Title

Two Half-Truths Make a Whole? On Bias in Self-reports and Tracking Data.

Authors

Pascal Jürgens, Dept. of Communication, University of Mainz, Germany (p@atlasnovus.net)

Birgit Stark, Dept. of Communication, University of Mainz, Germany (birgit.stark@uni-mainz.de)

Melanie Magin, Dept. of Sociology and Political Science, NTNU, Norway (melanie.magin@ntnu.no)

Abstract

The pervasive use of mobile information technologies brings new patterns of media usage, but also challenges to the measurement of media exposure. Researchers wishing to, for example, understand the nature of selective exposure on algorithmically driven platforms need to precisely attribute individuals' exposure to specific content. Prior research has used tracking data to show that survey-based self-reports of media exposure are critically unreliable. So far, however, little effort has been invested into assessing the specific biases of tracking methods themselves. Using data from a multi-method study, we show that tracking data from mobile devices is linked to systematic distortions in self-report biases. Further inherent but unobservable sources of bias, along with potential solutions, are discussed.

Keywords: tracking data; self-reports; media exposure; quantitative methods; non-reactive measurement; survey; digital traces

Manuscript: 7079 words including headings, tables, references, and footnotes. 3 Graphics * 307 words = 921 words. Total = 7079 + 921 = 8000 words.

Introduction

Media exposure, as a measurement of contact with media content, is so elementary to virtually any kind of media effects research that it almost seems mundane. Yet, the proliferation of digital communication, along with the "high choice media environment" (Prior 2007) it created, has spurred a search for new, accurate measurements of exposure such as automated tracking of online activity. The driving factor behind this methodological trend is not only the incremental improvement of existing, survey-based measures. Its goal is also to establish access to media exposure in new contexts and with increased precision where established methods of data collection are lacking: Mobile devices now offer convenient access to a wealth of content, extending the time and locations where media can be consumed throughout the day. The result is a potentially growing power of communication and platforms, along with new kinds of media effects (Iyengar & Hahn 2009, Bennett & Iyengar 2008). Because these developments lead to increasingly elaborate and specific effect assumptions (Garrett 2013, Hayes 2013), researchers are in need of exposure measurements that reliably capture content that recipients were exposed to – be it at home or on the go, on computers or smartphones.

A major and known challenge in this endeavor is the limited ability of recipients to accurately recall or estimate media consumption (Prior 2009b, Prior 2013, Scharkow 2016). Even if self-reports had been accurate in the past, remembering usage throughout the day is poised to become increasingly difficult, since it entails remembering the devices and specific platform and source that lead there (Niederdeppe 2016). A growing literature provides evidence that self-reports are severely and systematically skewed when compared to tracking data from desktop PCs (e.g. Scharkow 2016) or mobile phone usage (Boase & Ling 2013). Self-assessments, these studies contend, suffer from severe bias in the form of over- and underestimation that becomes visible only when gauging them against tracking data as a “true” baseline. Whether tracking methods warrant this amount of trust, however, is an open question: After all, any new method comes with specific challenges and consequences, which in this case have so far drawn little scrutiny. In this paper, we set out to complement our methodological picture of tracking by focusing on three key phenomena: The specific types of bias inherent in tracking methods, the potential effects of differential self-selection and the consequences of differential bias in tracking methods and/or participant responses.

The paper is structured as follows: We first review the existing state of research on the relation between self-reports and tracking data. Building on that literature, we discuss the theoretical and pragmatic limitations in the data collection process of various tracking methods, with a special focus on mobile devices. As that section will reveal, there are numerous potential sources of errors at various stages in the data collection process, which warrant an investigation into biases in tracking data. We go on to show empirically that such biases exist, drawing on original data from a multi-method study comprising survey and tracking data. In order to establish the validity of data and method, we first replicate existing findings of biased self-assessments (RQ 1). Using the differences between participants who provided mobile and/or desktop tracking data, we then show a genuinely new type of bias, namely a differential bias in self-reports of people willing to share mobile tracking data (RQ 2). Finally, we assess the impact of this bias through a simulation exercise (RQ 3), which builds on a realistic statistical model of perceived polarization to show how strong tracking bias will impact results.

Literature Review: Self-report Bias, Direction and Sources

A growing list of study designs aims to bypass the insufficient reliability of self-reports by directly capturing trace data of digital media usage through various means (Revilla et al. 2017, Araujo et al. 2017, Vraga et al. 2016, Scharkow 2016). The results show rather consistently that there are strong systematic biases present in self-reports across different devices, settings and operationalizations. An early study that served to draw attention to the issues is Prior's (2009b) investigation of time spent on TV. By comparing survey-based self-assessments to Nielsen people meter data (which are generated from custom tracking devices on TVs, see Napoli 2003), he shows that individuals on average overestimate their TV usage by a factor of three, with younger respondents doing worse. Tapping into an earlier debate in political communication (Price & Zaller 1993), the paper suggests either using alternative methods for measuring exposure or instead focusing on deeper layers of processing: Instead of merely recording exposure, researchers should rely on measures of reception, which presuppose exposure but capture the resulting effects. Subsequent investigations support and even extend Prior's pessimistic diagnosis. Studying the usage of cell phones for text messages and phone calls, Boase and Ling (2013) validated survey answers against a rarely available true baseline – the log data from mobile network operators. While the results confirm the existence of bias, the authors document cases of systematic underreporting, in addition to overreporting.

Regression across various socio-demographic variables shows few clear patterns, although males clearly tend to overestimate.

