing politicians with a comprehensive view of public preferences. Specifically, the results disconfirm the democratic representation hypothesis regrading polling and indicate that a majority of the public prefer a direct public-policy link through channels such as town hall meetings rather than the mediated public-polling-policy link through public opinion research.

With respect to the question of who believes public opinion polls can provide a useful contribution to the democratic process, our study showed the expected influence for respondents who perceived surveys as a valuable tool. This finding highlights the importance of perceptions of surveys when understanding the role of public opinion polling in the policy formation process (see De Leeuw et al., 2019; Gengler et al., 2019; Loosveldt & Storms, 2008; Stocké & Langfeldt, 2004).

Especially important from the perspective of actively engaging with society in order to educate people about surveys is the finding that perceptions of survey value appear to be a critical factor. This suggests that the educational efforts of professional public opinion research advocates may be well advised to focus at least as much on the societal value and impact of surveys as on their technical mastery. Building on that, combining surveys with other methods of democratic engagement (Delli Carpini, Cook, & Jacobs, 2004; Skocpol & Fiorina, 1999) might present a way to increase the value of surveys in the democratic process while at the same time offering political decision-makers a comprehensive view of the opinions of the public.

The preference of the public for alternative policy input sources other than polling may indicate that people do not think that standardized, indirect expression through surveys allows them to adequately contribute their opinion on (complex) policy issues. This interpretation is supported by the finding that the public did not think that polls provide the opportunity to say what they really think about an issue. At the same time, not all people think that polls are based on sound scientific methods. Consequently, preferences for alternative policy input sources are likely a mix of perceived shortcomings of polls and the limited role that polling is believed to play within the political decision-making process.

Besides the public opinion data examined, this analysis also included data from expert samples of politicians and journalists. In line with the citizen-elite congruence hypothesis (Hibbing & Theiss-Morse, 2001; André & Depauw, 2017), these additional data illustrated that all three groups see the policy formation process in a relatively consistent manner regarding the importance of various input sources. However, it also showed that public opinion polling as a policy input source was rated more favorably by the two expert groups than by the public. A possible reason for that is that both expert groups actually use public opinion polls for professional activities (legitimization hypothesis). While journalists use it for their news output, politicians rely on them as a source of policy input for their performance evaluation, and to actively shape public opinion (Jacobs & Shapiro, 2000; Shapiro, 2011). Thus, attributing a larger impact to surveys legitimazies their professional attention to them.

When comparing perceptions of polling in the democratic process to less scientific input sources such as connecting with the public at shopping malls and on the streets, we found that members of each of the three groups (the public, journalists, and politicians) who perceive surveys as more valuable understandably also prefer polling as a policy input source. However, only for politicians did party registration and political ideology have an impact as well. Again, the comparison of polling with a less scientific policy input source suggests that the perceived value of surveys appears to be most impactful when considering their democratic contribution.

Exploring the differences across party registration of politicians further, we found that especially Democrats and Independents perceived public opinion polls as the best option for the public to influence political decision making. In contrast, Republicans preferred alternative ways of public engagement. Those substantial differences across political party lines are likely to be even more visible today since American politics have become increasingly polarized (Alwin & Tufis, 2016).

Limitations

One limitation is that our data sets are from around 2000 and, therefore, about 20 years old. However, to our knowledge, no other available data allows the comparison of public perceptions of polling as a policy input source with other policy inputs to those of policy elites such as journalists and politicians. Also, when comparing the findings of our study to reports from other more recent studies from Kantar in 2013 and McClatchy-Marist in 2017, public perceptions about politicians and polling appear to be comparable to our results. Specifically, the Kantar study illustrates that the public still believes the best approach for politicians to obtain input from the public is through town hall meetings (see Online Appendix Table A.4). And the McClatchy-Marist results show that the public does not think that they are well represented while at the same time trust in public opinion polling remains low (see Tables A.5 and A.6). These two data sets are not publicly available and we were unable to access them, so we could not include either in the analyses reported here.

Another limitation, which is connected to the date of data collection, is that new developments in technology and communication are not included. One might think that the introduction of social media may have introduced an essential source of public engagement to the democratic process. However, again the Kantar study suggests that interaction via social media using Facebook or Twitter is considered the least optimal way for politicians to receive valuable policy input (see Table A.4).

A third limitation is that Study 3, the journalists data set, only includes a small number of Republicans and Conservatives. Thus, the results regarding political orientations of journalists should be treated with the necessary caution. Considering that we found substantial party differences regarding perceptions of the role of polling for politicians, the sample composition may have influenced the overall results for journalists in the direction of a more positive view toward public opinion polling.

Finally, limitations of the secondary data sources employed prevent the assessment of other potential explanations for the observed polling paradox. It may be, for example, that citizens who are more actively engaged – those who regularly vote, who attend rallies or town hall meetings, who contact their elected representatives, and/or who more closely monitor public events – see greater value in these approaches, relative to passive reliance on public opinion polling, as the more effective means for influencing public policy. The data sets examined in this paper unfortunately do not include the engagement indicators necessary to explore this possibility. Future research will thus need to address this question.

Conclusion

Our study shows significant differences between ideal and perceived public representation within the political system of the United States. Considering our findings, polling appears to be a straightforward and democratic way that policymakers can increase their attentiveness to public preferences. However, while this view is shared by most politicians (i.e., Democrats and Independents) and journalists, many members of the public paradoxically believe that more direct approaches to engaging with policymakers through town hall meetings and through similar channels are the preferred approach to informing policy decisions. Yet, if at all, the merit of public opinion polling within the democratic process is favored by those people who perceive polls as valuable and have trust in them. Consequently, efforts to improve the publics' perceptions about polling might be best advised to educate people about the function and contribution of polling within the democratic process. Ideally, this would include joint activities of public opinion researchers with journalists and politicians who appreciate the deliberative function of polling within democracies.

References

Alwin, D. F., & Tufiş, P. A. (2016). The changing dynamics of class and culture in American politics: A test of the polarization hypothesis. *The Annals of the American Academy of Political and Social Science*, *663*(1), 229–269.

André, A., & Depauw, S. (2017). The quality of representation and satisfaction with democracy: The consequences of citizen-elite policy and process congruence. *Political Behavior*, *39*(2), 377–397.

Beal, R. S., & Hinckley, R. H. (1984). Presidential decision making and opinion polls. *The Annals of the American Academy of Political and Social Science*, *472*(1), 72–84.

Bowler, S., Donovan, T., & Karp, J. A. (2007). Enraged or engaged? Preferences for direct citizen participation in affluent democracies. *Political Research Quarterly*, *60*(3), 351–362.

Brodie, M., Parmelee, L. F., Brackett, A., & Altman, D. E. (2001). Polling and democracy. *Public Perspective*, *12*(4), 10–24.

Burstein, P. (2003). The impact of public opinion on public policy: A review and an agenda. *Political Research Quarterly*, *56*(1), 29–40.

Burstein, P. (2010). Public opinion, public policy, and democracy. In K. T. Leicht & J. C. Jenkins (Eds.), *Handbook of politics* (pp. 63–79). New York: Springer.

Cayton, A. F. (2017). Consistency versus responsiveness: Do members of congress change positions on specific issues in response to their districts? *Political Research Quarterly*, *70*(1), 3–18.

Converse, P. E. (1964). The nature of belief systems in mass publics. In D. E. Apter (Ed.) *Ideology and discontent* (pp. 201-261). New York: Free Press.

Couper, M. P., & Miller, P. V. (2008). Web survey methods: Introduction. *Public Opinion Quarterly*, *72*(5), 831–835.

Curtin, R., Presser, S., & Singer, E. (2005). Changes in telephone survey nonresponse over the past quarter century. *Public Opinion Quarterly*, *69*(1), 87–98.

de Leeuw, E., Hox, J., Silber, H., Struminskaya, B., & Vis, C. (2019). Development of an international survey attitude scale: Measurement equivalence, reliability, and predictive validity. *Measurement Instruments for the Social Sciences*, *1*(9), 1–10.

Delli Carpini, M. X., Cook, F. L., & Jacobs, L. R. (2004). Public deliberation, discursive participation, and citizen engagement: A review of the empirical literature. *Annual Review of Political Science*, *7*, 315–344.

Delli Carpini, M. X., & Keeter, S. (1996). *What Americans know about politics and why it matters*. New Haven: Yale University Press.

Doherty, D. (2013). To whom do people think representatives should respond: Their district or the country? *Public Opinion Quarterly*, *77*(1), 237–255.

Doherty, D., Bryan, A. C., Willis, R., & Witry, P. (2019). Representation Imperatives in the Public Mind. *Social Science Quarterly*, *100*(6), 1963–1983.

Erikson, R. S. (1976). The relationship between public opinion and state policy: A new look based on some forgotten data. *American Journal of Political Science*, *20*(1), 25–36.

Gengler, J. J., Tessler, M., Lucas, R., & Forney, J. (2021). 'Why Do You Ask?' The Nature and Impacts of Attitudes towards Public Opinion Surveys in the Arab World. *British Journal of Political Science*, *51*(1), 115–136.

Gray, V., Lowery, D., Fellowes, M., & McAtee, A. (2004). Public opinion, public policy, and organized interests in the American states. *Political Research Quarterly*, *57*(3), 411–420.

Heith, D. J. (1998). Staffing the White House public opinion apparatus 1969-1988. *Public Opinion Quarterly*, *62*(2), 165–189.

Herbst, S. (1993). *Numbered voices: How opinion polling has shaped American politics*. Chicago: University of Chicago Press.

Hibbing, J. R. (2001). Process preferences and American politics: What the people want government to be. *American Political Science Review*, *95*(1), 145–153.

Hillygus, D. S. (2011). The evolution of election polling in the United States. *Public Opinion Quarterly*, *75*(5), 962–981.

Jacobs, L. R., & Shapiro, R. Y. (1995). The rise of presidential polling the Nixon White House in historical perspective. *Public Opinion Quarterly 59*(2), 163-195.

Jacobs, L. R., & Shapiro, R. Y. (2000). *Politicians don't pander: Political manipulation and the loss of democratic responsiveness*. Chicago: University of Chicago Press.

Jacobs, L. R., & Shapiro, R. Y. (2005). Polling politics, media, and election campaigns: Introduction. *The Public Opinion Quarterly*, *69*(5), 635–641.

Johnson, T. P. (2018). Legitimacy, wicked problems, and public opinion research. *Public Opinion Quarterly, 83*(3), 614-621.

Joly, J. K., Hofmans, J., & Loewen, P. (2018). Personality and party ideology among politicians. A closer look at political elites from Canada and Belgium. *Frontiers in Psychology*, *9*, 552.

Loosveldt, G., & Joye, D. (2016). Defining and assessing survey climate. In C. Wolf, D. Joye, T. W. Smith, & Y.-c. Fu (Eds.), *The sage handbook of survey methodology* (pp. 67–76). London: Sage.

Loosveldt, G., & Storms, V. (2008). Measuring public opinions about surveys. *International Journal of Public Opinion Research*, *20*(1), 74–89.

Lyberg, I., & Lyberg, L. (1991). Nonresponse research at statistics Sweden. *Paper presented at the Annual Meeting of the American Statistical Association, Atlanta, GA, USA*.

MacInnis, B., Anderson, S. E., & Krosnick, J. A. (2018). *Process approval and democratic legitimacy: How Americans want their elected representatives to decide how to vote*. Retrieved from: https://pprg.stanford.edu/wp-content/uploads/Democratic-Representation-2018.pdf

Moy, P., & Rinke, E. M. (2012). Attitudinal and behavioral consequences of published opinion polls. In C. Holtz-Bacha & J. Strömbäck (Eds.), *Opinion polls and the media* (pp. 225–245). Basingstoke, UK: Palgrave Macmillan.

Newport, F., Shapiro, R. Y., Ayres, W., Belden, N., Fishkin, J., Fung, A., Herbst, S., Lake, C., Page, B., & Page, S. (2013). Polling and democracy: Executive summary of the AAPOR task force report on public opinion and leadership. *Public Opinion Quarterly*, *77*(4), 853–860.

Page, B. I. (1994). Democratic responsiveness? Untangling the links between public opinion and policy. *PS: Political Science & Politics*, *27*(1), 25–29.

Page, B. I., & Shapiro, R. Y. (1983). Effects of public opinion on policy. *American Political Science Review*, *77*(1), 175–190.

Page, B. I., & Shapiro, R. Y. (2010). *The rational public: Fifty years of trends in Americans' policy preferences*. Chicago: University of Chicago Press.

Patterson, T. E. (2005). Of polls, mountains: US journalists and their use of election surveys. *Public Opinion Quarterly*, *69*(5), 716–724.

Percival, G. L., Johnson, M., & Neiman, M. (2009). Representation and local policy: Relating county-level public opinion to policy outputs. *Political Research Quarterly*, *62*(1), 164–177.

Princeton Survey Research Associates. (2001a). *Methodological report: Polling and democracy – Policy and media elite*. Prepared for Henry J. Kaiser Family Foundation. Unpublished report.

Princeton Survey Research Associates. (2001b). *Methodological report: Polling and democracy – Public*. Prepared for Henry J. Kaiser Family Foundation. Unpublished report.

Rosenstiel, T. (2005). Political polling and the new media culture: A case of more being less. *Public Opinion Quarterly*, *69*(5), 698–715.

Shapiro, R. Y. (1998). Public opinion, elites, and democracy. *Critical Review*, *12*(4), 501–528.

Shapiro, R. Y. (2011). Public opinion and American democracy. *Public Opinion Quarterly*, *75*(5), 982–1017.

Sheffer, L., Loewen, P. J., Soroka, S., Walgrave, S., & Sheafer, T. (2018). Nonrepresentative representatives: An experimental study of the decision making of elected politicians. *American Political Science Review*, *112*(2), 302–321.

Skocpol, T., & Fiorina, M. P. (1999). Making sense of the civic engagement debate. In T. Skocpol & M. P. Fiorina (Eds.), *Civic engagement in American democracy* (pp. 1–23). Washington: Brookings Institution.

Sobel, R. (2001). *Impact of public opinion on U.S. foreign policy since Vietnam*. New York: Oxford University Press.

Soroka, S. N. (2002). Issue attributes and agenda-setting by media, the public, and policymakers in Canada. *International Journal of Public Opinion Research*, *14*(3), 264–285.

Stocké, V., & Langfeldt, B. (2004). Effects of survey experience on respondents' attitudes towards surveys. *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique*, *81*(1), 5–32.

Strömbäck, J. (2012). The media and their use of opinion polls: Reflecting and shaping public opinion. In C. Holtz-Bacha & J. Strömbäck (Eds.), *Opinion polls and the media* (pp. 1–22). Basingstoke, UK: Palgrave Macmillan.

Verba, S. (1996). The citizen as respondent: Sample surveys and American democracy presidential address, American Political Science Association, 1995. *American Political Science Review*, *90*(1), 1–7.

Walgrave, S., Sevenans, J., Van Camp, K., & Loewen, P. (2018). What draws politicians' attention? An experimental study of issue framing and its effect on individual political elites. *Political Behavior*, *40*(3), 547–569.

Wlezien, C., & Soroka, S. N. (2012). Political institutions and the opinion–policy link. *West European Politics*, *35*(6), 1407–1432.

Zaller, J. R. (1992). *The nature and origins of mass opinion*. Cambridge: Cambridge University Press.

methods, data, analyses | Vol. 17(2), 2023, pp. 273-302 DOI: 10.12758/mda.2022.10

Measuring Students‘ Reading Behavior with an Ambulatory Assessment – A Field Report on a Smartphone-Based Reading Diary Study

Franziska Maria Locher[1], *Verena Angelika Schnabel*[2], *Valentin Unger*[1] *& Maximilian Pfost*[2]

[1] *Institute for Educational Assessment, St. Gallen University of Teacher Education*

[2] *Department of Educational Research, University of Bamberg*

Abstract

In prior research, reading behavior was predominantly measured using either a questionnaire, which is economical and easy to implement but imprecise, or paper-pencil diaries that document reading behavior quite accurately, but which are time consuming and costly. The present study aims to introduce and evaluate a precise and easy to implement measure of reading behavior, namely a reading diary app in which participants can record their reading behavior on a smartphone. To evaluate the development procedure, the first research question asked whether data gathered with the app is of high quality (e.g., reliability). The second research question asked how reading time recorded via the app is related to reading time assessed via different retrospective questionnaires. $n = 31$ German university students recorded their reading activities for 14 days. Different approaches were applied to estimate the data quality and reliability and yielded satisfactory results. Participants reported more time spent reading daily on the retrospective questionnaire than when recording their reading time using the app. The correlation between reading diary app data and questionnaire data was medium in size. Our findings are discussed in the light of future directions for reading research and the use of ambulatory assessments.