Applying the same approach to Internet tracking data, Scharkow (2016) used a large representative household sample and extended the analysis to include specific online platforms. His analysis confirmed an overall tendency for overestimation, along with low and heterogeneous correlations (.15 to .57) between self-assessed and tracked time spent online, usage of social networking sites, video platforms and auction sites. In an extension of this design, Araujo et al. (2017) included tracking data from Android tablets in addition to PCs and assessed the validity of both self-assessed recent and typical usage. In line with their expectations, higher Internet usage was correlated with lower accuracy of survey measures, whereas the answers given for typical usage were less biased than those for recent usage. Other methods have by and large supported the conclusion of highly unreliable self-reports plagued with over- and underreporting, even though there does not seem to be a consensus on the causes yet. In a lab experiment, Jerit et al. (2016) confirmed both types of bias through manipulated exposure, while Vraga, Bode & Troller-Renfree (2016) come to the same conclusion after using eye-tracking.

The newer literature also presents several suggestions for decreasing errors when using self-reports. Andersen, de Vreese and Albæk (2016) test existing ways to measure exposure to specific programs, in particular the program list technique (Dillplane et al. 2013), where recipients check programs they watch regularly. Contending that such a binary measure of regular exposure is inadequate, the authors argue in favor of a dual measurement of both information channel (such as TV) and usage frequency, an approach they call "List-Frequency Technique". Guess (2015) used tracking data from the browser-stored history to assess the performance of three different types of exposure questions: Open-ended (participants name news sources from the previous 30 days in a text entry box), check-all (a list of sources is presented and participants mark those they used) and a forced-choice condition where each source had a mandatory yes or no answer. Of all three, open-ended responses were least biased, whereas the forced-choice condition incurred the largest error. Finally, Ohme, Albæk and de Vreese (2016) assessed the reliability and feasibility of mobile-based daily surveys. They report a high reliability when comparing measurements of exposure to political news over the course of 15 days. However, reliabilities were lower for online sources (operationalized as social media) than for offline sources (p. 148). There is clearly a solid amount of evidence on biased self-reports across a myriad of settings and methodological approaches. Some wordings do better than others, but in the end they will still exhibit marked differences from non-reactive tracking measurements. Going deeper, things become much less clear. For example, we still lack a coherent theoretical explanation for self-report bias that is empirically well-supported. There are also mixed findings regarding the predictors of bias (Scharkow 2016, Araujo et al. 2017). Most crucially, perhaps, is the lack of insight into the specific errors of the tracking methods that serve as the analysis baseline. We hence discuss theoretical considerations of tracking method reliability, before empirically assessing bias.

The Pitfalls of Tracking Data Collection

((ABOUT HERE: Figure 1: Error Sources in Tracking Studies, Simplified Schema))

Despite its growing popularity, we posit, too little attention has been dedicated to the specific potential error sources in tracking data, which is why we want to highlight some of them[1]. In general, the new sources of bias in tracking methods can be thought of as an extension of error types known from other kinds of data collection. Figure 1 shows a simplified schematic of the stages that constitute a tracking study, along with corresponding occurrences of bias. As in any study involving a subset of the true population, the first important step is drawing a representative sample (1). However, the issue of sampling bias arises a second time, insofar as participation in a tracking study is dependent on the participant. If, for example, participants need to consciously opt in to the study, self-install software and/or perform other actions in order to enable data collection, we may see a secondary sampling bias: The presence of tracking data is then contingent on participants' willingness and ability to provide data. In study designs which combine data from different settings, such as tracking of mobile and desktop devices, this bias may additionally differ between the groups. The result is then a differential selection bias (2), where the makeup of one self-selected group, say mobile phone users, differs from the makeup of another group such as desktop pc users. Following the sampling stages (1 and 2) which may change the *composition of the sample*, additional errors may occur in the *generation and collection of the data itself* (3). On the one hand, technical issues may distort the collection. Other causes may lie within participants themselves, such as reactivity to the knowledge of being

[1] In this section, we mostly draw on proprietary documentation by tracking providers such as Wakoopa, which is unfortunately not publicly available.

tracked. As is the case with sampling, these response biases may vary between subpopulations for which the data collection methodology differs. Finally, all three sources of bias (general sampling, tracking sampling and response bias) will cumulatively propagate into any further analysis (4) done on the data.

Limiting bias is possible by evading self-selection effects in sampling and tightly controlling technical error sources – but this requires close control over the overall study design. Researchers will need to be more involved in the conception and execution of tracking studies, which is complicated in particular by the evolution of mobile devices.

Scharkow (2016) considers mobile devices to be prohibitively heterogeneous targets, noting that "the sheer diversity of devices, applications and differences in user behavior make it impossible to capture a complex phenomenon like Internet use entirely using client logs". Although this statement might seem overly cautious, it is getting more and more accurate. Largely due to a substantial security and privacy-related effort of mobile operating system producers Google (Android) and Apple (iOS), access to usage data has become more and more restrictive. In general, there are three different ways of capturing digital device usage, each of which carries specific challenges.

(1) Installing software on the device that records usage. This is a popular approach. The GfK media efficiency panel data used by Scharkow (2016) is produced by software running on desktop PCs. Araujo et al. (2017) also used this method for android tablets. However, it does not work on iOS devices and has only limited application on newer Android devices.

(2) Routing the devices' Internet access through a special server that logs usage. This solution requires end users to configure their devices themselves, which can hinder uptake and increase mortality. Furthermore, there are significant limits to tracking precision: Encrypted connections (HTTPS) do contain the domain that was requested[2] – such as washingtonpost.com – but not the full path of the URL. In some cases the encryption can be dropped by installing a special certificate on the device, but many modern apps (for example Facebook and Instagram) are resistant to this technique. In this setting where activity is merely recorded through URLs, page visit durations and app usage durations can furthermore only be approximated, since the true point when they are closed does not necessarily generate records. Worse, there are subtle variations in the possible failure scenarios for this kind of tracking. For example, when Android devices are

[2] That is because the relevant standard mandates the visible transfer of the domain name – a procedure called Server Name Identification.

tracked via a custom VPN connection, that connection will be dropped automatically when the battery is low. At the same time, iOS devices can be tracked via proxies, but those have to be set up for every Internet connection (mobile and each WIFI).

(3) Finally, more precise access to usage data may be gained by *physically accessing devices* in order to either extract existing log data or in order to tamper with its software[3].

Each of these technical limitations may introduce differential response biases at stage (3) in figure 1 above.

Measuring Bias in Self-Reports and Tracking

Objectively measuring bias in a tracking method itself is inherently difficult, because we know that other benchmarks (surveys) carry error, too. We therefore attack the problem in three stages: First, we assess the validity of our data and method. RQ 1 is hence: Can we generally support existing findings of bias in self-reports with our sample and study design? RQ 2 aims at the distinction between respondents providing desktop and mobile device data: Do we find systematic differences in the bias of self-reports between both groups? Do these differences appear because of different socio-demographic makeup of the groups (which would point to a sampling issue) or not (which would be an indicator of differences in behavior)? Finally (RQ 3), using simulation, we explore an example model built from the data to show how bias in exposure measurements propagates into analyses of other variables. The analyses are made possible by a particular multimethod study design, which is presented first.

Study Design

The data used in this study was collected in integrated multimethod design combining cross-sectional and daily surveys as well as tracking data of the same sample (Stark et al. 2017). We used a commercial access panel to recruit a quota sample of 459 participants who were representative of the German online population (14-69 years) in terms of age, gender, education and usage of Facebook (condition: having an account and using it at least once a week). Participation was incentivized through the

[3] Such tampering is possible with Android devices and to a lesser degree with iOS devices, due to their strong security architecture. For a practical example of such techniques, see the open source software "objection": https://github.com/sensepost/objection.

access panel's bonus point system, paying an equivalent of one Euro per day plus a base amount of 3.5 Euros for a total of roughly $15. Descriptive statistics and question wordings can be found in the online appendix.

(a) An initial cross-sectional online survey assessed socio-demographics along with media usage. For Internet usage, respondents gave time estimates of usage on a "typical day" on an ordinal scale ranging from "less than one hour" to "more than six hours" per day in steps of on hour each. Usage of other sources was collected but is not considered further here.

(b) The main study module consisted of a daily online survey over the period of 14 days. On each of those days, participants were asked to list the two subjectively most important political issues of the respective day (open-ended). These issue mentions were coded manually afterwards[4]. For both issue mentions per day, the respondents self-assessed five constructs related to opinion formation, among them the perceived political polarization (for the theoretical concept see Yang et al. 2016). Those variables will be used in the assessment and simulation of bias towards the end of the paper.

(c) In parallel to the 14 day longitudinal module, tracking data was collected from computers (PC and Mac), mobile phones and tablets (Android and iOS) of participants. The installation of tracking software was optional, self-selected and incentivized with additional bonus points with a value of approximately three euros. Upon agreeing to take part, users received instructions for installing and setting up the software, with help available upon request. The technical tracking solution was provided and administrated by the market research company who in turn sourced it from the specialized supplier Wakoopa, which has since been bought by GfK[5].

Overall, computers produced 2,098,299 records (URLs) from 411 of 459 total users (the rest either failed to install the software properly or did not use their devices at all), whereas mobile devices generated 266,594 records (URLs and App usage) from 163 users. For each of those requested URLs, the tracking software reported an active duration in seconds, which reflects how long a user was actively exposed to the content[6]. Further pruning was done for validity and completeness of the survey part of the dataset: A total of

[4] Drawing on a sample of the open-ended answers and a review of the media coverage over the 14 days under investigation, we designed a codebook consisting of 163 hierarchical categories.
4704 unique issues were hand-coded by three trained coders, roughly 1500 items each. Krippendorff's alpha for a random benchmark of 40 codings across all three coders was an acceptable 0.71.

[5] https://www.gfk.com/insights/press-release/gfk-drives-its-global-digitalization-through-acquisition-of-digital-panel-specialist-netquest/

[6] For mobile devices using iOS, this duration was estimated through the timing of network requests. Exposure times were not truncated and hence can grow very large. A total of 65 URLs from desktop devices were open for more than

354 participants from the entire sample had at least five days of survey responses with valid political issues; other datasets were dropped. In the initial recruitment, 1818 individuals were contacted, resulting in an AAPOR response rate type 1 of 25% (recruitment) and 19% (ready to analyze). The resulting working sample of 354 participants showed no significant skew in terms of the sampling quota and gave a total of 8866 valid issue-specific answers out of a theoretical maximum of 9912 (354 participants x 14 days x 2 issues per day). Corresponding to the possible combination of devices, the tracking data yielded four distinct groups: 73 participants produced no tracking data whatsoever, 243 had at least one URL recorded from a PC but no mobile device, 25 participants sent at least one URL from a mobile device but had no PC data, and 13 sent data from both types of devices. While the number of individuals in these groups is rather low, each of them provided up to 14 days of tracking measurements, yielding a dataset of adequate size.

Assessing Tracking Bias

Discrepancy between Self-Reports and Tracking Data *(RQ1)*.