Keywords: ambulatory assessment, reading diary, reading behavior, smartphone app, field report

Being able to effectively process written information is essential for cultural, social, and economic participation in our society. Reading facilitates self-exploration and self-enrichment. Therefore, reading competence is a central skill for today's society (e.g., Alexander, 2005; Artelt et al., 2001; Becker-Mrotzek et al., 2015; Marshall, 2000). The PIRLS 2021 study defines reading literacy as a functional construct capturing readers' ability to process written language in order to achieve personal or socially defined goals. It includes reading to learn from texts, reading to participate in society and reading for enjoyment (e.g., Mullis & Martin, 2019, see also OECD, 2019, for the PISA framework). The reading literacy construct encompasses both cognitive (knowledge and skills) and affective-motivational aspects of reading.

Reading behavior, defined as the sum of all activities related to reading (i.e., time spent reading, amount of reading, or being read to aloud in early childhood), is an important predictor of reading skill development. Many studies have provided convincing evidence of the positive relation between reading skills and reading behavior across the life course (e.g., Burgess, Hecht, & Lonigan, 2002; Bus, van IJzendoorn, & Pellegrini, 1995; Guthrie et al., 1999; Locher & Pfost, 2020; Mol & Bus, 2011; Pfost, Dörfler, & Artelt, 2013). However, beyond the well-replicated general finding of a positive relation between time spent reading and reading skills, there are still large areas of uncharted territory. For instance, there is scare evidence on what kind of reading material (e.g., with respect to text difficulty, content, type of text, writing style) individuals should read to facilitate the optimal development of their reading skills and reading motivation (e.g., Troyer et al., 2018). Thus, gaining deeper knowledge about the nature of people's reading development is of major interest to researchers and practitioners. This concerns above all reading behavior, which, as described above, is one of the most important predictors of reading skills. People at all stages of reading literacy development can face difficulties while reading: for example, while beginning readers might struggle to decode letters, advanced readers might struggle to extract information and construct meaning from the text (e.g., Chall, 1983; Kutner et al., 2007; OECD, 2021). Thus, it is important that research on reading does not end in adolescence. The better researchers

Acknowledgements

This research was supported by grants from the Bamberg Graduate School of Social Sciences (BAGSS) and the German Research Foundation (DFG: Grant number PF 840/2-2). We have no conflicts of interest to disclose.

We want to thank Peter Kuntner for his technical support in developing the reading diary app.

Direct correspondence to

Franziska Maria Locher, Institute for Educational Assessment,
St. Gallen University of Teacher Education, 9000 St.Gallen, Switzerland
E-mail: franziska.locher@phsg.ch

understand reading at different stages of individual development, the better interventions or instructional materials practitioners can develop to support readers in facing the challenges they encounter in later stages (e.g., Alexander, 2005).

Currently, in reading research as well as in psychology in general, most studies use global retrospective self-report data from questionnaires (e.g., an evaluation of average reading time per week; Fahrenberg et al., 2007b). Research questions such as the one above, however, can only be answered by taking a closer look at individuals' "real" reading activities, rather than merely relying on global retrospective measures that only provide information about average trends.

Continuous assessments of people's reading behavior (e.g., daily reading diaries) are seldom used because daily logs tend to be very time-consuming and can be a huge burden for participants, especially in paper-pencil studies. Therefore, the goal of this study was to develop a reading diary app for participants to record their reading behavior (e.g., reading time, reading material) and reading motivation with a smartphone in an economical way. In addition, the present study explores the reading behavior data collected via this smartphone app. The data on reading motivation will be analyzed in a further research project. The first aim was to examine the quality of the data (e.g., reliability) gathered with the reading diary app (ambulatory assessment). The second aim was to investigate how reading time assessed with the reading diary app is related to reading time assessed with global and retrospective questionnaire measures.

Theoretical Background

Conceptualization of Reading Behavior and How to Measure it

Reading behavior can be defined as the sum of all activities related to reading. As this definition can potentially include a wide range of reading-related activities, previous studies have operationalized reading behavior in many different ways. In order to clarify the concept of reading behavior, it seems worthwhile to differentiate between the quantitative aspects ("How much do people read?") and qualitative aspects ("What do people read?") of reading (Locher, Becker, & Pfost, 2019a). Quantitative aspects of reading behavior refer to the amount or volume of reading (e.g., number of books read in the last month) or time spent reading. Qualitative aspects of reading behavior are multifaceted. They comprise information about the nature of the reading material (e.g., type of text, text difficulty, text content, or medium, i.e., print or digital). Common ways to measure the quantitative aspect of reading behavior are global and retrospective self-reports of reading time (e.g., "About how much time do you usually spend reading outside of school?") such as those used in PISA (Programme for International Student Assessment; OECD, 2010), one of the largest and perhaps most well-established international large-scale comparison studies in the field of education.

However, in addition to these global self-report scales of time spent reading, recent research has provided evidence of differential effects of reading different types of texts on variables such as reading motivation or reading skills (e.g., Locher et al., 2019a; Jerrim & Moss, 2019; McGeown et al., 2015; McGeown et al., 2016; Pfost et al. 2013). For example, reading traditional fiction books (e.g., novels, short stories or tales) has been found to be more important for reading skill development than reading comics and newspapers or online media (e.g., Pfost et al., 2013). The finding that the type of text moderates the relation between reading behavior and reading skills was further supported by Jerrim and Moss (2019) in an analysis of PISA data: students who frequently read fiction books had better reading skills than their peers who do not read fiction. The authors did not find such an effect for other text types, such as magazines or non-fiction. Furthermore, with respect to reading motivation, recent research provides first evidence that reading classic literature, especially in comparison to modern fiction books, negatively relates to intrinsic situational reading motivation (Locher et al., 2019a). In addition, within the school context, students who read more difficult books were less motivated to read (Locher et al., 2019a). These results illustrate that more detailed insight into reading behavior is desirable.

Exploring qualitative aspects of students' reading behavior has often been neglected, probably because assessing such information comes at a high cost. Therefore, measures capturing the amount of time people spend reading different types of texts are a good complement to the global evaluation (Locher & Pfost, 2019b). Beyond the quantitative aspect of reading time, these measures provide at least some additional information about the average amount of time individuals spend reading fiction books, nonfiction books, newspapers, or other text types. Nevertheless, whether global retrospective self-reports from questionnaires (global and text type-specific measures of people's reading time) accurately capture individuals' behavior is doubtful, because this methodology "records mental representations rather than the actual experience and behavior" (Fahrenberg et al., 2007b, p. 207), and several potential biases might occur (Fahrenberg et al., 2007a). Data can be affected by cognitive schemata, response tendencies, judgment heuristics, or memory effects (Fahrenberg et al., 2007a; Gershuny, 2012). For instance, when determining daily reading time, some people might use the last week as a reference, whereas others might use the past month. Another source of bias might arise when the days selected are not representative with respect to the behavior being assessed (Kan & Pudney, 2008). The experience of time is also subjective, as persons perceive time use differently (Juster, Ono, & Stafford, 2003). This means that individuals' responses to a question about their normal daily reading time might be based on different heuristics. Moreover, individuals might only remember longer reading activities (e.g., reading a book for 3 hours on the weekend) and fail to recall brief reading activities (e.g., reading a newspaper for 5 minutes in a waiting room). The fact that people might not consider all of their reading activities might lead to biases in response behavior (memory effect). Another issue is that recalling all reading activities correctly and assigning them to the appropriate text category listed in the questionnaire can be a very difficult task, especially for children and young adolescents (Locher & Pfost, 2019b).

Thus, alternative approaches are required to obtain data that better captures a person's actual reading behavior. Options used in research fields such as psychology include the experience sampling method, which asks about experiences in the moment (e.g., Csikszentmihalyi & Larson, 2014; Hektner, Schmidt, & Csikszentmihalyi, 2007; Shumow, Schmidt, & Kackar, 2008; Zirkel, Garcia, & Murphy, 2015), as well as the day reconstruction method, in which participants reflect on their activities that day (e.g., Kahneman et al., 2004; Lucas et al., 2019). A third way to gather information that more closely approximates people's "real" reading behavior and is less error-prone is to use reading diaries in which people document their reading activities, namely how long and which books, magazines, newspapers, or other texts they read (e.g. Akbar et al., 2015; Anderson, Wilson, & Fielding, 1988; Nieuwenboom, 2008; Stoffelsma, 2018). Reading diaries are often seen as a kind of gold standard because they offer a quite precise documentation of people's reading behavior, provide concrete information about the books or texts the person read, and yield information that can be used in further analyses.

Paper-Pencil versus Digital Diaries – Using an Ambulatory Assessment

Most daily diary studies in psychology and educational research have relied on paper-pencil methods (Akbar et al., 2015; Bolger, Davis, & Rafaeli, 2003; Fahrenberg et al., 2007b; Wilhelm, Perrez, & Pawlik, 2012). However, this method is quite time- and space-consuming and can also be a huge burden for participants (Bolger et al., 2003). For example, people have to carry their documents/diaries with them at all times, or might not have their diary available when they need it due to the cumbersome nature of the paper documents. In the worst case, this results in missing information. Another possibility is that people enter their reading activities later. However, filling out the diary after too much time has passed increases the risk of a retrospective bias, as people have to estimate their activities (Bolger et al., 2003; Wilhelm et al., 2012). This drawback also applies to the day reconstruction method and so-called "end-of-day diaries", in which all activities are documented once a day. Therefore, although reading diaries are seen as the gold standard, they are seldom used as a method for continuously assessing people's reading behavior. One way to deal with the issues associated with the paper-pencil method is electronic documentation of reading behavior via an ambulatory assessment. This promising and innovative method “refers to the use of computer-assisted methodology for self-reports, behavior records, or physiological measurements, while the participant undergoes normal daily activities” (Fahrenberg et al., 2007b, p. 206). In other words, ambulatory assessments aim to conduct research (i.e., monitoring people’s psychological, emotional, behavioral, or biological processes) in daily life and in people’s natural environment with digital assistance (Trull & Ebner-Priemer, 2014). In particular, smartphones have become more and more important for ambulatory assessments because a large amount of data can be collected very economically and easily via apps. In addition, there is no need to carry around anything extra, like a paper-pencil diary (Conner & Lehman, 2012; Miller, 2012; Trull & Ebner-Priemer, 2014; Zhang et al., 2018). Due to these advantages, apps have been used to measure behavior in various disciplines, for example in the field of health (e.g., Ahram, 2019; Ebner-Priemer et al., 2007; Glomann et al., 2019; McLaws et al., 1990).

In summary, an ambulatory assessment to assess reading behavior (e.g., a specific reading diary app) has numerous advantages over paper-pencil diaries. First, most adolescents and young adults use smartphones and therefore already carry one with them at all times (Lampert, Sygusch, & Schlack, 2007). This means that data can be collected in daily life at low cost, and participants do not need extra materials or extra recording devices or computers (Fahrenberg et al., 2007b). Second, participants can easily fill out the reading diary whenever and wherever they want. This means that it is much more convenient and requires less effort for participants to document their activities, presumably leading to better data quality. Third, questions and questionnaires can be adapted based on people’s responses, opening up a broader range of possibilities and creating flexibility (Fahrenberg et al., 2007b). For example, participants can only be given information and questions that are relevant for them. This approach reduces the text load in digital reading diaries and may result in less workload, as everything “unimportant” can be hidden and data entry is made easier. Fourth, with paper-pencil diaries, researchers would likely never know if participants filled out the whole diary retrospectively at the very end of the study. In ambulatory assessment/reading diary app, researchers have better control over the timing and reliability of the entries, meaning that the use of electronic data should result in higher compliance (Bolger et al., 2003; Fahrenberg et al., 2007b). And fifth, it might be easier to recruit more study participants for studies using smartphone apps than paper-pencil diaries (Zhang et al., 2018). This could directly affect the generalizability of the diary data collected.

Of course, the use of diaries also comes with several challenges. Depending on the type of research question, software and programming costs can occur (Conner & Lehman, 2012). Thus, developing an app could be more costly and would probably require more resources than developing and implementing a questionnaire or a paper-pencil reading diary. In addition, completing a digital diary requires certain technical skills that are not necessarily present in all individuals within society and thus may lead to sampling bias. Finally, it should be noted that apps can always produce errors (e.g., in data storage or data transmission).

Despite the huge potential of ambulatory assessments, we do not know of any studies that have used this approach to collect data on people's reading behavior in daily life. At least two exploratory studies used electronic diaries to assess reading behavior. Keller (2010) had 12 college students document their reading behavior over three days by taking digital photographs of their reading activities (e.g., pictures of books). Raith (2008) had ninth-grade students use weblogs to document their reading behavior. This qualitative study compared paper-pencil and weblog reading dairies, finding that students who documented their reading behavior with weblogs reflected on the content of the book better than students who documented the content with paper-pencil diaries. However, both studies used desktop computers and did not leverage the possible advantages of ambulatory assessment using personal smartphones to monitor daily reading behavior (e.g., flexibility, diary always readily available, thus yielding a large amount of precisely documented data).

Measurment and Data Quality in Ambulatory Assessments

A sufficient measurement and data quality is an important precondition that is needed for further analyses and the correct interpretion of results. Therefore, as for any empirical measure, in ambulatory assessment studies it is important to evaluate the quality of the collected diary data and measures used (Calamia, 2019). To date, only a small number of studies using ambulatory assessments have reported quality criteria such as the reliability of the self-reported data (Calamia, 2019). For instance, well-regarded diary studies such as Greaney and Hegarty (1987) or Allen, Cipielewski, and Stanovich (1992) lack information on measurement and data quality. This might be because there are no clear standards for evaluating measures and data quality within ambulatory assessments like there are for survey scales or test development.

One way of evaluating data quality in ambulatory assessments is to use postmonitoring interviews (e.g., reaction quesionnaires: Nieuwenboom, 2008; Stone et al., 2003). In such questionnaires, which are conducted after the monitoring period, participants answer questions about, for instance, whether they found the ambulatory assessment to be a huge burden, whether they think they behaved differently in some situations because of the ambulatory assessment, or whether they think their behavior differed from their average behavior throughout the ambulatory assessment. If this is the case (i.e., the majority of participants agree with the statements), the researcher must conclude that the data quality is not acceptable. Postmonitoring questionnaires therefore provide some indication of the reliability and validity (e.g., does the fact of observation change the nature of the phenomenon being measured) or generalisability of a given assessment.

Another method for demonstrating measurement and data quality in reading diaries or ambulatory assessments (e.g., Anderson et al., 1988; McLaws et al., 1990; Nieuwenboom, 2008) is to demonstrate the reliability of the reading diary app by examining correlations in the goal construct (e.g., reading time) between the observed weeks as well as between even- and odd-numbered days. This approach is similar to the idea of split-half reliability, where a test is divided into two halves, scores on which should be correlated with one another. Furthermore, reliability can be demonstrated by examining internal consistency (e.g., Cronbach's alpha values for the amount of time spent reading per day across all days).

Study Aims and Research Questions

Two research questions were formulated.

(a) First: Does the ambulatory assessment of reading behavior via a smartphone-based diary app have satisfactory measurement and data quality? For an optimal quality check, existing approaches used in previous studies were combined. Measurement and data quality can be assumed to be sufficient when two conditions hold: 1) postmonitoring questionnaire results reveal no issues, and 2) the reading diary app measure is sufficiently reliable. The first condition is achieved if only a small proportion of study participants report irregularities and problems with app use (e.g. Category 4: "I very often forgot to make an entry in the diary"). With respect to the second condition, in accordance with previous research examining the reliability of diary measures (e.g., Anderson et al., 1988; McLaws et al., 1990; Nieuwenboom, 2008), the reading diary app measure can be assumed to be reliable if the correlation in reading time between Week 1 and Week 2 and between even- and odd-numbered days is around $r = .70$ or higher. Finally, our reading diary app measure can be assumed to be reliable if Cronbach's alpha (for the amount of time spent reading per day across all days) exceeds $\alpha = .80$.

(b) Second: How is reading time measured via the ambulatory assessment related to reading time measured via different global retrospective questionnaire measures? An important criterion for the quality of an instrument is construct validity, and convergent validity is one way to check for construct validity. Convergent validity refers to overlap in the results of different tests for the same or similar constructs (Moosbrugger & Kelava, 2012). Thus, high correlations between the results of two tests or measures of the same construct reflect high convergent validity (Pospeschill, 2010). Global retrospective questionnaire measures are widely used and therefore can be considered well established; thus, they seem to be well-suited as a criterion for testing the quality of our reading diary app data. According to the literature, however, data obtained from reading diaries most closely reflect “real” reading behavior. Therefore, low correlations between these two measures may not necessarily indicate low validity of the reading diary app data, but may also indicate that the two instruments partially measure different constructs. Consequently, this second research question is examined in an exploratory manner.

Method

Participants

All analyses rely on data from a convenience sample of n = 31[1] German university students (77% women) with a mean age of 20.71 (SD = 2.60) years. The university students were in their third semester of higher education on average (M = 3.01); 23% had an immigrant background, meaning that at least one parent was born abroad. Nearly all students (n= 28) had taken courses in the fields of psychology and education science. The participants had received an average grade of 2 (10-12 points) in German language arts on their secondary school completion exams (Abitur), which reflects “good” performance according to the German grading system.