Following the strategy of prior studies, we first assess the discrepancy between self-assessment and tracking data by computing the difference between both values. Since our study design is based on days as the basic time unit, we aggregated both desktop usage times and mobile usage times into one daily tracking sum per user . The resulting scores, in the hypothetical absence of any technical measurement errors, reflect the precise amount of time a participant spent on digital devices on a particular day.

Self-assessment of time spent online originates from the initial cross-sectional survey, where Internet usage times for a typical day were collected on ordinal scales. Since those scales consist of hourly intervals, we transformed the answers into metric records placed in the middle of these intervals (0 -under 1 hour=0.5 hours; 1 - under 2 hours=1.5 hours etc.). The deviation from actual tracking data can then be calculated by subtracting the tracking estimate from the subjective estimate, and further decreasing that number by 0.5 to give a distance to the one-hour interval, not the artificially chosen center.

[[FIGURE 2 ABOUT HERE]]

one hour. We manually investigated these cases and found them to be valid: The very long exposure times stem from video streaming, online games, live tickers of sports events and online auctions.

In line with previous results, we find significant differences between tracking and self-reports of Internet usage (mean=.85, median=.68, sd=2.3) . Whereas the average amount of bias is lower than one might expect, this is due to the fact that both over- and underreporting are present.

Figure 2 shows box plots[7] of the relation between self-reports (x-axis) and tracking data (y-axis). Panel A shows data for the *average* daily Internet use per user; the box plots depict variation between users. Panel B instead uses the un-aggregated days, showing within-user variation and allowing inferences about the general volatility of measurements day-to-day. The dashed line in each panel represents the perfect precision, i.e. the points where both measurements match. Cases above the dashed line have more tracked time than expected (underreporting time spent). Consistent with Prior (2009b), Scharkow (2016) and Araujo et al. (2017), we see that the vast majority of people who say they are online for more than two hours do actually spend much less time. The error also grows along with peoples' own estimates: Self-declared heavy users overreport more than moderate users.

Mobile Tracking: Self-Selection or Response Bias? *(RQ2)*

One particularly interesting aspect of our design is the integration of tracking across both PCs and mobile devices. The sample contains subsets of users who only agreed to desktop tracking, only to mobile or both. Any systematic differences will point to either a differential selection effect (stage 2 in the tracking study schematic in figure 1) or a differential response effect (stage 3). The latter might occur when participants perceive mobile device tracking to be more invasive – for example due to the fact that mobile devices are more personal (as opposed to computers which might be shared) and used in many personal contexts (see works on everyday information seeking, e.g. Savolainen 1995).

[[FIGURE 3 ABOUT HERE]]

Figure 3 depicts the overestimation of Internet usage in hours (self-report in survey minus actual usage time), by availability of tracking data from devices; the dashed line again denotes an unbiased perfect

[7] Per convention, boxplots show the median as the central, horizontal black bar. The edges of boxes correspond to the lower and upper quartiles, whereas the narrow black lines represent the distance of 1.5 times the interquartile range. Outliers outside those ranges are plotted as dots.

match between survey answer and tracking data. The primary visible trend is the aforementioned relation between the expected usage time and the magnitude of the error. The longer people think they use the Internet (x-axis), the more do they tend to overestimate their online time. Participants who don't think they use the internet much at all (leftmost bin), on the other hand, exhibit negative bias, meaning that they underestimate actual time spent. The result is an inflation of distance from the center.

A second visible phenomenon is the difference between the desktop and mobile tracking groups. Users with only mobile data have a higher average bias than both other groups[8]. Even though they may in fact spend more time online (for example due to the simple availability of the device), they still exaggerate this difference, resulting in an even stronger overestimation. This difference does in fact support our assumption of second-order bias through sampling or response errors. The mixed group seems to fall in between both, but a clear assessment is unfortunately prevented by the low number of measurements and high variance.

[[TABLE 1 ABOUT HERE]]

We can further estimate the magnitude of this effect and reason about the underlying mechanisms through a proper model. That is, the self-assessment error can be used as a dependent variable to be explained by various socio-demographic variables and other predictors, in a similar vein to Scharkow (2016) and Araujo et al. (2017). Table 1 shows effect estimates with credible intervals from a bayesian linear mixed model predicting self-assessment bias on an average day. The model explains daily bias as a function of the individual level variables age, sex, education, income, self-assessed internet usage, and tracking type (dummy-coded with desktop as baseline) while accounting for group-level random intercepts within participants, days and political issues[9]. As expected, the number of hours spent online is a major predictor of bias – this represents the roughly linear trend seen in figure 3. The second important factor is the presence of

[8] Desktop tracking: mean 0.10, sd 2.61; mobile tracking: mean 2.56, sd 1.9; both tracking: 0.75, sd 2.56. Due to non-normality (Anderson-Darling normality test has $A = 17.313$ and $p \simeq 0$), a non-parametric Kruskal-Wallis (Kruskal-Wallis chi-squared $= 687.99$, $df = 2$, $p \simeq 0$) and subsequent Dunn post-hoc test were performed, revealing that all groups differ significantly from each other (each $p \simeq 0$).

[9] As DV, bias was coded as right-censored where values hit the theoretical ceiling. Left-censoring (where usage reaches 24 hours per day) did not occur. Convergence was unproblematic with nominal trace plots and minimal scale inflation (R hat).

mobile tracking data, while socio-demographic variables play no significant role. Hence, we see that even when controlling for potential differences in the demographic makeup of the groups of desktop and mobile device users, there still remains a significant difference in the strength of their self-assessment bias.