Study Design

The study included three measurement points.

First Measurement Point (M1). The first measurement point involved a one-hour session in small groups of three to nine study participants. A reading achievement test was administered at the beginning of the session. Afterwards, participants filled out a questionnaire with different reading behavior measures, which lasted about 15 minutes. Afterwards, the reading diary app was introduced and explained. The app was installed on the smartphones the participants had brought with them, and the app’s functions were tested. Finally, the further study procedure was explained and participants were instructed on how to use the app (see measures section for further details). The first session followed a standardized script.

Second Measurement Point (M2). The second measurement point represents the ambulatory assessment period via the reading diary app. For the ambulatory

[1] The study began with a total of 35 participants. Three participants could not install the app on their smartphone and could thus not further participate in the study. It later turned out that the reading diary app did not work on devices with an older Android operating system. Of the 32 participants who used the reading diary app during the two-week survey period, one person did not fill out the final questionnaire and therefore was not considered either in the further analyses.

assessment, we followed an event-based design in which every reading activity was documented immediately after participants completed a reading event. All participants recorded their reading activities for 14 days. Based on arguments made by Foasberg (2014), the start of the ambulatory assessment period was placed in the middle of the semester so that there would be no bias due to final exam stress.

Third Measurement Point (M3). The day after the reading diary app was used for the last time (Day 15), participants received a link via e-mail to an online questionnaire created using SoSci Survey (Leiner, 2019). The questionnaire included different reading behavior measures as well as the postmonitoring questionnaire. The participants had four days to complete the online questionnaire, which lasted about ten minutes. The incentives for complete participation were made available the following week. We elected to conduct an online survey at M3. Participants were not required to return to the lab, reducing the effort required. This was of high importance in order to avoid non-response and missing data in the postmonitoring questionnaire.

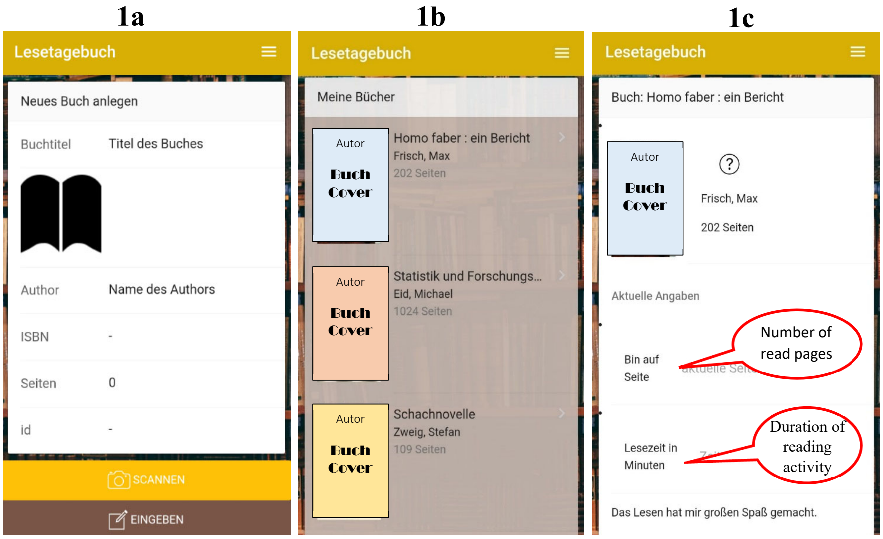
The study was advertised and participants were recruited in university seminars, lectures and via the student council email list at the University of Bamberg. For technical reasons, only students with Android devices could participate in the study. For complete participation in the study, students received a 15€ voucher for a local bookstore. Bachelor's degree students in psychology could alternatively opt to receive four credit hours for their participation.

Measures

Reading diary app data

The reading diary app was developed by the authors of the present paper at the Department of Educational Research at the University of Bamberg for the purpose of the present study. The app has a clear structure and can be used intuitively[2]. As previously mentioned, the study followed an event-based design, meaning that participants were to document their reading behavior directly after each reading activity occurred. Personal communication such as emails and text messages were not to be taken into account. No restrictions were applied concerning text type, medium (print or digital device), or whether the reading activity was for enjoyment or for one's studies, and participants were requested to document all reading activities. Furthermore, browsing and looking things up on the internet was not explicitly excluded.

[2] At M3, participants answered questions about user friendliness (e.g. "Scanning new books worked without problems", "The entries for reading were uncomplicated"). The feedback was good. On a four-point Likert scale ("strongly disagree" to "strongly agree"), participants gave each statement an average rating of M = 3.0.



Neues Buch anlegen

Buchtitel	Titel des Buches
Author	Name des Authors
ISBN	-
Seiten	0
id	-

Meine Bücher

Autor Buch Cover — Homo faber : ein Bericht, Frisch, Max, 202 Seiten

Autor Buch Cover — Statistik und Forschungs..., Eid, Michael, 1024 Seiten

Autor Buch Cover — Schachnovelle, Zweig, Stefan, 109 Seiten

Buch: Homo faber : ein Bericht

Autor Buch Cover — Frisch, Max, 202 Seiten

Aktuelle Angaben

Bin auf Seite

Lesezeit in Minuten

Das Lesen hat mir großen Spaß gemacht.

Figure 1 Screenshots showing the interface of the reading diary app

See Figure 1 for an illustration of what the reading diary app looked like. The key element of the reading diary app is each participant's personal library. Participants were able to add books/texts to their personal libraries by scanning the barcode on the back of the book with an app tool (see Figure 1a). This automatically entered book-related data such as book title, author's name, ISBN number, and the number of pages in the app, as the app was able to pull this information from the German National Library. Participants also had the opportunity to type in the title and author of the book manually. This also applied to all other texts (e.g., newspapers, digital articles, magazines) where no barcode was available. Then, every time the participant opened the reading diary app, his or her personal library appeared (see Figure 1b), encompassing all previously added books and texts (labeled "reading projects"). To measure the time participants spent reading, they were asked to make an entry every time they completed a reading event/activity. To make such an entry, participants first chose the reading project (i.e., the book or text they had read) from their library, and then they answered a short question about the duration of the reading event (in minutes) and the number of pages they had read (see Figure 1c). Participants were also asked to answer four short questions regarding aspects of situational reading motivation every time they completed a reading event. Log data provided additional information about the date and time of day when participants indicated they had read something (i.e., when they made an entry). All elements in the library (books or other texts) were recorded and visible to the participants until they had completed a reading project. When participants indicated that they had finished a reading project (as well as at the end of the two weeks of using the app), they had to conclude the reading project. To do so, they had to answer questions about the reading project in general, such as the reading purpose.

Reading Behavior Measures from the Paper-pencil Questionnaire

Global evaluation of reading time. The global evaluation of reading time was captured in a manner comparable to the PISA study (Hertel, Hochweber, Mildner, Steinert, & Jude, 2014) by asking participants to answer the following question: "How much time do you normally spend reading per day?" A 5-point Likert scale was used (1 = *never*, 2 = *up to 30 min*, 3 = *between half an hour and 1 hour*, 4 = *1 to 2 hours*, 5 = *more than 2 hours*). The global evaluation of reading time was measured in the M1 and M3 questionnaires.

Evaluation of reading time for different types of texts. Equivalent to the global evaluation, the evaluation of reading time with respect to different types of texts was captured with the item: "How much time do you spend per day reading the following text types?," again on a 5-point Likert scale (1 = *never* to 5 = *more than 2 hours*). This time, however, participants were asked to indicate how much time they spent per day reading different types of texts. Similarly to the PISA study, this study asked about the following categories: (a) fiction books, (b) nonfiction books, (c) newspapers, (d) magazines, and (d) comic books. This variable was also measured at M1 and M3. Although more text types exist than the five categories mentioned, a recent study by Locher & Pfost (2019b) showed that a too fine-grained differentiation between text types tends to become counterproductive. Therefore, these broad text categories were used.

Comparative Reading Habits (CRH)

The CRH is a measure of reading habits developed by Acheson, Wells, and MacDonald (2008). Participants were asked to rate their reading habits in comparison with their peers (e.g., reading time: "Compared to other college students, how much time do you spend reading all types of materials?" or reading speed: "Compared to other college students, how fast do you normally read?"). For each of the five questions on the CRH, participants chose a number on a scale ranging from 1 to 7, with higher numbers indicating greater amounts of the quantity in question (e.g., reading time, speed). Similarly to the study by Acheson et al. (2008), a 7-point Likert scale was used to ensure sufficient variance in responses. The CRH was measured at M1 only.

Postmonitoring Questionnaire

Conscientiousness. To check whether participants regularly documented their reading activities, we asked participants at M3: "How regularly did you make entries in the app after reading?” They were asked to rate this question on a 4-point Likert scale (1 = *forgot very often*, 2 = *sometimes forgot*, 3 = *regularly documented*, 4 = *always documented*).

Generalizability. To check whether participants’ reading activities during the 2 weeks of using the reading diary app were comparable to their normal daily reading habits, participants were asked the following question: “Do you think you spent more or less time reading during the 2 weeks of using the reading diary app than you normally do?” A 5-point Likert scale (1 = *much less*, 2 = *a bit less*, 3 = *exactly the same*, 4 = *a bit more*, 5 = *much more*) was used to ensure sufficient differentiation.

Reaction. Three items were used to check for possible reaction effects caused by the reading diary app. The first item refers to boredom effects (“Filling out the reading diary app was boring”), the second item to the burden (“Filling out the reading diary app was a burden in everyday life”), and the third item to an unintentional intervention effect (“Filling out the reading diary app influenced my usual reading behavior”). All three items were rated on a 4-point Likert scale (1 = *disagree*, 2 = *somewhat disagree*, 3 = *somewhat agree*, 4 = *agree*).

Analysis Strategy

To explore the correlation between reading time in Weeks 1 and 2 and on even- and odd-numbered days, and to examine internal consistency, the total duration of all reading events each person documented throughout the 14 days was aggregated. In so doing, information was collected about the amount of time participants spent reading on each individual day as well as during each of the 2 weeks. In general, internal consistency measures whether different items that aim to measure the same construct produce similar scores. To compute the internal consistency of reading time as measured in the reading diary, the reading time on each day was treated as a single “item” measuring the same construct, namely, the amount of time spent reading.

For the second research question regarding the relation between reading time measured via the reading diary app and reading time measured via the questionnaire, the reading time data from the ambulatory assessment was transformed. In a first step, to differentiate between the different types of texts, each reading project from a participant’s personal library was assigned to one of the categories from the questionnaire: fiction books, nonfiction books, newspapers, magazines, or comics. Some reading projects could not clearly be assigned to one of the text categories (e.g., lecture notes from university courses). These titles formed the category “other books” or the category “other texts.” While categorizing the reading projects, it became apparent that the app failed to transfer title names from a substantial number of manually added reading projects to the server. Due to this technical problem, these reading projects could not be categorized. These titles formed the category “texts with missing title.” In a second step, daily reading time in minutes from the reading diary app was classified into one of the five response categories from the paper-pencil questionnaire: 1 = *never*, 2 = *up to 30 min*, 3 = *between half an hour and 1 hour*, 4 = *1 to 2 hours*, 5 = *more than 2 hours*. This made it possible to compute the average daily reading time across the 2 weeks on a categorical level. This was also done separately for each text type. After preparing the data in this manner, repeated-measures ANOVAs and correlation analyses were computed in SPSS (IBM-Corporation, 2012).

Results

Before presenting the results for the two research questions, some descriptive results will be highlighted to provide a first impression of the information we were able to collect with the reading diary app.

Descriptive Results for the Reading Diary Data

A total of 416 event-based entries were made during the ambulatory assessment. This means the 31 participants indicated that they had engaged in reading activities 416 times during the 2-week period. Figure 2 shows the times of day that were the most popular reading times. Most reading events were documented between 7 pm and 12 am, meaning that people mostly indicated spending time reading in the evening. Most reading and most entries were made on Mondays. Other than that, peak reading time was rather equally distributed (see Figures A and B from the electronic supplement for further information). On average, participants documented one reading event per day (M = 0.96, SD = 0.66). These reading events lasted an average of M = 31.78 min (SD = 16.10). The duration of a reading event ranged from 5 to 240 min. Approximately 17% of all reading events lasted 10 min or less, while 16% of all reading events lasted 1 hour or longer.

During the 2 weeks of the ambulatory assessment, participants added an average of four books (M = 3.61, SD = 4.19) and two additional texts, meaning newspapers, magazines, online articles, and so forth (M = 2.10, SD = 2.17), to their library. On average, four reading events per book (M = 4.23, SD = 3.90, Min = 1, Max = 19) and two reading events per text (M = 2.28, SD = 1.61, Min = 1, Max = 16) were recorded. Some participants did not document any reading time for some of the reading projects they entered into their library, meaning they did not read every book/text they entered. Therefore, Table 1 shows the number of reading projects overall and the number of reading projects with valid reading times. One possible explanation for this is that participants added some books they planned to read into

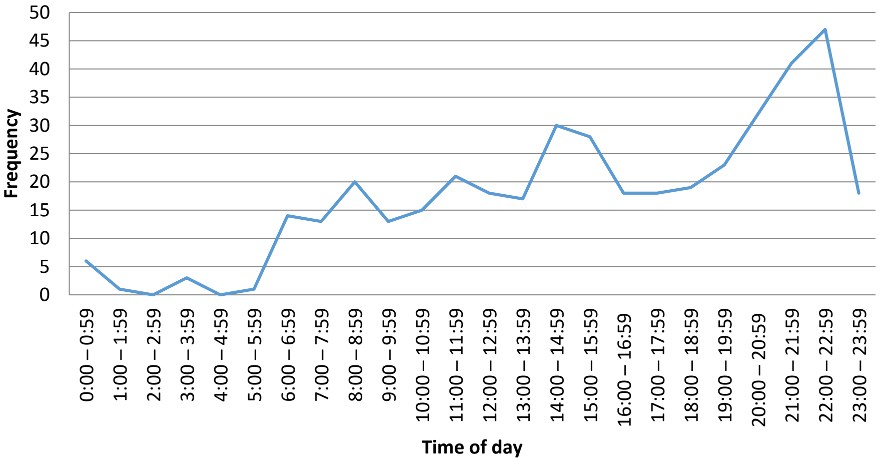


Figure 2 Frequencies of entries in the reading diary app per time of day

their library in advance, but then they did not actually spend time reading them. In-depth analyses revealed that it was often nonfiction books that were added to the library but had no reading time (see Table 1). Furthermore, the students added a higher number of books than other texts (e.g., newspapers) during the 2 weeks.

Table 1 Numbers of Reading Projects by Type of Text

Type of reading project	Type of text	*N*	n_T
Books	Fiction	46	42
	Nonfiction	54	36
	Other books	12	11
	Sum of books	112	89
Texts	Newspapers and magazines	15	13
	Other texts	15	15
	Texts with missing titles	35	33
	Sum of texts	65	61
Sum of reading projects		177	150

Note. N = Number of reading projects overall, n_T = Number of reading projects with valid reading time.

Table 2 Average Daily Reading Time in Minutes By Text Type Measured via the Reading Diary App

	n	*M*	*SD*	Min	Max
Fiction	42	16.82	25.37	0.00	119.29
Nonfiction	36	3.88	11.92	0.00	66.43
Newspapers	4	0.72	2.18	0.00	9.29
Magazines	9	1.61	5.81	0.00	32.14
Comics	1	0.03	0.19	0.00	1.07
Other books	11	1.81	4.35	0.00	16.43
Other texts	14	1.12	2.96	0.00	15.00
Texts with missing title	33	5.06	9.91	0.00	49.86
All	150	31.05	32.20	0.00	139.29

Note. $N = 31$ participants. n = number of reading projects to which this reading time can be subsumed.

Table 2 shows the average time people spent reading across all types of texts as well as differentiated by type of text. First, the results showed that on average, participants spent about half an hour a day ($M = 31.05$, $SD = 32.20$) reading. Differentiated by type of text, participants predominantly spent time reading fiction ($M = 16.82$, $SD = 25.37$) and nonfiction books ($M = 3.88$, $SD = 11.92$), whereas other types of texts such as newspapers and magazines were only read for a few minutes a day. On average, participants spent more time (in minutes) reading in the first week ($M = 34.18$, $SD = 37.90$) than in the second week ($M = 27.91$, $SD = 32.45$). However, this difference was not significant, $t(30) = 1.21$, $p > .05$, $r = .67$. Individual differences between participants in average reading time were large, as seen in the large standard deviation.