There are three possible interpretations for this, all of them troubling: First, there could be technical issues with the data collection. A possible indicator is the strikingly differing variance: Mobile device usage times have much lower variance, which might for example imply that they are artificially constrained (such as by the automatic lock function of smartphones). Second, people who participate in invasive tracking might have an extremely exaggerated impression of their own habits – even more so that "ordinary" respondents. If true, this would mean that mobile tracking is not any more biased than other tracking, only that mobile phone users overestimate how much they use the devices – possibly because they handle their phones so often. Finally, knowledge of the recording of their activity could motivate participants to suppress their usual behavior, leading to unnaturally low tallies. In any case, we find that the issue at hand seems to be caused not by distorted sampling (at least not along the controlled variables) but rather by some inherent difference in either the tracking method or differing user behavior.

Consequences of Tracking Bias: An Exploration (RQ 3).

The fact that both self-reports and tracking data suffer from distortions leads to a pressing need for methods that can mitigate any or both of those issues. While we have high hopes for future methodological advances, there are existing strategies even today that can help in limiting the harmful interference. In addition to seeking out innovative designs (see the section on theoretical issues above), tracking studies should try to explicitly incorporate different device types and model them as a "canary" variable in the analysis. If an analysis shows a strong effect from any of the devices, then there is likely either a methodologically induced bias or a source of response bias that should be addressed. The strength of this device effect can be interpreted as an indicator of the variance and hence uncertainty inherent to the tracking method. The arguably easiest way to do this is by using multilevel models, which allow for nested predictors that make efficient use of available information (Gelman & Hill 2006).

While a full investigation into statistical consequences and mitigation of the differential bias we identified goes beyond the scope of this paper, we still want to provide at least an outlook of these issues. We therefore conclude by building a realistic but simple model from the data used above and then simulate a higher level of bias in order to show its impact on results. The model uses socio-demographic variables, exposure measurements and participants' attitude towards an issue to explain "perceived polarization". Perceived polarization (see Yang et al. 2016) is a concept that captures peoples' impression of how deeply a society is split regarding an issue – how strongly the different camps oppose each other. In contrast to other forms of polarization (such as ideological or affective polarization), and with media coverage being an important source for people's assessment of the state of public opinion, perceived polarization is thought to be directly malleable by media effects. A comparative analysis of ten countries (ibid.) supports this assumption, revealing a strong effect from online (but not offline) news. Drawing on and extending this idea, one might assume that using the internet in general harbors the same potential, since much online activity is spent on sites which include social network elements that are susceptible to amplifying perceptions of differences (Stark et al. 2017). We therefore build models that use the base variables from the original study (Yang et al. 2016) while, for the purpose of this paper, replacing news sources with internet usage measures.

[[Table 2 about here]]

Using the "authentic" base model, we can simulate the effects of strong differential bias on the results by selectively increasing the existing overestimation of mobile device users through a higher self-assessed time[10]. Table 2 shows three different analyses: Model 0, which is computed using the authentic, unchanged data; model 1 which uses the data with an artificially increased bias for mobile device users but no information about the type of tracking data used, and model 2 which still uses simulated data, but contains dummy coded variables for mobile device tracking and mobile and computer tracking. Comparing the three, we can see that most predictors do not change dramatically. However, the role of tracking data is portrayed differently across the three models (see highlighted row): In the actual data (model 0), there is no strong relationship between online time tracked and perceived polarization (CI crosses zero). In model 1, where the self-assessment was inflated for mobile users, there is a strong, positive link between the two variables. Once we introduce the device factor, however, this exaggerated estimator shrinks to almost its realistic level,

[10] To accomplish this, we add a normally distributed delta with mean=10 and sd=1 to the survey-based usage time variable only for mobile device users. While it first might seem intuitive to subtract a delta of identical sum from the other, non-mobile cases, this would not serve the goal: Simulating differential bias means, after all, artificially raising a measurement whose true value is lower, while not changing the values of the other measurement type.

whereas we see a very large estimate for the mobile dummy variable. Thus, the “canary” device variable helped catch the hypothetical increased differential bias. Note that in this simple simulation exercise, the exposure variables and the other predictors are not strongly correlated, resulting in a relatively robust behavior of the regression models. Were there stronger interrelations (which we did not simulate for sake of the clarity of this demonstration), then other effect estimates would change as well.

Conclusions

This article is concerned with systematic bias in self-reports and tracking data on media exposure. Using a multi-method study design, we replicated prior studies' findings to show that self-reports are on average strongly exaggerated. Respondents tend to diverge from the center, with light users underestimating and heavy users overestimating the duration. By using differences in the groups of participants supplying data from computer and mobile devices, we were further able to show that there is a systematic second-order bias: Those participants who are willing to share mobile device data exhibit significantly higher overreporting than those who do not.

There are important conclusions to be drawn from this re-assessment of the state of combining tracking and survey data. First and foremost, it is crucial to admit that apart from very few exceptions (such as Boase and Ling 2013 or Dunn et al. 2012), tracking data should not by default be considered an unbiased source of "true" media exposure. Second, our findings of differential bias depending on the device type suggest that we cannot treat the results from different devices equally. In essence, tracking methods might result in exchanging lower measurement bias (than surveys) against a fragmentation of the meaning, where tracking results from separate devices incur different biases.

What, then, is a good way of attacking the problem of biased exposure measurements, if no one method proves faultless? Recapitulating the state of the literature, we see three options that could lead to increasingly valid, well-understood measurements:

(1) Development of improved tracking tools. Today's shortcomings of mobile device tracking not only arise from the design of smartphone operating systems, but also from the opaque nature of proprietary tracking solutions and restricted access to study participants. With more greater knowledge of and control over technical errors, along with direct access to recruitment, researchers will be able to counteract visible and reveal previously hidden error sources. A remaining challenge lies in the “black box“ of platforms and

their algorithms (Jürgens & Stark 2017). Clearly, the limitations to scholarly data collection are a serious impediment – they limit the ability to understand platform effects, and add further methodological uncertainty (Bachl 2018).