Measurement and Data Quality

Detailed results of the postmonitoring questionnaire can be found in Table 3. With respect to conscientiousness, only one person indicated that he or she often forgot to make entries in the reading diary app (= response option 1), whereas more than 75% stated that they regularly or almost always made entries (response option 3 or 4). Once again, the categories were 1 = "*forgot very often*", 2 = "*sometimes forgot*", 3 = "*regularly documented*", 4 = "*always documented*". Regarding the generalizability of the documented reading activities, the results of the postmonitoring questionnaire were satisfactory. Participants indicated how their reading activities during the ambulatory assessment compared to their normal daily reading habits

Table 3 Descriptive Statistics from the Postmonitoring Questionnaire (M3)

	M	*SD*	Cat 1 %	Cat 2 %	Cat 3 %	Cat 4 %	Cat 5 %
Conscientiousness	3.26	0.89	3.2	19.4	25.8	51.6	-
Generalizability	2.97	0.91	3.2	32.3	29.0	35.5	0.0
Reaction 1: Boredom	2.26	0.82	16.1	48.4	29.0	6.5	-
Reaction 2: Burden	1.68	0.65	41.9	48.4	9.7	0.0	-
Reaction 3: Intervention	2.23	0.81	19.4	41.9	35.5	3.2	-

Note. Data were collected from $N = 31$ participants. Cat in %= percentage of people selecting this category. Conscientiousness: Category 1 = *forgot very often* to Category 4 = *always documented*; Generalizability: Category 1 = *much less* to Category 5 = *much more*; Reaction: Category 1 = *disagree* to Category 4 = *agree.*

with the following categories: 1 = “*much less*”, 2 = “*somewhat less*”, 3 = “*exactly the same*”, 4 = “*a bit more*”, 5 = “*much more*”. About 29% of participants indicated that they spent exactly the same amount of time reading during these 2 weeks compared with their usual reading time. A total of 32% stated that they usually read slightly less and 3% much less than in the 2 weeks of data collection. On the other hand, 36% of participants indicated that they read slightly more than they usually read. Consequently, participants’ reading time during the 2 weeks of data collection seemed to be comparable on average to participants’ usual reading activities. With respect to the reaction items (categories: 1 = “*disagree*”, 2 = “*rather disagree*”, 3 = “*rather agree*”, 4 = “*agree*”*)*, only 7% of participants indicated that they got bored (Reaction 1) while using the reading diary app, whereas 65% stated that they were not bored or only slightly bored. About 90% of participants disagreed or somewhat disagreed that the task of monitoring their reading behavior with the reading diary app every day was a burden (Reaction 2). Furthermore, only one participant indicated that the ambulatory assessment influenced his or her daily reading behavior (i.e., he or she read a lot more than average). About 60% of participants stated that they did not think they changed their reading behavior due to the ambulatory assessment (Reaction 3).

As an additional quality measure, we calculated the internal consistency for reading time across all days. The reading diary app data had satisfactory internal consistence ($\alpha = .87$). This was also supported by the correlations between the sum total daily reading time on even- and odd-numbered days ($r = .81$, $p < .01$) and between the sum total of reading time in Weeks 1 and 2 ($r = .67$, $p < .01$). Both results serve as indicators of reliability.

Comparing Different Reading Behavior Measures

To compare the reading diary app data with the global retrospective questionnaire data, we used information about daily reading time from the event-based entries during the 2 assessment weeks. This data was transformed to match the response categories for the questionnaire items. Because the number of reading projects in the categories of newspapers and magazines was too small to analyze separately, a joint category was built for both the app and questionnaire data. No analyses were conducted regarding comic books because only one comic book was mentioned in the reading diary app.

Comparing the average amount of reading time per week measured as a global evaluation on the questionnaire and the average weekly reading time measured via the reading diary app (Table 5 and Figure 3) yielded a significantly lower average for the reading diary app data ($M = 2.06$, $SD = 0.70$) compared to the questionnaire data (M1: $M = 3.29$, $SD = 0.97$; M3: $M = 2.94$, $SD = 0.85$).

Furthermore, there were significant differences in the global retrospective questionnaire measure before (M1) and after (M3) the 2 weeks of reading behavior documentation, with participants indicating less time spent reading after the ambulatory assessment period. Comparable results were found when differentiating between text types (Table 4). For all text categories, the evaluation of reading time before using the reading diary app was descriptively but not significantly higher than the evaluation after the ambulatory assessment period. Turning to global reading time (i.e., summing up all reading activities across all types of texts), no participants indicated that they spent no time reading when asked in the questionnaire. However, according to the reading diary data, 23% of participants fell into the lowest category, which meant that on most days during the 2 survey weeks, they did not spend any time reading. The results of the correlational analyses in Table 5 revealed that the reading diary app data were significantly associated with the global retrospective evaluation from the questionnaire (M1: $r = .39$, $p < .05$ and M3: $r = .58$, $p < .01$). The global retrospective evaluation *after* the 2 weeks of data collection was more strongly related to the reading diary app data than the global retrospective evaluation *before* the 2 weeks of data collection. Comparable results were found when differentiating by type of text, with the exception of the newspapers and magazines category. Table 5 also shows the correlations with the CRH scale. One can see that reading time measured via the reading diary app was significantly correlated with the CRH ($r = .38$, $p < .05$). As the internal consistency of the CRH was quite low ($\alpha = .49$), and the CRH also includes items that refer to reading skills and reading speed, we additionally computed correlations with the item referring to reading time only ("Compared to other college students, how much time do you spend reading all types of materials?"). This item had a considerably stronger correlation with the reading diary app data ($r = .51$, $p < .01$).

Table 4 Average Time Spent Reading (Questionnaire and App Data) in General and by Type of Text

Type of text		*M*	*SD*	M1- M2	M2-M3	M1-M3	Never (%)	> 2 hr (%)
Fiction ($n = 42$)	M1: Questionnaire	2.84	0.93	$p < .01$	$p < .01$	*ns*	0.0	6.5
	M2: Reading diary app	1.59	0.64				51.6	0.0
	M3: Questionnaire	2.68	0.83				3.2	0.0
Nonfiction ($n = 36$)	M1: Questionnaire	2.21	0.90	$p < .01$	$p < .01$	*ns*	19.4	3.2
	M2: Reading diary app	1.13	0.27				96.8	0.0
	M3: Questionnaire	2.07	0.53				19.4	0.0
Newspapers and Magazines ($n = 13$)	M1: Questionnaire	1.60	0.46	$p < .01$	$p < .01$	*ns*	19.4	0.0
	M2: Reading diary app	1.05	0.11				100.0	0.0
	M3: Questionnaire	1.48	0.38				29.0	0.0
Global/All text types ($n = 150$)	M1: Questionnaire	3.29	0.97	$p < .01$	$p < .01$	$p < .05$	0.0	12.9
	M2: Reading diary app	2.06	0.70				22.6	0.0
	M3: Questionnaire	2.94	0.85				0.0	3.2

Note. $N = 31$ participants. n = entries/events. Reading time measured with the reading diary app was transformed into the same response categories used in the questionnaire. Analyses regarding differences between M1, M2, and M3 were computed using post hoc comparisons after computing repeated-measures ANOVAs. 5-point Likert scale: 1 = *I never read* to 5 = *I read for more than 2 hr.*

Table 5 Pearson Correlations for Reading Behavior Measures in General and by Type of Text

	Fiction			Nonfiction			Newspapers and magazines			Global (all text types)		
	1)	2)	3)	1)	2)	3)	1)	2)	3)	1)	2)	3)
1) M1: Questionnaire	-			-			-			-		
2) M2: Reading diary app	.43*			.45*			.18			.39*		
3) M3:Questionnaire	.66**	.58**		.64**	.48**		.64**	.23		.67**	.58**	
4) M1: CRH scale	.35	.26	.54**	.20	.32	.33	-.24	-.28	-.20	.38*	.38*	.31
5) M1: CRH (1 item)	.48**	.51**	.68**	.03	.22	.27	-.25	-.17	.08	.49**	.51**	.43*

Note. $N = 31$ participants. Cronbach's alpha of CRH was $\alpha = .49$. Correlation M1: CRH Scale and M1: CRH (1st item: "Compared with other college students, how much time do you spend reading all types of materials?"): $r = .65^{**}$. $^{*}p < .05$. $^{**}p < .01$.

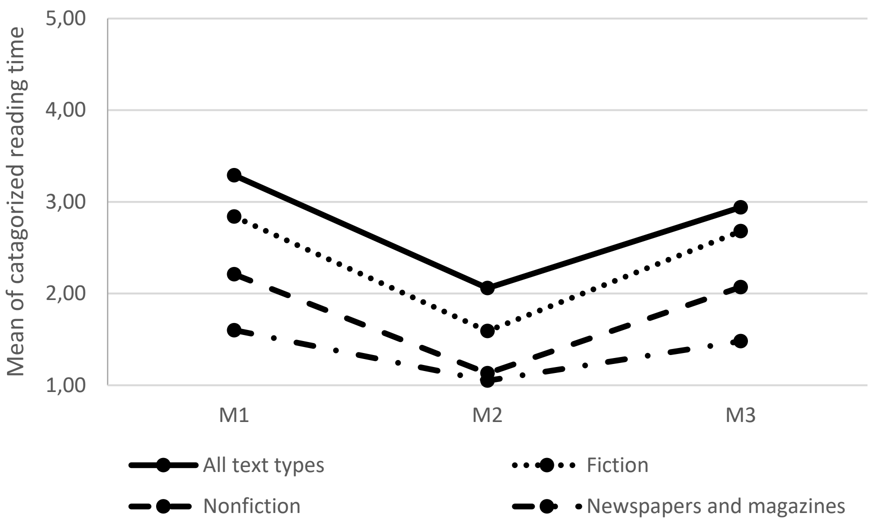


Figure 3 Presentation of differences in reading time between the three measurement points: M1 = prequestionnaire, M2 = reading diary app, M3 = postquestionnaire

Discussion

The main interest of this study was to gain deeper insight into people’s “true” reading behavior using an ambulatory assessment. Therefore, a reading diary app was developed to monitor people’s reading behavior in daily life. In this field report, we explored whether the ambulatory assessment had satisfactory measurement and data quality, which is an important precondition for further analyses. In addition, it was explored how the ambulatory assessment data were related to data from global retrospective questionnaire measures.

First, a reading diary app/ambulatory assessment seems to be an appropriate method for collecting data of satisfactory quality about reading behavior in daily life. As smartphone use allows for a relatively easy and economical data collection process, it seems feasible for participants to document their reading behavior continuously rather than, for instance, just once a day (e.g., as often the case for end-of-day diaries). Therefore, quite precise and detailed information about individuals’ daily reading behavior can be obtained. The results showed that internal consistency was very good, and the correlations of time spent reading in Weeks 1 and 2 as well as between even- and odd-numbered days, another reliability indicator, were strong and in the expected direction (e.g., Anderson et al., 1988; McLaws et al., 1990; Nieuwenboom, 2008). Moreover, most participants reported that documenting their reading behavior with the reading diary app was not a burden for them and that they made regular entries. However, the generalizability of this finding requires further research. For example, some persons (e.g., older people) may be less familiar with smartphone apps than university students, making the use of electronic reading diaries more difficult for this population (Conner & Lehman, 2012). Furthermore, some people may have privacy concerns related to an app that collects information on their personal behavior, including their reading behavior. It is also important to consider that participants only monitored their reading behavior for 2 weeks. Therefore, it might be argued that participants' reading behavior during those 2 weeks was not representative of their reading behavior in general. Nevertheless, two weeks are a common and sometimes recommended time period for diary studies (Conner & Lehman, 2012). Moreover, we chose a time period for our diary study in the middle of the semester when no exams had to be taken, as exam stress could affect college students' reading behavior. Furthermore, in the postmonitoring questionnaire, nearly all participants indicated that the amount of time they spent reading did not deviate significantly from their general reading behavior.

Second, the results showed that the reading diary app data and global retrospective questionnaire data (collected before and after the ambulatory assessment period) were closely related. However, despite the significant correlations, there were substantial differences between the reading time data collected via the reading diary app and the questionnaires. The average daily time spent reading was significantly lower in the ambulatory assessment compared to the questionnaire self-report scales. One possible explanation for this is that participants tend to overestimate the amount of time they spend reading each day when asked to make a global retrospective self-report on a questionnaire. This is in line with Nieuwenboom (2008), who found in a sample of third to fifth graders that students tended to overestimate their reading time in a questionnaire compared to a paper-pencil reading diary. Nevertheless, it must be noted that both the questionnaire and reading diary measures rely on self-reports. A third, independent source of data might be helpful in order to confirm whether participants really overestimated their reading time in the retrospective questionnaire or whether reading diaries tend to underestimate reading time. Although we implicitly assume that participants continuously documented their reading activities during the ambulatory assessment, which should result in less bias due to memory effects, some participants might have forgotten to document their reading time and then did not respond honestly to the conscientiousness question.

Finally, our results found significant differences in the global retrospective self-report before and after the 2 weeks of ambulatory assessment. Furthermore, the correlation between the reading diary app data and the global evaluation of reading time from the questionnaire was stronger after the ambulatory assessment.

One possible explanation is that after participants monitored their own reading behavior for 2 weeks, they revised their reading time estimates, leading to different and possibly more precise responses to the global retrospective question. This change in participants' responses could again be interpreted as a sign that global retrospective questionnaires are influenced by aspects such as heuristics and memory effects. Nevertheless, it should be kept in mind that all results are based on a rather small sample. Therefore, in order to confirm the results regarding the quality of the reading diary app as well as the relations with retrospective questions, the app would need to be applied in a larger sample and complemented with further sources of information, such as interview data. Consequently, future research should try to replicate the study with a larger and more heterogeneous sample with persons at different stages of life, from school students to older adults. Furthermore, data collection periods of varying length might be explored. Whereas longer time periods would help the diary data better capture habitual behavior, longer time periods might also lead to more missing data, measurement error, unwillingness to participate in the study and boredom effects (Bolger et al., 2003). Therefore, the effects of shorter data collection periods might also be examined.

Limitations of the Study

The present study also has some limitations. First, a small convenience sample of college students in psychology and educational science was used. Due to sample selectivity, the results may not generalize to the general population. Second, there were some unexpected technical problems with the reading diary app. The biggest issue was that the titles of 33 reading projects were not transferred to the server correctly. This information about the title/type of text was necessary for the differential analyses by type of text. Because it was not possible to restore this missing information, reading projects lacking information about the title/type of text had to be excluded from these analyses. This led to a reduction in the sample size in these categories and to a relatively small sample size in the newspapers and magazines category, which could be an explanation for the nonsignificant correlation between the app and questionnaire data. Third, it was not possible to determine whether participants made fake entries in the diaries for reasons such as social desirability (Carels et al., 2006; Gershuny, 2012). For example, social desirability effects have been found when parents report reading times with their children, with parents often exaggerating this reading time (Hofferth, 2006). Accordingly, participants might have indicated spending more time reading in the app than they actually spent reading. However, it might be seen as less likely for a participant to continuously make invalid statements for several entries across a two-week period compared to a single questionnaire response. Nevertheless, to address this limitation, it would be useful for future research to examine a third source of information, such as interview data. Finally, there may be individual differences in data accuracy due an imprecise definition of the construct “reading event”. Therefore, some study participants might have recorded reading events in the diary that other participants did not record. Hence, future studies should develop a more precise working definition of the term reading event and communicate this to study participants in order to improve data accuracy.

Conclusion

The present study is among the first to use an ambulatory assessment in the form of a reading diary smartphone app to examine people’s reading behavior. In doing so, this study addresses the often-discussed necessity to use more innovative methods to study behavior in daily life (e.g., Fahrenberg et al., 2007b), and the need to obtain new and deeper insights into people’s common reading activities and qualitative aspects of reading behavior (e.g., Troyer et al., 2018). Global retrospective measures often do not provide information that would allow for such insights because they only reflect average trends and tendencies rather than concrete information about the books and texts a person has actually read.

The present field report illustrates that a reading diary app is a promising method for economically collecting detailed data about people’s reading behavior in daily life. However, ambulant assessment via a smartphone app also involves many challenges (e.g. susceptibility to technical problems, relatively large effort required to develop the app), as shown in this study and documented in this field report. While this study has taken a small first step in the direction of resolving these challenges, and there is a lot of work still to do and improvements to be made. The present study clearly illustrates that reading behavior is a much more complex construct than just the average time spent reading as measured in global retrospective questionnaires. The results showed that the amount of time individuals spent reading each day varied substantially across days. Furthermore, there is great variation in participants’ reading material, which typically remains invisible in global retrospective data – except with respect to very general types of texts. But perhaps it is not just reading a lot but reading diverse books and texts (varying in content, complexity, and writing styles) that makes a competent reader (Kirsch et al., 2002). Given that existing evidence on the relation between reading behavior and reading skills or reading motivation is predominantly based on studies using global retrospective questionnaire data (e.g., Locher et al., 2020; Pfost et al. 2013; Troyer et al., 2018), future research should examine whether these findings can be replicated with more fine-grained measures of reading behavior. It might also be fruitful to further develop the reading diary app to promote increased reading behavior, e.g., by using a token system for the amount of reading time reached (see Robinson, Newby, & Ganzell, 1981). Akbar et al. (2015), for example, found that reading apps can help to improve reading speed. Such interventions might be a further perspective for future research with reading diary apps.

References

Acheson, D. J., Wells, J. B., & MacDonald, M. C. (2008). New and updated tests of print exposure and reading abilities in college students. *Behavior Research Methods, 40*(1), 278-289. https://doi.org/10.3758/BRM.40.1.278

Ahram, T. (Ed.). (2019). *Advances in artificial intelligence, software and systems engineering.* Cham: Springer International Publishing.

Akbar, R. S., Taqi, H. A., Dashti, A. A., & Sadeq, T. M. (2015). Does e-reading enhance reading fluency? *English Language Teaching, 8*(5), 195-207. https://doi.org/10.5539/elt.v8n5p195

Alexander, P. A. (2005). The Path to Competence: A Lifespan Developmental Perspective on Reading. *Journal of Literacy Research, 37*(4), 413–436. https://doi.org/10.1207/s15548430jlr3704_1

Allen, L., Cipielewski, J., & Stanovich, K. E. (1992). Multiple indicators of children’s reading habits and attitudes: construct validity and cognitive correlates. *Journal of Educational Psychology, 84*(4), 489-503.

Anderson, R. C., Wilson, P. T., & Fielding, L. G. (1988). Growth in reading and how children spend their time outside of school. *Reading Research Quarterly, 23*(3), 285-303. Retrieved from http://www.jstor.org/stable/748043

Artelt, C., Stanat, P., Schneider, W. & Schiefele, U. (2001). Lesekompetenz: Testkonzeption und Ergebnisse. In J. Baumert, E. Klieme, M. Neubrand, M. Prenzel, U. Schiefele, W. Schneider et al. (eds.), *PISA 2000* (Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich, pp. 69-137). Opladen: Leske + Budrich.

Becker-Mrotzek, M., Brinkhaus, M., Grabowski, J., Hennecke, V., Jost, J., Knopp, M., Schmitt, M., Weinzierl, C. & Wilmsmeier, S. (2015). Kohärenzherstellung und Perspektivübernahme als Teilkomponenten der Schreibkompetenz. Von der diagnostischen Absicherung zur didaktischen Implementierung. In A. Redder, J. Naumann & R. Tracy (eds.), *Forschungsinitiative Sprachdiagnostik und Sprachförderung – Ergebnisse*. (pp. 177–205). Münster: Waxmann.

Bolger, N., Davis, A., & Rafaeli, E. (2003). Diary methods: Capturing life as it is lived. *Annual Review of Psychology, 54*(1), 579-616. https://doi.org/10.1146/annurev.psych.54.101601.145030

Burgess, S. R., Hecht, S. A., & Lonigan, C. J. (2002). Relations of the home literacy environment (HLE) to the development of reading related abilities: A one year longitudinal study. *Reading Research Quarterly, 37*(4), 408-426. https://doi.org/10.1598/RRQ.37.4.4

Bus, A. G., van IJzendoorn, M., & Pellegrini, A. (1995). Joint book reading makes for success in learning to read: A meta-analysis on intergenerational transmission of literacy. *Review of Educational Research, 65*(1), 1-21. https://doi.org/10.3102/00346543065001001

Calamia, M. (2019). Practical considerations for evaluating reliability in ambulatory assessment studies. *Psychological assessment, 31*(3), 285. https://doi.org/10.1037/pas0000599

Carels, R. A., Cacciapaglia, H. M., Rydin, S., Douglass, O. M., & Harper, J. (2006). Can social desirability interfere with success in a behavioral weight loss program? *Psychology & Health, 21*(1), 65-78. https://doi.org/10.1080/14768320500102277

Chall, J. S. (1983). *Stages of reading development.* New York, NY: McGraw-Hill.

Conner, T. S., & Lehman, B. J. (2012). Getting started: Launching a study in daily life *Handbook of research methods for studying daily life.* (pp. 89-107). New York, NY, US: The Guilford Press.

Csikszentmihalyi, M., & Larson, R. (2014). Validity and Reliability of the Experience-Sampling Method. In *Flow and the Foundations of Positive Psychology* (pp. 35-54). Springer, Dordrecht. https://doi.org/10.1007/978-94-017-9088-8_3

Ebner-Priemer, U. W., Kuo, J., Kleindienst, N., Welch, S. S., Reisch, T., Reinhard, I. et al. (2007). State affective instability in borderline personality disorder assessed by ambulatory monitoring. *Psychological Medicine*, 37(7), 961–970. https://doi.org/10.1017/S0033291706009706

Fahrenberg, J., Myrtek, M., Pawlik, K., & Perrez, M. (2007a). Ambulantes Assessment - Verhalten im Alltagskontext erfassen [Ambulatory assessment - recording behaviour in an everyday context]. *Psychologische Rundschau, 58*(1), 12-23. https://doi.org/10.1026/0033-3042.58.1.12

Fahrenberg, J., Myrtek, M., Pawlik, K., & Perrez, M. (2007b). Ambulatory assessment-Monitoring behavior in daily life settings: A behavioral-scientific challenge for psychology. *European Journal of Psychological Assessment, 23*(4), 206. https://doi.org/10.1027//1015-5759.23.4.206

Foasberg, N. M. (2014). Student reading practices in print and electronic media. *College & Research Libraries*, *75*(5), 705–723. https://doi.org/10.5860/crl.75.5.705

Gershuny, J. (2012). Too many zeros: A method for estimating long-term time-use from short diaries. *Annals of Economics and Statistics* (105/106), 247-270. https://doi.org/10.2307/23646464

Glomann, L., Hager, V., Lukas, C. A. & Berking, M. (2019). Patient-centered design of an e-mental health app. In T. Ahram (Ed.), *Advances in artificial intelligence, software and systems engineering* (pp. 264–271). Cham: Springer International Publishing.

Greaney, V., & Hegarty, M. (1987). Correlates of leisure-time reading. *Journal of Research in Reading, 10*(1), 3-20. https://doi.org/10.1111/j.1467-9817.1987.tb00278.x

Guthrie, J. T., Wigfield, A., Metsala, J. L., & Cox, K. E. (1999). Motivational and cognitive predictors of text comprehension and reading amount. *Scientific Studies of Reading, 3*(3), 231-256. https://doi.org/10.1207/s1532799xssr0303_3

Hektner, J. M., Schmidt, J. A., & Csikszentmihalyi, M. (2007). *Experience Sampling Method.* Thousand Oaks, London, New Delhi: Sage Publications

Hertel, S., Hochweber, J., Mildner, D., Steinert, B., & Jude, N. (2014). *PISA 2009 Skalenhandbuch [PISA 2009 scaling handbook].* Münster; New York: Waxmann.

Hofferth, S. L. (2006). Response Bias in a Popular Indicator of Reading to Children. *Sociological Methodology, 36*(1), 301–315. https://doi.org/10.1111/j.1467-9531.2006.00182.x

IBM-Corporation. (2012). IBM SPSS Bootstrappin 21. Retrieved from ftp://public.dhe.ibm.com/software/analytics/spss/documentation/statistics/21.0/en/client/Manuals/IBM_SPSS_Bootstrapping.pdf

Jerrim, J. & Moss, G. (2019), The link between fiction and teenagers' reading skills: International evidence from the OECD PISA study. *British Educational Research Journal*, 45: 181-200. https://doi.org/10.1002/berj.3498

Juster, F. T., Ono, H., & Stafford, F. P. (2003). An Assessment of Alternative Measures of Time Use. *Sociological Methodology, 33*, 19–54.

Kahneman, D., Krueger, A. B., Schkade, D. A., Schwarz, N., & Stone, A. A. (2004). A Survey Method for Characterizing Daily Life Experience: The Day Reconstruction Method. *Science, 306*(5702), 1776-1780.

Kan, M. Y., & Pudney, S. (2008). 2. Measurement Error in Stylized and Diary Data on Time Use. *Sociological Methodology, 38*(1), 101–132. https://doi.org/10.1111/j.1467-9531.2008.00197.x.

Keller, A. (2010). *Einsatz von digitalen Foto-Lesetagebüchern zur Erforschung des Leseverhaltens von Studierenden. [Use of digital photo reading diaries to research the reading behaviour of students].* Paper presented at the 5. Konferenz der Zentralbibliothek, Forschungszentrum Jülich.

Kirsch, I., de Jong, J., Lafontaine, D., McQueen, J., Mendelovits, J., & Monseur, C. (2002). *Reading for change. Performance and engagement across countries. Results from PISA 2000.* Paris: OECD.

Kutner, M., Greenberg, E., Jin, Y., Boyle, B., Hsu, Y., & Dunleavy, E. (2007). *Literacy in everyday life: Results from the 2003 National Assessment of Adult Literacy (NCES 2007–480).* Washington, DC: National Center for Education Research.

Lampert, T., Sygusch, R., & Schlack, R. (2007). Nutzung elektronischer Medien im Jugendalter [Use of electronic media in youth]. *Bundesgesundheitsblatt - Gesundheitsforschung - Gesundheitsschutz, 50*(5), 643-652. https://doi.org/10.1007/s00103-007-0225-7

Leiner, D. J. (2019). SoSci Survey (Version 3.1.06) [Computer software]. Available at https://www.soscisurvey.de

Lucas, R. E., Wallsworth, C., Anusic, I., & Donnellan, B. (2019). *A Direct Comparison of the Day Reconstruction Method and the Experience Sampling Method.* https://doi.org/10.31234/osf.io/cv73u

Locher, F. M., Becker, S., & Pfost, M. (2019a). The Relation Between Students' Intrinsic Reading Motivation and Book Reading in Recreational and School Contexts. *AERA Open*, 5(2), 1-14. https://doi.org/10.1177/2332858419852041

Locher, F. M., & Pfost, M. (2019b). Erfassung des Lesevolumens in Large-Scale Studien. Ein Vergleich von Globalurteil und textspezifischem Urteil. [Measuring reading volume in Large-Scale Assessments: A comparison of an overall evaluation and a differentiated evaluation relating different text types.]. *Diagnostica(1)*, 26-36. https://doi.org/10.1026/0012-1924/a000203

Locher, F., & Pfost, M. (2020) The relation between time spent reading and reading comprehension throughout the life course. *Journal of Research in Reading*, 43: 57– 77. https://doi.org/10.1111/1467-9817.12289.

Marshall, J. (2000). Research response to literature. In M. L. Kamil, P. B. Mosenthal, P. D. Pearson, & R. Barr (Eds.), *Handbook of reading research: Vol. III* (pp. 381-402). Mahwah, NJ: Lawrence Erlbaum Associates.

McGeown, S. P., Duncan, L. G., Griffiths, Y. M., & Stothard, S. E. (2015). Exploring the relationship between adolescent's reading skills, reading motivation and reading habits. *Reading and Writing, 28*(4), 545-569. https://doi.org/10.1007/s11145-014-9537-9

McGeown, S. P., Osborne, C., Warhurst, A., Norgate, R., & Duncan, L. G. (2016). Understanding children's reading activities: Reading motivation, skill and child characteristics as predictors. *Journal of Research in Reading, 39*(1), 109-125. https://doi.org/10.1111/1467-9817.12060

McLaws, M. L., Oldenburg, B., Ross, M. W., & Cooper, D. A. (1990). Sexual behaviour in aids-related research: Reliability and validity of recall and diary measures. *The Journal of Sex Research, 27*(2), 265-281. https://doi.org/10.1080/00224499009551556

Miller, G. (2012). The smartphone psychology manifesto. *Perspectives on Psychological Science, 7*(3), 221-237. https://doi.org/10.1177/1745691612441215

Mol, S. E., & Bus, A. G. (2011). To read or not to read: A meta-analysis of print exposure from infancy to early adulthood. *Psychological Bulletin, 137*(2), 267-296. https://doi.org/10.1037/a0021890

Moosbrugger, H. & Kelava, A. (Hrsg.). (2012). *Testtheorie und Fragebogenkonstruktion* (2.Aufl.). Berlin, Heidelberg: Springer.

Mullis, I. V. S. & Martin, M. O. (2019). PIRLS 2021 reading assessment framework. In I. V. S. Mullis & M. O. Martin (eds.): *PIRLS 2021 Assessment Frameworks.* (pp. 5-25). Retrieved from Boston College, TIMSS & PIRLS International Study Center website: https://timssandpirls.bc.edu/pirls2021/frameworks/

Nieuwenboom, J. W. (2008). *Wie viel lesen Kinder? Die Erfassung von Leseaktivitäten mit Hilfe von strukturierten Tagebüchern-eine methodologische Studie [How much do kids read? The recording of reading activities using structured diaries - a methodological study]*: Tectum Verlag.

OECD (2010). *PISA 2009 results: Learning to learn – Student engagement, strategies and practices. Volume III.* https://doi.org/10.1787/9789264083943-en

OECD (2019). *PISA 2018 Assessment and Analytical Framework*, PISA, OECD Publishing, Paris. https://doi.org/10.1787/b25efab8-en.

OECD (2021), 21st-Century Readers: Developing Literacy Skills in a Digital World, PISA, OECD Publishing, Paris. https://doi.org/10.1787/a83d84cb-en

Pfost, M., Dörfler, T., & Artelt, C. (2013). Students' extracurricular reading behavior and the development of vocabulary and reading comprehension. *Learning and Individual Differences*, 26, 89-102. https://doi.org/10.1016/j.lindif.2013.04.008

Pospeschill, M. (2010). *Testtheorie, Testkonstruktion, Testevaluation*. Munchen: Reinhardt.

Raith, T. (2008). Weblogs als Lesetagebücher im aufgabenorientierten Fremdsprachenunterricht–Ergebnisse einer Vergleichsstudie [Weblogs as reading diaries in task-oriented foreign language teaching - results of a comparative study]. *Aufgabenorientiertes Lernen und Lehren mit Medien: Ansätze, Erfahrungen, Perspektiven in der Fremdsprachendidaktik, 15*, 297.

Robinson, P. W., Newby, T. J., & Ganzell, S. L. (1981). A token system for a class of underachieving hyperactive children. *Journal of Applied Behavior Analysis, 14*(3), 307-315. https://doi.org/10.1901/jaba.1981.14-307

Shumow, L., Schmidt, J. A., & Kackar, H. (2008). Reading In Class & Out of Class: An Experience Sampling Method Study. *Middle Grades Research Journal, 3*(3), 97-120.

Stoffelsma, L. (2018). Short-term gains, long-term losses? A diary study on literacy practices in Ghana. *Journal of Research in Reading, 41*(S1), S66-S84. https://doi.org/10.1111/1467-9817.12136

Stone, A., Broderick, J., Schwartz, J., Shiffman, S., Litcher-Kelly, L., & Calvanese, P. (2003). Intensive momentary reporting of pain with an electronic diary: reactivity, compliance, and patient satisfaction. *Pain, 104*(1), 343-351. https://doi.org/10.1016/S0304-3959(03)00040-X

Troyer, M., Kim, J., Hale, E., Wantchekon, K., & Armstrong, C. (2018). Relations among intrinsic and extrinsic reading motivation, reading amount, and comprehension: a conceptual replication. *Reading and Writing*. https://doi.org/10.1007/s11145-018-9907-9

Trull, T. J., & Ebner-Priemer, U. (2014). The role of ambulatory assessment in psychological science. *Current Directions in Psychological Science, 23*(6), 466-470. https://doi.org/10.1177/0963721414550706

Wilhelm, P., Perrez, M., & Pawlik, K. (2012). Conducting research in daily life: A historical review *Handbook of research methods for studying daily life*. (pp. 62-86). New York, NY, US: The Guilford Press.

Zhang, J., Calabrese, C., Ding, J., Liu, M., & Zhang, B. (2018). Advantages and challenges in using mobile apps for field experiments: A systematic review and a case study. *Mobile Media & Communication, 6*(2), 179–196. https://doi.org/10.1177/2050157917725550

Zirkel, S., Garcia, J. A., & Murphy, M. C. (2015). Experience-Sampling Research Methods and Their Potential for Education Research. *Educational Researcher, 44*(1), 7-16. https://doi.org/10.3102/0013189X14566879

methods, data, analyses | Vol. 17(2), 2023, pp. 303-324 DOI: 10.12758/mda.2023.03

Large-Scale Comparative School-Based Survey Research: Challenges and Solutions for Sampling, Fieldwork and Informed Consent

Natalia Waechter[1,2], *Veronika Kalmus*[3], *Giovanna Mascheroni*[4], *& Signe Opermann*[3]

[1] *University of Graz*

[2] *Ludwig-Maximilian University Munich*

[3] *University of Tartu*

[4] *Università Cattolica del Sacro Cuore*

Abstract

Based on our experiences with implementing the comparative school-based ySKILLS survey in six European countries, this article investigates the preparation of fieldwork in school-based surveys. This includes the sampling strategies and recruitment of schools and (secondary level) students, the continuous collaboration with schools, as well as collecting parental consent. By interviewing the national survey experts, we found that previously described challenges of school-based survey research have become specifically relevant during the COVID-19 pandemic. Our results further show that collaborating with schools is demanding and that collecting active parental consent involves problems regarding a non-response bias as well as ethical concerns about children's rights. For future research, we have identified seven general preconditions and facilitating factors regarding the recruitment and collaboration with schools for a successful implementation of school-based surveys. Regarding informed consent, we provide seven ethical and practical recommendations for research policy and future studies.