(2) Moving towards more precise survey methods, including high-frequency longitudinal designs, source- and issue-specific questions, as done in our study, along with measures of reception rather than exposure (Price & Zaller 1993).

(3) Combining data sources. Joining different methods, as we have shown, ultimately requires modeling their specific characteristics if one wants to reduce the resulting overall error. An especially interesting approach for doing this is the experimental assessment of bias by Jerit et al. (2016). Study designs can contain interventions, whose effects in turn are measured through tracking data. By compelling participants to use a particular stimulus and then gauging over- and underreporting, it may be possible to create a benchmark for the rest of the study.

Finally, the methodological discussion cannot and should not be detached from the crucial ethical considerations underlying not only scientific but also commercial tracking of user behavior. It is clear that advances in one area are directly related to the development of the other, albeit with both beneficial and harmful consequences.

This paper highlighted theoretical as well as empirical concerns about the use of new methods for measuring media exposure. Rather than being a panacea of precise measurement, tracking data itself is subject to errors, many of which are not well understood today. Even though our results reinforce earlier warnings, the fact that different data sources are available today will contribute to a better understanding of problems that existed all along, and lead to future methodological progress. In that vein, we hope for future research to unveil more of the cognitive and technical mechanisms at work.

Data and code availability

The authors are committed to transparent and replicable research. In accordance with the journal's guidelines, replication data may be obtained by request to the primary author. Supplementary information including R code is available in the online appendix.

Literature

Andersen, K., Vreese, C. H. de, & Albæk, E. (2016). Measuring Media Diet in a High-Choice Environment - Testing the List-Frequency Technique. Communication Methods and Measures, 10(2–3), 81–98. https://doi.org/10.1080/19312458.2016.1150973

Araujo, T., Wonneberger, A., Neijens, P., & Vreese, C. de. (2017). How Much Time Do You Spend Online? Understanding and Improving the Accuracy of Self-Reported Measures of Internet Use. Communication Methods and Measures, 11(3), 173–190. https://doi.org/10.1080/19312458.2017.1317337

Bachl, M. (2018). An Evaluation of Retrospective Facebook Content Collection. Open Science Framework. https://doi.org/10.17605/OSF.IO/6TXGE

Bennett, W. L., & Iyengar, S. (2008). A New Era of Minimal Effects? The Changing Foundations of Political Communication. Journal of Communication, 58(4), 707–731. https://doi.org/10.1111/j.1460-2466.2008.00410.x

Boase, J., & Ling, R. (2013). Measuring Mobile Phone Use: Self-Report versus Log Data. Journal of Computer-Mediated Communication, 18(4), 508–519. https://doi.org/10.1111/jcc4.12021

Bode, L., & Vraga, E. K. (2018). Studying Politics Across Media. Political Communication, 35(1), 1–7. https://doi.org/10.1080/10584609.2017.1334730

Cook, R. D. (1977). Detection of Influential Observation in Linear Regression. Technometrics, 19(1), 15–18. https://doi.org/10.2307/1268249

Dilliplane, S., Goldman, S. K., & Mutz, D. C. (2013). Televised Exposure to Politics: New Measures for a Fragmented Media Environment. American Journal of Political Science, 57(1), 236–248. https://doi.org/10.1111/j.1540-5907.2012.00600.x

Dunn, C. W., Gupta, M., Gerber, A., & Spatscheck, O. (2012). Navigation Characteristics of Online Social Networks and Search Engines Users. In Proceedings of the 2012 ACM Workshop on Workshop on Online Social Networks (pp. 43–48). New York, NY, USA: ACM. https://doi.org/10.1145/2342549.2342560

Garrett, R. K. (2013). Selective Exposure: New Methods and New Directions. Communication Methods and Measures, 7(3–4), 247–256. https://doi.org/10.1080/19312458.2013.835796

Gelman, A., & Hill, J. (2007). Data Analysis Using Regression and Multilevel/Hierarchical Models. Cambridge University Press.

Groves, R. M. (2006). Nonresponse Rates and Nonresponse Bias in Household Surveys. Public Opinion Quarterly, 70(5), 646–675. https://doi.org/10.1093/poq/nfl033

Guess, A. M. (2015). Measure for Measure: An Experimental Test of Online Political Media Exposure. Political Analysis, 23(1), 59–75. https://doi.org/10.1093/pan/mpu010

Hayes, A. F. (2013). Methodology of Selective Exposure Research: Introduction to the Special Issue. Communication Methods and Measures, 7(3–4), 145–146. https://doi.org/10.1080/19312458.2013.845500

Iyengar, S., & Hahn, K. S. (2009). Red Media, Blue Media: Evidence of Ideological Selectivity in Media Use. Journal of Communication, 59(1), 19–39. https://doi.org/10.1111/j.1460-2466.2008.01402.x

Jerit, J., Barabas, J., Pollock, W., Banducci, S., Stevens, D., & Schoonvelde, M. (2016). Manipulated vs. Measured: Using an Experimental Benchmark to Investigate the Performance of Self-Reported Media Exposure. Communication Methods and Measures, 10(2–3), 99–114. https://doi.org/10.1080/19312458.2016.1150444

Jürgens, P., & Stark, B. (2017). The Power of Default on Reddit: A General Model to Measure the Influence of Information Intermediaries: The Influence of Information Intermediaries. Policy & Internet, 9(4), 395–419. https://doi.org/10.1002/poi3.166Napoli, P. M. (2003). Audience Economics: Media Institutions and the Audience Marketplace. New York: Columbia University Press.