Keywords: School-based survey, fieldwork, informed consent, sampling, response rate, COVID-19

The European research project Youth Skills (ySKILLS) investigates digital skills of young people in a longitudinal perspective, surveying students aged 12 to 15 in three consecutive years (until they become 14 to 17, respectively). In order to understand which factors influence young people's acquisition of digital skills and how, in turn, digital skills influence young people's wellbeing, we have developed a quantitative, longitudinal research design and a questionnaire with new digital skills indicators. We decided on a non-probability sample with data collection in schools because previous research has led to expect much higher response rates in longitudinal school-based surveys (e.g., Schreiner & Haider, 2006) compared to out-of-school surveys with children and young people (e.g., Brix et al., 2017).

In the ySKILLS project, survey data is being collected in secondary schools in six European countries (Estonia, Finland, Germany, Italy, Poland and Portugal) in three waves (2021, 2022, 2023). For each wave we aim at a sample with at least N=6,000 students in secondary education (n=1,000 per country). This article focuses on our experiences and insights related to the first wave, which was successfully accomplished in 2021 with N=6,221 students aged 12 to 15.

While our overall response rate of 60.8% is higher than average response rates in out-of-school surveys,[1] it is still smaller than we had expected based on our previous experiences with (voluntary) school-surveys. In this article, we investigate possible reasons for the non-response and reflect on the particular challenges of school-based surveys and data collection with children and young people, as well as on the challenges related to the fieldwork during the COVID-19 pandemic. In fact, data collection in schools took place in spring and autumn 2021, when different restrictions to contain the health emergency were adopted by the survey countries. Consequently, in some countries the online-survey was not only administered in class but also fully online with students at home (as it was the case in Estonia, Germany, and Italy, when a class was quarantined or schools were closed[2]), or in a hybrid mode (with some students in class and some at home, as in Estonia and Italy). In Portugal, Poland and Finland, the survey was administered mainly face to face in class, except for certain classes in quarantine in Poland.

The diverse restrictions in place to contain the pandemic did alter not only data collection but also the recruitment and collaboration with schools. Previous research has already pointed out the complexity of school-based large-scale survey research (e.g., Bartlett et al., 2017; Madge et al., 2012). For a successful implemen-

[1] A meta-analysis of studies published in academic journals revealed an average response rate of 52.7% (Baruch & Holtom, 2008).

[2] In Italy, one of the participating schools was closed due to a power cut.

Direct correspondence to
Natalia Waechter, Ludwig-Maximilian-University Munich and University of Graz,
Department for Educational Science (ORCID: 0000-0001-6828-6517)
E-mail: natalia.waechter@uni-graz.at

tation of the school-based survey, we found fieldwork preparation regarding the recruitment of schools, the collaboration with participating schools and parental consent to be particularly important. In addition to the challenges in recruiting and working with schools and collecting parental consent as described in the literature so far, we also faced new challenges related to the COVID-19 pandemic.

In this article we will reflect on our experience and new insights based on qualitative expert interviews with all project leaders in the respective countries and team experts of the ySKILLS data collection consortium. The interviews concern the recruitment of schools, the collaboration with participating schools as well as the methodological and ethical problems with parental consent.

The following section summarizes previous findings regarding fieldwork preparation of school-based surveys. "The ySKILLS survey" section provides basic information on the ySKILLS survey and fieldwork, while in "The Expert Interview" section the methodological outline of our expert interviews is described. In the following sections we present and discuss the results of our experiences, reflections and insights regarding fieldwork preparation of large-scale school surveys. Finally, in the "Conclusions and Recommendations" we have developed recommendations for future school-based survey research.

Previous Findings and Scholarly Debate on Fieldwork Preparation: Recruitment of Schools and Parental Consent

Recruitment of Schools and Collaboration with Participating Schools

With some notable exceptions (Madge et al., 2012; Mishna et al., 2012; Rice et al., 2007), only few publications provide critical commentary on the process of doing school-based large-scale surveys in its complexity, including all the background work that usually remains invisible. In fact, the challenges of the "non-empirical work" and emotion work (Lindsay, 2005) involved in recruiting schools, getting them on board and negotiating access is often glossed over in articles reporting on school-based survey findings.

Gaining access to schools in order to undertake research with children is often a lengthy and sensitive process. Schools are busy institutions, increasingly overwhelmed with both requests for participating in academic research and growing administrative tasks (Madge et al., 2012). Getting schools to take part in research involves identifying the best contacts in the school, establishing collaborative relations with all the parties involved, and negotiating participation.

A major challenge in conducting school-based surveys lies in ensuring the cooperation of a variety of gatekeepers, including local authorities, school principals, teachers and parents (Barker & Weller, 2003; Bartlett et al., 2017). Gaining access to the key contacts in school is demanding: as Madge et al. (2012, p. 422) explain, "Making and maintaining contact involved school visits, telephone calls and emails, many of which did not elicit any response".

Once they receive a response from school, researchers have to cultivate collaborative relationships with each level of authority, and persuade them that participating in the research is not in conflict with the school's educational mission (Madge et al., 2012; Mishna et al., 2012; Rice et al., 2007). Negotiating participation of all interested parties is vital: if school principals are not motivated in taking part in the research, getting the support of teachers and solving any logistics issue may be complicated as well. Similarly, if the school principal is on board but key teachers are not committed to supporting the recruitment process, getting the consent of parents and children will also be difficult. For these reasons, Rice et al. (2007) suggest to motivate each group of gatekeepers within the school community separately.

Managing the logistical aspects of a school-based survey also requires a deep understanding of the school organization (daily timetable, the school's calendar of events and the national education calendar, including the PISA test administration) and access to computers.

Ultimately, recruitment and collaboration with schools is best achieved if research in schools is conceived of as a "give and take" process (Madge et al., 2012, p. 423). Researchers need to emphasize how the school community can benefit from participating in the survey and follow up with the school on a regular basis. For example, promising feedback on initial findings, offering teachers' training or education initiatives aimed at children and parents, and recognizing that each school has distinctive needs are all suggested ways to maintain a positive relationship with schools, and, as a consequence, gain their commitment over time (Clary et al., 2021; Madge et al, 2012; Mishna et al., 2012; Rice et al., 2007).

Parental Consent

Although a school-based survey is a viable method to reach adolescents' populations and to obtain important data on various aspects of their lives and surroundings, researching underage adolescents cannot be done without their parents' or guardians' permission as well as the voluntary and informed consent from the adolescents themselves. Ideally, these two decisions should be in harmony or at least negotiated, but there may appear situations where the researchers face the dilemma of having to choose between the parents' and the child's views.

The primary ethical norms and principles of research integrity (e.g., Ryan et al., 1979; The European Code of Conduct for Research Integrity, 2017) rightly prioritize the need to respect the persons involved in research, and the protection of their wellbeing, autonomy, privacy and the best interests, stating that the research activity must avoid any harm to their health and dignity. The ethical approaches also warn against the risk of vulnerability, marginalization and stigmatization that may damage research participants' interests. However, depending on the topic and design of research, there may be cases of underage children being declined participation by their parents or guardians even if the children themselves would be in favor of taking part in the research. This may result in negative effects on the rights of adolescents as young citizens, as well as on the response rate, sampling bias, and, thus, validity of the research findings.

One of the central topics in scholarly discussion on parental consent (e.g., Baker et al., 2001; Dent et al., 1997; Cavazos-Regh et al., 2020; Courser et al., 2009; Liu et al., 2017) has been the methodological effects of "active" versus "passive" consent procedures. The first type of consent means an explicitly given, in most cases written and manually signed permission (an "opt-in" procedure), while "passive" type of consenting means receiving information about the study and having an opportunity to "opt-out" by returning a "non-consent" form. In the respective academic debate, Courser et al. (2009, p. 2) refer to the changing research environments and increasingly demanding requirements for school-based student surveys in several countries (e.g., the United States). They explicitly regret the shift from passive consent procedures, which have for a long time fulfilled ethical and statutory requirements when participating students remain anonymous and which have usually guaranteed high response rates, to active parental consent for all research. As Courser et al. (ibid.) clarify, under active consent procedure, "an unreturned consent form is equivalent to refusal of consent" and can mean several things including explicit refusal by the parents, neglecting to return the form, the loss of the form in transit back to the school, etc. The authors' (Courser et al., 2009, p. 3) main concerns about this shifting regulatory and research environment for school-based surveys are that it leads to low student participation rates and a non-response bias in survey data.

Courser et al. (2009, p. 4) also emphasize that there is a higher tendency for vulnerable students to be excluded from the surveys if their parents do not take much effort to interact with the school or research team. Systematic comparative analysis confirms that studies accepting only active forms of parental consent lead to silencing the voices of some (vulnerable) groups, for example, boys with lower academic achievement, adolescents belonging to certain ethnic minority groups or with risk-taking behavior (Liu et al., 2017, p. 46). Such systematic error in sampling based only on written parental consent and leaving out many participants may be associated with parents who do not provide active consent because they are not so engaged with the school and lack awareness of the benefits of their child's contribution to research, or who are skeptical of science, but also whose own educational attainment may be lower, or who are facing challenges in their everyday lives.

The authors of meta-analyses (e.g., Liu et al., 2017) and studies (e.g., Cavazos-Regh et al., 2020) point out another problem. They emphasize that active parental consent can act like a potential barrier, keeping some adolescents away from research who would be quite happy to participate. Both cited studies pay special attention to research on students with depression and anxiety, eating disorders and risk behaviors and who may not feel free to talk about these issues with their parents. Cavazos-Regh et al. (2020, p. 4) conclude that the adolescents attempted to retain privacy by not allowing researchers to contact parents about active consent.

Currently, the ethical regulations of empirical research among adolescents consider this issue mainly from a juridical perspective and associate it with parental responsibility and authority over their child in all matters until the children legally become adults. However, the overall regulations, largely influenced by medical research and sciences, do not address the key dilemma of parental authority versus children's agency, which is being faced by the researchers in the field of social studies. For example, Iltis (2013, p. 333) discusses the controversy in research ethics policy and guidelines "regarding who ought to make decisions involving children" in research, and proposes that the traditional approach of parents being "the default decision-makers for children" with regard to various matters including education needs to be revised, so that children could have greater authority over themselves and be treated as rights-bearers, as the United Nations Convention on the Rights of the Child states, especially when children's "best interests" tend to be threatened (ibid.).

The ySKILLS Survey: Sampling Schools, Classes, and Students

The ySKILLS longitudinal survey is based on a quantitative questionnaire administered in schools in six European countries in three waves (2021, 2022, 2023). We aimed at a purposive, non-probability sample (at least n=1000 per wave and country) that would allow for a diverse and inclusive sample of respondents. Our basic population in the first wave were 12- to 15-year-old adolescents attending secondary school (ISCED 2 and ISCED 3). The first wave was successfully accomplished in 2021 – despite the pandemic and restrictive measures which also affected schools we were able to collect data from N=6,221 participants (final sample size after data cleaning) (Bedrosova et al., 2022).

Funding regulations required that instead of using public opinion institutes, the national researchers of the ySKILLS consortium directly recruited schools for participation in the survey and carried out data collection in schools. Furthermore, longitudinal, large-scale data collection in schools is a complex process, which we found can only be accordingly implemented and carried out by the researchers themselves. This applies even more since we also collected network data.

After our research had gotten approved by the IBR committee of the project coordinator's university (KU Leuven) (Application Dossier Social and Societal Ethics Committee, 2020), the project partners responsible for the longitudinal school-based data collection in their countries applied for ethical approval according to national regulations. In Germany and Portugal, the survey had to be approved by the (Federal) Ministry of Education, and in Finland, Italy, and Poland, approval was required by the ethical commission of the project partners' universities (University of Helsinki; Università Cattolica del Sacro Cuore; Adam Mickiewicz University). The Finnish team obtained another ethical approval by the city of Salo. In Estonia, no further ethical approval was necessary as KU Leuven's procedure was considered adequate and applicable.

In each of the six participating countries, the schools at secondary level were recruited in specific regions, usually the city and the surrounding districts of the partner university in the project. Regarding the *sampling of the schools*, we had decided for a non-probability sample because data collection would have required too many resources if carried out in schools across whole countries and there was no evidence leading to expect regional differences.

A systematic evidence review of the antecedents and consequences of digital skills (Haddon et al., 2020) has shown that some studies point to a direct association of families' socio-economic status (SES) with children's digital skills (Paus-Hasebrink et al., 2019; Zilka, 2019). Other research, instead, found an indirect effect of household SES on digital skills, mediated by access (Fizeşan, 2012): children from higher-income families seem to benefit from more autonomy of use and better quality of access. Overall, Haddon et al. (2020) found more studies showing a positive effect of household as well as school SES on digital skills of children than studies showing no significant effect.[3] Therefore, we aimed at collecting a diverse sample regarding SES and applied two sampling strategies. Basically, we selected schools in different school districts characterized by varying degrees of urbanization and wealth (as in Estonia, Finland, Italy, Poland, and Portugal) (Bedrosova et al., 2022). In countries with a segregated school system (Germany and Italy), we also selected different types of schools (professional/vocational education on the one side and

[3] The authors conclude that the mixed results of household SES may derive from different measurements/proxies (e.g., parents' education, income). Regarding school SES, they state that the causality remains unclear: "Do these schools lead to more skills or do the type of children likely to develop such skills go to particular schools?" (Haddon et al., 2020, p. 72).

grammar schools on the other side, because each type is usually attended by students with a similar SES background).[4]

In each school, we sampled the *classes* by grades (in the first wave: classes with students aged 12 to 15 which corresponds with grade 6 or 7 to grade 9 or 10)[5] and availability (depending on the timetables, exams, etc.). In all countries, classes were sampled in four grades (grades 6–9 or grades 7–10) and the grades were equally distributed in each regional sample (e.g., two classes in each of the four grades). In smaller schools, all classes in a specific grade had to be surveyed.

In Germany, Estonia, and Finland, students transition after grade 9 from lower to upper secondary education (from ISCED 2 to ISCED 3). At this point, the majority of students are 15 years old. This means that in the first wave, the surveyed students in all four grades were in lower secondary education in Germany and Estonia (ISCED 2). Only in Finland, the students from grade 6 still belong to ISCED 1.

In Italy, Poland, and Portugal, students were surveyed in ISCED 2 grades as well as ISCED 3 grades in the first wave because the majority is only 14 years old when transitioning from ISCED 2 to ISCED 3.

In all classes, we aimed at a *full sample*, but because it was planned to survey all students per class at once, in the first wave we had expected a non-response rate (due to illness, etc.) of about 10%.[6] The actual (individual) non-response in wave 1, however, turned out to be 39.2% (ranging from 20.1% in Germany to 61.9% in Finland), mainly due to eligible students without active parental consent as well as more students having been absent from class during the pandemic (for response rates of each country see *Table 4*).

Considering the non-probability sample, we assessed possible limitations by considering population statistics and estimates. Regarding gender (50.2% male, 48.1% female, 1.7% other), the sample does not significantly differ from the population of 12- to 15-year-old adolescents in the surveyed countries. Regarding age,

[4] The aim of this sampling strategy was not to define the SES for the individual students but to receive a diverse sample. For estimating the SES of the individual surveyed students, we used a child-friendly variable on the financial situation of the family ("the people with whom you live"). We asked them to choose from five items, from (1) "We live very well – We can purchase luxury items and still have money left over" to (5) "We struggle to get by – We sometimes do not have enough money to afford basic needs, such as food and clothes".

[5] The European schooling systems vary somewhat in the age of school entry; therefore, a particular grade does not correspond with the exact same age group across all countries.

[6] The consortium of the Programme of International Student Assessment (PISA) sets response thresholds for data quality. The threshold for pupil response is 80% (Micklewright et al., 2010). Authors of the German PISA study, for example, reported a non-response of only 6.4% (Schreiner & Haider, 2006). While participation in our survey was not mandatory, the school-based data collection during regular class with teachers being present led to expecting high participation rates, not much lower than those of PISA.

there are significantly less 12-year-olds and 15-year-olds than 13- and 14-year-olds in the sample, which is due to data collection per grade. In one grade, there are always two cohorts, so the lowest grade surveyed consists of students aged 12 and 13, while the next grade consists of students aged 13 and 14, etc. This means there were fewer chances for 12-year-olds to become part of the sample than for 13-year-olds who were presented in two grades. Furthermore, as described above, the sample does not represent the population regarding regional diversity within countries. Also the country- and region-specific school systems, their embeddedness in the respective political systems and the different political systems themselves as well as different social contexts represent limitations, above all, for comparing the regional samples in different countries. Regarding the language spoken at home, the distribution in our sample seems to correspond with the official national (or, if available, regional) population statistics. For example, in the German first wave sample (2021) collected in Bavarian schools, 17.3% reported a language other than German, while German microcensus data from 2017 reveals that 15.0% of the 12- to 17-year-old Germans and 17.6% of children and young people in Bavaria aged 17 and younger live in foreign language households (Geis-Thöne, 2021). Finally, in all countries where our survey was implemented, there is compulsory education until the age of 16 (in Poland: until 15). This means that there are no limitations regarding educational participation because in the surveyed age group (12 to 15) all boys and girls are obliged to be in school. In this paper, we will further address in which way the requirement of active parental consent represents a limitation.