Niederdeppe, J. (2016). Meeting the Challenge of Measuring Communication Exposure in the Digital Age. Communication Methods and Measures, 10(2–3), 170–172. https://doi.org/10.1080/19312458.2016.1150970

Ohme, J., Albaek, E., & Vreese, C. H. de. (2016). Exposure Research Going Mobile: A Smartphone-Based Measurement of Media Exposure to Political Information in a Convergent Media Environment. Communication Methods and Measures, 10(2–3), 135–148. https://doi.org/10.1080/19312458.2016.1150972

Price, V., & Zaller, J. (1993). Who Gets the News? Alternative Measures of News Reception and Their Implications for Research. Public Opinion Quarterly, 57(2), 133. https://doi.org/10.1086/269363

Prior, M. (2007). Post-broadcast democracy: how media choice increases inequality in political involvement and polarizes elections. New York: Cambridge University Press.

Prior, M. (2009a). Improving Media Effects Research through Better Measurement of News Exposure. The Journal of Politics, 71(3), 893–908. https://doi.org/10.1017/S0022381609090781

Prior, M. (2009b). The Immensely Inflated News Audience: Assessing Bias in Self-Reported News Exposure. Public Opinion Quarterly, 73(1), 130–143. https://doi.org/10.1093/poq/nfp002

Prior, M. (2013). Media and Political Polarization. Annual Review of Political Science, 16(1), 101–127. https://doi.org/10.1146/annurev-polisci-100711-135242

Revilla, M., Ochoa, C., & Loewe, G. (2017). Using Passive Data From a Meter to Complement Survey Data in Order to Study Online Behavior. Social Science Computer Review, 35(4), 521–536. https://doi.org/10.1177/0894439316638457

Savolainen, R. (1995). Everyday life information seeking: approaching information seeking in context of “way of life”. Library and Information Science Research, 17(3), 259-294.

Scharkow, M. (2016). The Accuracy of Self-Reported Internet Use—A Validation Study Using Client Log Data. Communication Methods and Measures, 10(1), 13–27.https://doi.org/10.1080/19312458.2015.1118446

Stark, B., Magin, M. & Jürgens, P. (2017). Ganz meine Meinung? Informationsintermediäre und Meinungsbildung – Eine Mehrmethodenstudie am Beispiel von Facebook. Düsseldorf: LfM.

Singer, E. (2018). Survey Incentives. In The Palgrave Handbook of Survey Research (pp. 405–415). Palgrave Macmillan, Cham. https://doi.org/10.1007/978-3-319-54395-6_50

Vraga, E., Bode, L., & Troller-Renfree, S. (2016). Beyond Self-Reports: Using Eye Tracking to Measure Topic and Style Differences in Attention to Social Media Content. Communication Methods and Measures, 10(2–3), 149–164. https://doi.org/10.1080/19312458.2016.1150443

Yang, J., Rojas, H., Wojcieszak, M., Aalberg, T., Coen, S., Curran, J., … Tiffen, R. (2016). Why Are “Others” So Polarized? Perceived Political Polarization and Media Use in 10 Countries. Journal of Computer-Mediated Communication, 21(5), 349–367. https://doi.org/10.1111/jcc4.12166

Tables

Table 1: Predictors of Bias in Self-Reports (Bayesian Mixed Model)

Predictors	**Estimates**	**Credible Interval (95%)**	**Inflation Factor (R hat)**
Intercept	0.44	0.16 – 0.73	1.00
Tracking: Mobile	**1.96**	**1.47 – 2.45**	1.00
Tracking: Mobile and desktop	0.19	-0.07 – 0.45	1.00
Age (standardized)	**-0.32**	**-0.53 – -0.10**	1.00
Sex: Female	-0.11	-0.52 – 0.29	1.00
Education (standardized)	0.10	-0.13 – 0.32	1.00
Household income (standardized)	0.15	-0.05 – 0.37	1.00
Self-assessed average time spent online per day (standardized)	**1.44**	**1.24 – 1.63**	1.00
Random Effects			
$\sigma 2$	1.64		1.00
$\tau 00$ day	0.02		1.00
$\tau 00$ id	2.49		1.00
ICC day	0.01		
ICC id	0.60		
Observations	3059		
Bayes R2 / Standard Error	0.768 / 0.004		

Bayesian model with gaussian family and identity link. NUTS sampler from stan via brms package in R

Formula: bias_daily | cens(bias_censored) ~ device_factor + scale(age) + sex + scale(education) + scale(income) + scale(online_time_survey) + (1 | day) + (1 | id)

Samples: 8 chains, each with iter = 8000; warmup = 4000; thin = 1; total post-warmup samples = 32000; seed=112358

Table 2: Bayesian Models for Bias Simulation Response variable: Perceived polarization (0-100 low-high)