The Expert Interviews: Reflecting on Fieldwork

As anticipated above, the challenges faced in recruiting participating schools, collecting data amidst social distancing measures and the higher non-response rate have left us with many unanswered questions. Therefore, in order to investigate possible reasons for the unexpected non-response and to reflect on the particular challenges of school-based surveys and data collection with children, as well as on challenges related to doing fieldwork during the COVID-19 pandemic, we used the qualitative method of expert interviews (see e.g., Bogner et al., 2014; Doeringer, 2020). Our aim was not only to learn about facts and processes that apply to the specific national and regional contexts but also to gain knowledge about interpretations and recommendations by the project leaders who were responsible for data collection in their countries. These stem from different academic disciplines within the field of social sciences (Sociology, Educational Science, and Media and Communication Studies), were all experienced in international survey research and had taken part in various collaborative research projects before (EU Kids Online, Mediappro, EUYOUPART, ENRI-East, CATCH-EyoU). Their previous experience in survey research allowed them to compare and to detect changes and particularities.

The authors of this article have developed a qualitative interview guide in written form (see Appendix) regarding the preparation and the implementation of fieldwork, containing mainly open questions as well as interview topics for elaborating on them. It was sent to all team leaders in the respective countries after the data collection had been completed in all countries (in November 2021). The team leaders and experts of the six countries answered extensively in written form (on average, more than 10 pages) and their answers were coded based on the topics (e.g., recruitment of schools) and subtopics of the questionnaire (e.g., recruitment strategies, changes of strategies, cancellation of schools, personal contacts, sampling strategy regarding SES, COVID-related problems, etc.). In accordance with the principles of qualitative research and problem-centered interviews (Doeringer, 2020), we were also open to new subtopics when coding the material, above all, regarding practical and political implications. During the process of qualitative content analysis (Kuckartz, 2014), we also used the possibility to contact the experts again for clarifying questions (both orally and in writing). Additionally, we used the national technical reports that had been written in the frame of the project for documenting data collection, as summarized in Bedrosova et al. (2022).

Findings

Recruiting Schools and Collaborating with Schools

Getting schools on board was a lengthy and challenging process: beyond the usual challenges of overburdened schools, already highlighted in prior school-based research (Madge et al., 2012; Mishna et al., 2012; Rice et al., 2007), the COVID-19 pandemic, with social distancing measures and schools switching to remote learning during surges in infection, played a role in schools' refusal to take part in the research. The national research teams had to make vigorous efforts to find schools for collaboration:

> "Ensuring schools' participation was highly demanding: especially for upper secondary schools, this involved several email exchanges, plus several phone calls and online meetings between the researcher and the reference teacher (up to 5 meetings lasting 1 to 2 hours for each school)." (Italian expert)

The German experts pointed out that they had prepared individual presentations for each school. In the meetings, typically, two researchers, the school principal, and the contact teacher had been taking part.

As shown in *Table 1*, in most countries, researchers contacted a higher number of schools than effectively participated in the studies. Non-response from schools

Table 1 Recruitment and participation of schools (ySKILLS survey 2021)

Country	Contacted schools	Recruited schools	Cancellation during data collection	Prolonged data collection	School response rate
Estonia	14	9	3	no	64%
Finland	11	11	0	no	100%
Germany	14	6	0	no	43%
Italy	20	8	1	yes	40%
Poland	33	12	2	yes	36%
Portugal	7	7	0	no	100%
Total	99	53	6 (6%)	2 countries	54%

was highest in Poland, Italy and Germany. In total, 99 schools were contacted and the final response rate was 54% (with great variations from 36% in Poland to 100% in Finland and Portugal).[7]

Refusal to participate in the study was due both to increasing institutional pressure on schools, schools having been over-researched in recent years, and the challenges associated with managing the COVID-19 uncertainties. Denying participation in the survey took either the form of lack of response to researchers' emails and phone calls, polite refusals or even annoyed and angry feedback. When the researchers had not collaborated with the principal or teachers prior to this project, contacts with schools through emails were more likely to fail eliciting any response from the school. Beyond numerous explicit or silent refusals, some schools in Estonia, Poland and Italy retracted their participation to the study after data collection had already started, due to the uncertain and constantly evolving pandemic situation.

Moreover, the participation of schools was also challenged by time pressure and a mismatch between the researchers' timeline and the schools' calendar: for example, since the start of data collection was postponed to April and May due to the pandemic situation, in some countries (including Finland, Italy, and Poland) the

[7] Reflecting on the school response rate, our response rate (54%) seems high compared to school-based health related behaviour surveys (less than 40% in ESPAD, HBSC, and ISRD in the investigated countries Germany, the Netherlands, England, and USA), but relatively low compared to other school-based surveys on academic performance (52–93% in PISA and TIMSS for the same countries), which have a high public profile, "translating into pressure on schools to participate" (van der Gaag et al., 2019, p. 394–396). However, it is difficult to compare surveys with a different degree of voluntary participation. Furthermore, in our sample, the national school response rates of 100% (Finland and Portugal) did not result in high pupil response rates in these countries.

survey administration clashed both with OECD PISA tests, final exams and the end of the school year. Accordingly, in Italy and Poland data collection was postponed to the beginning of the following academic year (fall 2021).

The first contact with schools was made by the following two main patterns. Firstly, in most countries, researchers relied on prior collaborations with school principals or teachers. Having already built a collaborative and trustful relationship with certain teachers who were highly motivated in participating in the study meant that the school principals were also more easily persuaded on the value of getting on board. Secondly, and in addition to or as an alternative to prior collaborations, researchers pursued more institutional pathways. This includes contacting a researcher responsible for a given school district, who has strong relationships with both schools and local authorities (Finland), representatives from the city councils, who contacted schools and organized a first collective meeting with all the schools interested in taking part in the study (Portugal), and a formal endorsement of the project from the city council in order to approach schools where no prior collaboration existed with an institutional support (Italy). All survey partners acknowledge that prior collaboration with schools facilitated both the first contact and the following collaboration. From the countries' school response rates (*Table 1*) it seems that professional and official networks linking research, schools and local administration (as in Finland and Portugal) were most helpful for recruiting schools.

Negotiating access required multiple contacts with different levels of authority in the school, including online meetings with the principal, the teachers, and, sometimes, parents aimed at presenting the projects and highlighting the benefits for the school and the children in taking part in the study. Depending on the size of the school and the role of the first contact within the school's organizational structure, the contact person remained the school principal themselves, taking on the organization of the project directly, or teachers with key responsibilities (for example, teachers responsible for the digital citizenship curriculum or cyber-bullying prevention). In some schools, the responsibility to support the organization of the survey was delegated to the IT specialists, e.g., for remote learning platforms and sessions (Estonia), or to school counsellors and psychologists (Poland). Survey partners agreed that the commitment of teachers was crucial to the success of the survey, as the teachers mediated the information flow from researchers to children and their parents and could support the sensitive and problematic process of getting parental consent. Supportive teachers would also help researchers to deal with logistical and other unforeseen problems which might emerge during data collection (including when students needed extra-time to fill in the survey).

Consistent with the literature on recruitment and collaboration with schools, therefore, our experiences point to the importance of motivating the school principal, teachers and parents with a "take and give" approach (Madge et al., 2012). In many cases, the topic of the survey itself – digital skills, online risks and children's wellbeing – represented a major source of motivation for schools. For example, Finnish researchers agreed with their contacts that the information produced by ySKILLS would integrate, or partly replace, the cities' own annual measurements of pupils/school wellbeing, etc. In Italy, since the National Plan for Digital Education implemented in 2015 introduced compulsory digital citizenship education at all levels and curricula, teachers would integrate the survey into digital citizenship education activities.

Researchers promised concrete benefits to the participating schools, including school-specific feedback on each wave's findings, training sessions for teachers and school staff, and awareness initiatives for parents. The possibility of using the "ySKILLS quiz" as well as another research instrument developed in the project in the future, namely the "performance test", as educational tools for the development of digital skills, also contributed to ensuring the school's commitment to the project. Incentive strategies may further increase participation (McGonagle, 2020), but since EU regulations did not allow remunerating schools with tangible gifts, we partially compensated this by distributing symbolic gifts such as appreciation certificates for the students, personal appreciation letters to the school principals, teachers, and staff who assisted in the fieldwork, and ySKILLS banners which schools could place on their websites.

While persuading schools to get on board presented a major challenge in each country, researchers reported high support from their contact persons in schools, even if the fieldwork meant additional administrative work for teachers, principals and other staff. Obtaining parental consent, organizing the data collection in order to minimize disruptions of ordinary teaching activities, and preparation of the list of nicknames necessary for network data collection required a huge effort on their part.

Problems with Obtaining Parents' Informed Consent

In planning the longitudinal survey of the ySKILLS project, we initially aimed at obtaining the participating students' informed assent and informed consent of one of their parents or legal representatives in all six survey countries. Based on this condition, the clearance from the Social and Societal Ethics Committee of KU Leuven was obtained for the whole project and for the longitudinal survey. As a general principle, the Board of the Ethics Committee discourages passive (opt-out) consent procedures (Application Dossier Social and Societal Ethics Committee, 2020, p. 7); however, they do not exclude passive informed consent procedures under certain circumstances.

The project team developed information and consent forms for students and parents, providing, inter alia, detailed information about linking the data across the three waves and the pseudonymization process (the code system) in place to lessen the participants' and their parents' concerns. Furthermore, contact information of national researchers was provided and it was pointed out that consent can be withdrawn anytime without further explanation. The information and consent forms were prepared by the national project teams based on the specific national regulations (e.g., providing national legal contact information). Due to the length of the questionnaire (it was designed for a full teaching unit of 45–50 min), we did not attach it to the information and consent forms. Instead, parents and students were informed that the questionnaire is available at the schools on approval ahead of time.

According to the national regulations regarding the age of consent *(Table 2),* both child and parental consent had to be asked for in all grades (involving 12- to 15-year-old students) in three survey countries (Italy, Poland, and Portugal). In the other three countries, the older students (aged at least 14 in Germany, and at least 15 in Estonia and Finland) could give consent themselves. In Germany, the researchers communicated with parents by class or grade levels (via teachers) which made it feasible to ask for parental consent only in grades 6–8. In Estonia and Finland, the researchers could communicate with parents only via the schools' online systems, and parental consent was asked in all grades to streamline research and to simplify the procedure for schools.

Regarding the form of the parental consent procedure *(Table 2),* national and/or school regulations required obtaining *active* parental consent in five countries. In Estonia, the form of parental consent (active or passive) is not explicitly stated in regulations, and research practices vary. In the ySKILLS survey research practice, the country teams started by asking active parental consent via different, mostly online, channels *(Table 2).* Estonian schools, for example, were on distance learning mode at the beginning of the data collection, and parents could be reached only via online communication platforms. The initial endeavor of obtaining active written consent from parents through such platforms resulted in a very low response rate (26% in one school) as most parents were exhausted by online communication and/or indifferent or not used to consent actively online. The Estonian team decided to follow the suggestion by some schools to switch to passive parental consent which is a common and culturally accepted practice in the country context (for more details see Kalmus et al., 2022). Other countries stuck to the form of active consent procedure, as this was also required by schools' administrators, possibly reflecting a "free of troubles" line of thought (cf. Liu et al., 2017), or by national regulations (as is the case in Germany).

The requirement of active parental consent not only challenged the implementation of field work but also involved ethical problems and raised questions of how to solve them, as the following quote from the expert interviews illustrates:

> "Unexpectedly for researchers, obtaining parental consent via online channels sometimes accidentally excluded the child from the teacher-parent communi-

> cation on this matter. A few children, unaware of their parent's refusal, turned up in online sessions, willing to participate in the survey. This raised an ethical dilemma about respecting the child's rights and dignity versus parental will. Teachers and researchers tried to solve those cases as discreetly as possible, e.g., by letting the child fill in the survey and deleting the data later." (Estonian expert)

Following the active consent requirement in the research practice brought further problems. In several countries, the response rate was very low (e.g., 38.1% in Finland) as many parents did not give their consent (most of them simply did not respond via the schools' online systems) *(Table 2).* In Italy and Portugal, parental non-consent ranged between 90% and 100% in some school classes, and in Poland, two schools had to withdraw from the survey due to the low parental consent rate. The low response rate added to the problems with scheduling data collection units as described in the previous section and led to the need to recruit new schools for the survey and to partially postpone fieldwork in two countries (Italy and Poland).

Table 2 provides an overview of our actual research practices regarding parental consent across the six countries:

Table 2 Parental consent (ySKILLS survey 2021)

Country	Age-related requirements	Active vs passive in national regulations	Form of procedure	Rejection rate	Students' response rate
Estonia	< 15 years (grades 6–8); was asked in all grades	Unspecified	Active and passive, via online channels	5% (actively)	74.9%
Finland	Grades 6–8; was asked in all grades	Active required	Active, via online system / paper	62% (parent + child; 11% actively; 51% passively)	38.1%
Germany	< 14 years (grades 6–8)	Active required	Active, via online system / email / paper	11%	79.9%
Italy	All grades (6–9)	Active required by schools	Active, via online forms / platform / email	45% (90% in some classes); fieldwork postponed	50.2%
Poland	All grades (6–9)	Active required by schools	Active, organized by schools	7%; 2 schools withdrew; fieldwork postponed	69.9%
Portugal	All grades (7–10)	Active required	Active, on paper	39% (two classes collectively)	61.7%

Discussion

Collaboration with Schools in Times of the COVID-19 Pandemic

While the challenges of collaborating with schools have already been documented in the literature on doing research with children and young people (Madge et al., 2012; Mishna et al., 2012; Rice et al., 2007), the COVID-19 pandemic added an additional layer of complexity. Indeed, schools were cautious of starting new projects and collaborating in an uncertain and rapidly changing situation, where moments of full remote learning were followed by equally complex periods of hybrid teaching. Although the first wave of data collection took place in the second year of the pandemic, schools were still facing high degrees of uncertainty and had to switch teaching modes several times during the school year. For example, when schools were approached in Italy at the beginning of 2021, students were taught in a classroom setting, but all grades switched to remote schooling in March for four consecutive weeks. Such uncertainties had repercussions on both the fieldwork schedule and school's willingness to cooperate.

Therefore, recruiting and collaborating with schools under such circumstances required additional background preparation. Some national survey teams had to increase the number of team members in order to carry out the data collection. Others had to invest more working hours into fieldwork preparation than expected. Moreover, national teams had to approach more schools in order to reach the agreed sample size. These challenges have also had implications for the research findings. As explained above, partners adopted various strategies to ensure collecting a (even if not a representative) diverse and inclusive sample. In the face of (at times last minute) refusals from the selected schools to participate in the survey, additional schools had to be recruited. Thereby it was not always possible to strictly follow the original country-specific selection criteria (above all, regarding SES). Furthermore, in two countries data collection had to be postponed to a later stage (to the beginning of the following academic year) which might cause problems of comparability and interpretation.

Active Parental Consent and Implications for Data Quality and Interpretation

Problems with obtaining active parental consent have direct implications for research outcomes. It is probable that some systematic sampling biases result from the non-random selection of students for the study, by which some segments of the student population are over-represented while others are under-represented (see Liu et al., 2017). In our survey, for instance, the parental consent rate was highest in a religious school in Italy (with 98–100% of students per class). Also, variation in the form of parental consent procedures and rejection rates between the countries needs to be analyzed and considered in the interpretation of findings. In the context of the Open Access Data policy, this means that in order to avoid misinterpretations, secondary analyses of the data collected under such complex and nationally varying circumstances cannot be encouraged without proper awareness and consideration of all contextual factors.

Our experiences also have some wider political implications. Firstly, we should keep in mind that participation in social research is more than just being a "data subject"; it is also a way and opportunity for expressing one's opinions and preferences, and exercising voice, agency, and power (Houghton, 2018). Therefore, the requirement of active parental consent procedure may conflict with children's civic rights, tending to discriminate against more vulnerable children. Furthermore, the requirement of active parental consent may result in biased samples and unreliable research findings, which, in turn, lead to inadequate policy recommendations that, again, are more likely to be inconsiderate of the concerns and needs of more vulnerable groups (Anderman et al., 1995).

We need to assume that the ethical dilemma concerning active parental consent and children's rights has become more acute in the "post-truth" and "(post)-pandemic" society. While on the global level general trust in science has risen during the COVID-19 pandemic, considerable differences in trust levels between the world's regions and social groups exist (Wellcome, 2020). For instance, in the United States, confidence in scientists is significantly stronger among Democrats and those with high self-evaluated science knowledge (Pew Research Center, 2019). Thus, we may assume that the attitudes of parents towards science and hence, their children's participation in research, may be diverse, perhaps even polarized. Therefore, considering the transforming information and political environment, the stakeholders in social research should revisit the ethical requirements concerning active parental consent and make efforts to enhance what we call "research literacy" – a set of knowledge and attitudes necessary for informed and active participation in scientific research – as an important new dimension of students' and parents' active citizenship.

Conclusions and Recommendations

Our experience has challenged the idea that school-based surveys are a more effective and less time-consuming way to collect data about and from children. While our response rates were still better than what could have been expected from collecting data from 12- to 15-year-olds in non-school-based surveys (e.g., with quota samples), the efforts for the researchers were higher than expected, calculated and budgeted for in the project. We conclude that, in order to reduce non-response and non-response biases, (school-based) surveys benefit from data collection by the researchers but require appropriate time and personnel resources. Taking into account our longitudinal approach which means aiming at surveying the same sample over three years, we assume, however, that for minimizing non-response and sampling biases in wave 2 and 3, collaborating with schools promises best outcomes.

We were able to identify the following preconditions and facilitating factors for a successful recruitment process and fruitful collaboration with schools: 1) personal contacts with school principal or teachers prior to the project, 2) existing professional networks between schools, local administration and research, 3) committed principals and teachers, 4) measures for increasing parental consent, 5) respecting the school calendar and school events, 6) a school-relevant topic of research (such as digital skills which is in line with general national educational programs), and 7) further benefits for the schools (such as educational tools). Schools are unlikely to take part in the research if they do not find it worthwhile and feasible. However, this judgement is contingent upon a number of conditions: it may be that some teachers find the research topic of particular interest to them personally; alternatively, the research topic could fill in a gap in the curriculum or help teachers plan innovations in the curriculum. While it has its costs, nonetheless, the researchers-schools collaboration can be mutually beneficial: researchers can get access to the same group of children over years with less drop-outs, while offering support in forms of teacher training, and/or meetings with teachers and parents to present the initial findings and issues that concern them most (e.g., cyber-bullying, etc.).

We found that active parental consent as required by national and/or university regulations in many countries is problematic regarding ethical concerns about children's rights to express their own views (cf. The United Nations, 1989) and an assumed non-response bias, i.e., socially disadvantaged children and adolescents seem to be more likely to be excluded from participation in the survey. Therefore, a flexible, culture- and context-sensitive approach is needed to enable weighing the pros and cons of active parental consent procedures against the aims, focus and methods of each study. In school-based social research, it is sufficient to rely on one main gatekeeper (for instance, the school), parents' passive consent and adolescents' own informed consent.

For data analysis, it is important to consider possible limitations due to exclusions of students if their parents did not allow them to participate in the survey. Whether and in which ways such a sample bias limits the meaningfulness of data depends on the research questions. Assuming that children whose parents were skeptical about scientific research and considering that the majority of the excluded children were from lower SES backgrounds, leads to consider that the children excluded by parental non-consent might have fewer digital skills than the average of those participating (cf. Paus-Hasebrink et al., 2019). This means that the bias based on parental consent might be a relevant limitation in certain aspects of ySKILLS data analysis.

This article, furthermore, contributes to and complements the literature about the challenges of doing research in the context of a pandemic. While some research has addressed the ways in which COVID-19 restrictions shaped the research process by focusing on the design of the survey instrument (Dales & Kottman, 2021), we focused on the process of data collection and the challenges of collaborating with schools and obtaining parental consent. The challenges caused by the COVID-19 pandemic have affected the process of conducting this school survey and social distancing rules and restrictions that led to the temporary closure of schools added to the usual complexities of doing survey research in schools. In pre-pandemic times, schools were already over-exploited in a research context. During the pandemic, teachers and staff, but also parents experienced an increasing amount of communicative activity and administrative work (Beilmann et al., 2023). The high rejection rates of parents in our study can be related to the COVID-19 situation to some extent. Further, COVID-19 related restrictions were implemented at slightly different times across Europe, adding an additional layer of complexity to the usual challenges involved in doing cross-cultural research. For example, due to problems with recruiting schools and students (including supportive parents), data collection had to be postponed in two countries, which might affect comparability and complicate interpretation. In summary, however, doing research under the COVID-19 pandemic provided valuable lessons in terms of increasing the resilience of all parties (including the young participants themselves), and improving methodological reflexivity as well as creativity. Therefore, although the circumstances were exceptional under many respects, we believe the lessons learned from this project can be extended to doing fieldwork in schools increasingly overburdened with research and bureaucratic demands.

For research policy and future studies employing school surveys, we provide the following seven ethical and practical recommendations.

(1) In designing or reconsidering ethics regulations, setting the age of consent for social science research should be consistent with the evolving capacities of children to enable them to express their views freely (in accordance with the UN Convention on the Rights of the Child; The United Nations, 1989);

(2) Asking for active parental consent, if not required by regulations, should be avoided to respect young people's rights, agency, and dignity to the fullest;

(3) In cross-cultural studies, the country/regional context of research regulations and practices has to be taken into account when deciding on the mode of parental consent;

(4) In the contexts where active parental consent in school surveys is obligatory, researchers have to consider that it can lead to a non-response bias and therefore should employ (well-prepared) practices to inform and encourage parents;

(5) The mode of communication when informing parents and asking their consent has to be technically accessible and convenient;

(6) Researchers have to negotiate carefully between providing students and parents as much information on the research and the data collection instruments as possible, and not producing an information overload;

(7) Researchers and educators should make efforts to enhance students' and parents' research literacy to encourage informed participation in social studies.

References

Anderman, C., Cheadle, A., Curry, S., Diehr, P., Shultz, L., & Wagner, E. (1995). Selection bias related to parental consent in school-based survey research. *Evaluation Review, 19*, 663–674.

Application Dossier Social and Societal Ethics Committee (2020). Leuven: Social and Societal Ethics Committee of KU Leuven.

Baker, J.R., Yardley, J.K., & McCaul, K. (2001). Characteristics of responding-, nonresponding- and refusing-parents in an adolescent lifestyle choice study. *Evaluation Review*, *25*(6), 605–618.

Barker, J., & Weller, S. (2003). 'Never work with children?': the geography of methodological issues in research with children. *Qualitative research, 3*(2), 207–227.

Bartlett, R., Wright, T., Olarinde, T., Holmes, T., Beamon, E. R., & Wallace, D. (2017). Schools as sites for recruiting participants and implementing research. *Journal of community health nursing, 34*(2), 80–88.

Baruch, Y., & Holtom, B. C. (2008). Survey response rate levels and trends in organizational research. *Human Relations*, *61*(8), 1139–1160.

Bedrosova, M., Zlamal, R., Dufkova, E., Machackova, H., & Waechter, N. (2022). *Longitudinal Survey 1st Wave Technical Report: Estonia, Finland, Germany, Italy, Poland, Portugal.* ySKILLS Work package 4; Task 4.2. Leuven: KU Leuven.

Beilmann, M., Opermann, S., Kalmus, V., Vissenberg, J., & Pedaste, M. (2023). The role of school-home communication in supporting the development of children's and adolescents' digital skills, and the changes brought by Covid-19. *Journal of Media Literacy Education*, *15*(1), 1–13. doi: 10.23860/JMLE-2023-15-1-1.

Bogner, A., Littig, B., & Menz. W. (2014): *Interviews mit Experten. Eine praxisorientierte Einführung.* Wiesbaden: Springer VS.

Brix, J., Wich, P., & Schneekloth, U. (2017). *Beziehungen und Familienleben in Deutschland (2016/2017). Welle 9.* Technischer Report. München: TNS Infratest Sozialforschung. Retrieved November 10, 2022, from the TNS report: https://www.pairfam.de/fileadmin/user_upload/uploads/Neu_10/Method%20Reports/Methodenbericht%2C%20pairfam%20Welle%209%202016-17.pdf

Cavazos-Regh, P., Min, C., Fitzsimmons-Craft, E. E., Savoy, B., Kaiser, N., Riordan, R., Krauss, M., Costello, S., & Wilfley, D. (2020). Parental consent: A potential barrier for underage teens' participation in an mHealth mental health intervention. *Internet Interventions*, *21*, 100328, 1–7.

Clary, K. L., Reinhart, C. A., Kim, H. J., & Smith, D. C. (2021). Improving recruitment procedures for school-based surveys: Through the lens of the Illinois Youth Survey. *Journal of school health*, *91*(3), 250–257.

Courser, M. W., Shamblen, S. R., Lavrakas, P. J., Collins, D., & Ditterline, P. (2009). The impact of active consent procedures on nonresponse and nonresponse error in youth survey data: Evidence from a new experiment. *Evaluation Review*, *33*(4), 370–395.

Dales, L., & Kottman, N. (2021). Surveying singles in Japan: qualitative reflections on quantitative social research during COVID time. *International Journal of Social Research Methodology* (published online Nov 2021). doi: 10.1080/13645579.2021.1998758.

Dent, C.W., Galaif, J., Sussman, S., Stacy, A., Burton, D., & Flay, B.R. (1993). Demographic, psychosocial and behavioral differences in samples of actively and passively consented adolescents. *Addictive Behavior*, *18*(1), 51–56.

Doeringer, S. (2020). 'The problem-centred expert interview'. Combining qualitative interviewing approaches for investigating implicit expert knowledge. *International Journal of Social Research Methodology*, *24*(4), 1–15.

Fizeşan, B. (2012). Digital engagement among Eastern European children. *Studia Universitatis Babes-Bolyai-Sociologia, 57*(1), 83–99.

Geis-Thöne, W. (2021). Lebenslagen von Kindern und Jugendlichen mit fremdsprachigen Elternhäusern. *IW-Trends. Vierteljahresschrift zur empirischen Wirtschaftsforschung 48*(1), 1–22. doi: 10.2373/1864-810X.21-01-01.

Haddon, L., Cino, D., Doyle, M-A., Livingstone, S., Mascheroni, G., & Stoilova, M. (2020). *Children's and young people's digital skills: a systematic evidence review.* KU Leuven, Leuven: ySKILLS. https://doi.org/10.5281/zenodo.6921674

Houghton, C. (2018). Voice, agency, power: A framework for young survivors' participation in national domestic abuse policy-making. In S. Holt, C. Øverlien & J. Devaney (Eds.), *Responding to Domestic Violence: Emerging Challenges for Policy, Practice and Research in Europe* (pp. 77–96). London: Jessica Kingsley Publishers.

Iltis, A. (2013). Parents, Adolescents, and Consent for Research Participation. *The Journal of Medicine and Philosophy*, *38*(3), 332–346.

Kalmus, V., Opermann, S., & Tikerperi, M-L. (2022). Conducting school-based online survey during the COVID-19 pandemic: Fieldwork practices and ethical dilemmas. In: Kotilainen, Sirkku (Ed.). *Methods in practice: Studying children and youth online.* (pp. 53−56). Hamburg: Leibniz-Institut für Medienforschung | Hans-Bredow-Institut (HBI); CO:RE - Children Online: Research and Evidence. doi: 10.21241/ssoar.83031.

Kuckartz, U. (2014). *Qualitative Inhaltsanalyse. Methoden, Praxis, Computerunterstützung.* Weinheim, Basel: Beltz Juventa.

Lindsay, J. (2005). Getting the numbers: The unacknowledged work in recruiting for survey research. *Field Methods*, *17*(1), 119–128.

Liu, C., Cox, R.B. Jr., Washburn, I. J., Croff, J. M., & Crethar, H. C. (2017). The effects of requiring parental consent for research on adolescents' risk behaviors: A Meta-analysis. *Journal of Adolescent Health*, *61*, 45–52.

Madge, N., Hemming, P. J., Goodman, A., Goodman, S., Kingston, S., Stenson, K., & Webster, C. (2012). Conducting large-scale surveys in secondary schools: The case of the Youth on Religion (YOR) Project. *Children & Society*, *26(*6), 417–429.

McGonagle, K.A. (2020). The effects of an incentive boost on response rates, fieldwork effort, and costs across two waves of a panel study. *methods, data, analyses*, *14*(2), 241–250.

Micklewright, J., Schnepf, S., & Skinner, C. (2010). Non-response biases in surveys of school children: The case of the English PISA samples. Discussion Paper No. 4789. Bonn: Institute for the Study of Labor.

Mishna, F., Muskat, B., & Cook, C. (2012). Anticipating challenges: School-based social work intervention research. *Children & Schools*, *34*(3), 135–144.

Paus-Hasebrink, I., Kulturer, J., & Sinner, P. (2019). *Social inequality, childhood and the media: A longitudinal study of the mediatization of socialization*. Palgrave Macmillan.

Pew Research Center (August 2019). *Trust and mistrust in Americans' views of scientific experts*. Retrieved January 7, 2021, from the Pew Research Center website: https://www.pewresearch.org/science/2019/08/02/trust-and-mistrust-in-americans-views-of-scientific-experts/

Rice, M., Bunker, K. D., Kang, D. H., Howell, C. C., & Weaver, M. (2007). Accessing and recruiting children for research in schools. *Western journal of nursing research*, *29(*4), 501–514.

Ryan, K.J., Brady, J.V., Cooke, R.E., Height, D. I., Jonsen, A. R., King, P., Lebacqz, K., Louisell, D. W., Seldin, D. W., Stellar, E., & Turtle, R. H. (1979). *The Belmont Report*. Retrieved January 5, 2021 from the U.S. Department of Health and Human Services website:
https://www.hhs.gov/ohrp/regulations-and-policy/belmont-report/read-the-belmont-report/index.html

Schreiner, C., & Haider, G. (Eds.) (2006). *PISA 2006. Internationaler Vergleich von Schülerleistungen*. Technischer Bericht. Salzburg: Österreichisches Projektzentrum für Vergleichende Bildungsforschung. Retrieved November 10, 2022 from
https://www.yumpu.com/de/document/view/22304869/pisa-2006-technischer-bericht-bifie

The European Code of Conduct for Research Integrity (2017). Berlin: ALLEA. Retrieved January 5, 2021 from the All European Academies website:
https://allea.org/code-of-conduct/

The United Nations Convention on the Rights of the Child (1989). London: UNICEF. Retrieved January 5, 2021 from the UNISEF website:
https://www.unicef.org.uk/what-we-do/un-convention-child-rights/

van der Gaag, R. S., Herlitz, L., & Hough, M. (2019). Contemporary challenges in school recruitment for criminological survey research: lessons from the international self-report delinquency study in England, Germany, the Netherlands, and the United States. *Journal of contemporary criminal justice*, *35(*4), 386–409.

Wellcome (2020). *Wellcome Global Monitor: How Covid-19 Affected People's Lives and Their Views about Science*. Retrieved January 7, 2021 from the Wellcome Global Monitor website:
https://cms.wellcome.org/sites/default/files/2021-11/Wellcome-Global-Monitor-Covid.pdf

Zilka, G.C. (2019). The digital divide: Implications for the eSafety of children and adolescents. *International Journal of Technology Enhanced Learning, 11*(1), 20–35.

Information for Authors

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

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

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

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

- Please submit your manuscript via www.mda.gesis.org.
- The total length of the manuscript shall not exceed 10.000 words.
- Manuscripts should…
  - be written in English, using American English spelling. Please use correct grammar and punctuation. Non-native English speakers should consider a professional language editing prior to publication.
  - be typed in a 12 pt Roman font, double-spaced throughout.
  - be submitted as MS Word documents.
  - start with a cover page containing the title of the paper and contact details / affiliations of the authors, but be anonymized for review otherwise.
  - should be anonymized ("blinded") for review.
- Please also send us an abstract of your paper (approx. 300 words), a front page with a brief biographical note (no longer than 250 words as supplementary file), and a list of 5-7 keywords for your paper.
- Acceptable formats for Graphics are
  - pdf
  - jpg (uncompressed, high quality)
- Please ensure a resolution of at least 300 dpi and take care to send hiqh-quality graphics. Line art images should have a resolution of 500-1000 dpi. Please note that we cannot print color images.
- The type area of our journal is 11.5 cm (width) x 18.5 cm (height). Please consider this when producing tables or graphics.
- Footnotes should be used sparingly.
- Please number the headings of your article. Doing so will make the work of the layout editor easier.

- If your text includes formulas we would like to ask you to upload your text also as a PDF, additional to the Word document.
- By submitting a paper to mda the authors agree to make data and program routines available for purposes of replication.
- Response rates in research papers or research notes, where population surveys are analyzed, should be calculated according to AAPOR standard definitions.
- Please follow the APA guideline when structuring and formating your work.

When preparing in-text references and the list of references please also follow the APA guidelines:

Entire Book:

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

Journal Article (with DOI):

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

Journal Article (without DOI):

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

Chapter in an Edited Book:

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

Internet Source (without DOI):

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

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

gesis
Leibniz Institute for the Social Sciences

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