	Model 0 (with device factor)		Model 1 (no device factor)		Model 2 (with device factor)	
Data Type	**Authentic Data**		**Simulated**		**Simulated**	
Predictors	Estimates	CI (95%)	Estimates	CI (95%)	Estimates	CI (95%)
Intercept	43.87	41.24 – 46.13	43.63	41.28 – 45.85	46.58	44.03 – 49.16
Age (standardized)	-0.65	-1.63 – 0.32	**-1.08**	**-2.07 – -0.11**	-0.66	-1.64 – 0.31
Sex: Female	**-2.45**	**-4.10 – -0.67**	**-2.78**	**-4.55 – -1.13**	**-2.40**	**-4.11 – -0.62**
Education (standardized)	**1.94**	**1.02 – 2.91**	**1.85**	**0.89 – 2.78**	**1.94**	**1.00 – 2.90**
Household income (standardized)	-0.81	-1.72 – 0.06	-0.84	-1.75 – 0.03	-0.81	-1.68 – 0.08
Own position on issue	**6.33**	**5.46 – 7.16**	**6.40**	**5.61 – 7.20**	**6.32**	**5.53 – 7.17**
Simulated Bias Increase: Time spent online: Survey	**2.44**	**1.49 – 3.27**	-0.38	-1.16 – 0.36	**7.41**	**4.69 – 10.01**
Time spent online:Tracking	0.66	-0.22 – 1.61	**1.53**	**0.65 – 2.36**	0.66	-0.24 – 1.55
Tracking: Mobile	**-3.46**	**-6.00 – -1.00**			**-24.51**	**-33.01 – -16.76**
Tracking: Mobile and desktop	0.82	-1.67 – 3.38			0.80	-1.56 – 3.31
Observations	4892		4892		4892	
Bayes R2 / Standard Error	0.115 / 0.009		0.107 / 0.008		0.115 / 0.008	
WAIC	46848.28		46882.59		46848.56	
σ2	829.80		835.94		829.84	
τ00 (day, issue)	7.59, 44.88		7.62, 43.32		7.71, 44.96	

Bayesian model with gaussian family and identity link. NUTS sampler from stan via brms package in R

Formula: p.polarization ~ device_factor + scale(age) + sex + scale(education) + scale(income) + scale(online_time_survey) + scale(online_time_tracking_daily) + (1 | day) + (1 | issue) *[Model 1 omits device_factor]*

Samples: 4 chains, each with iter = 2000; warmup = 1000; thin = 1; total post-warmup samples = 4000; seed=112358. Convergence unproblematic, all R hat = 1.00

Figures

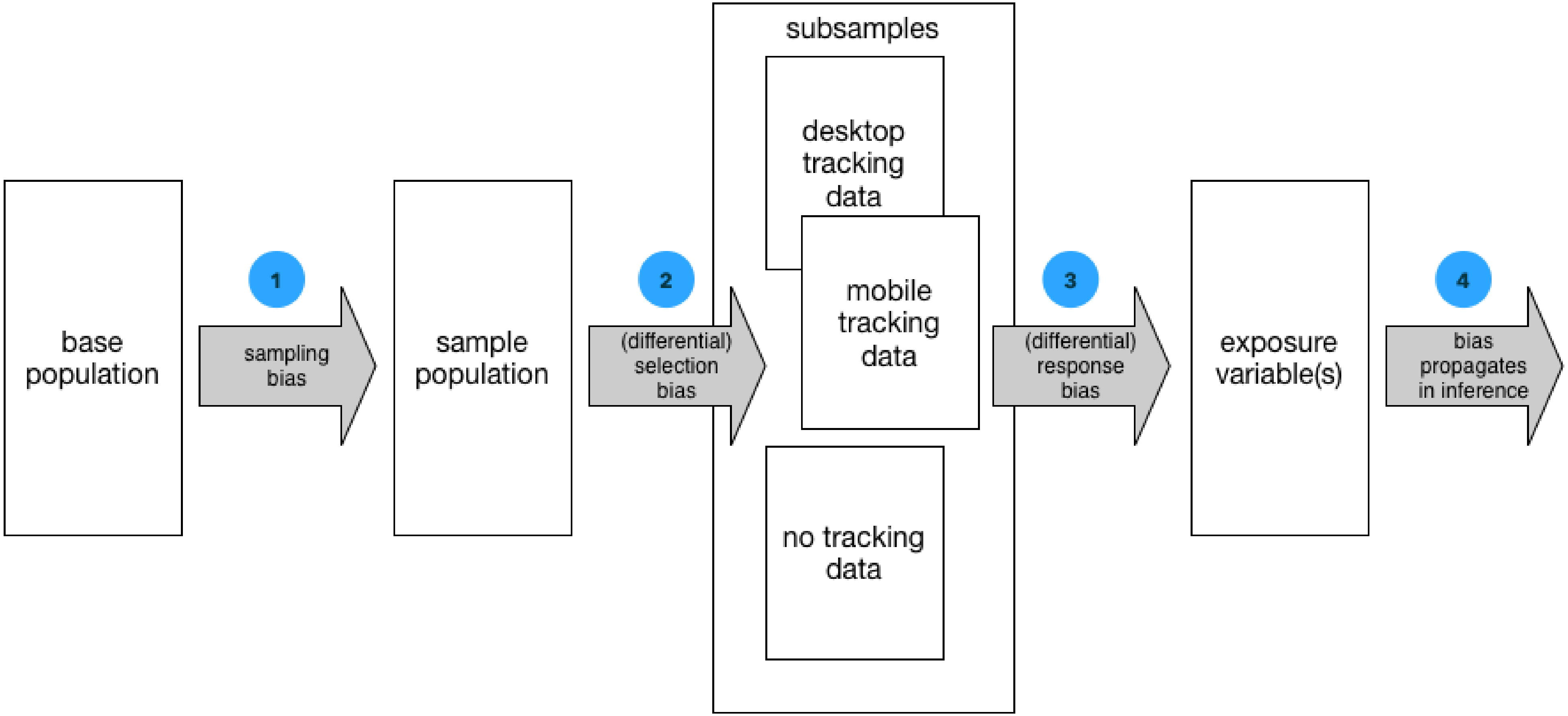


Figure 1: Error Sources in Tracking Studies, Simplified Schema

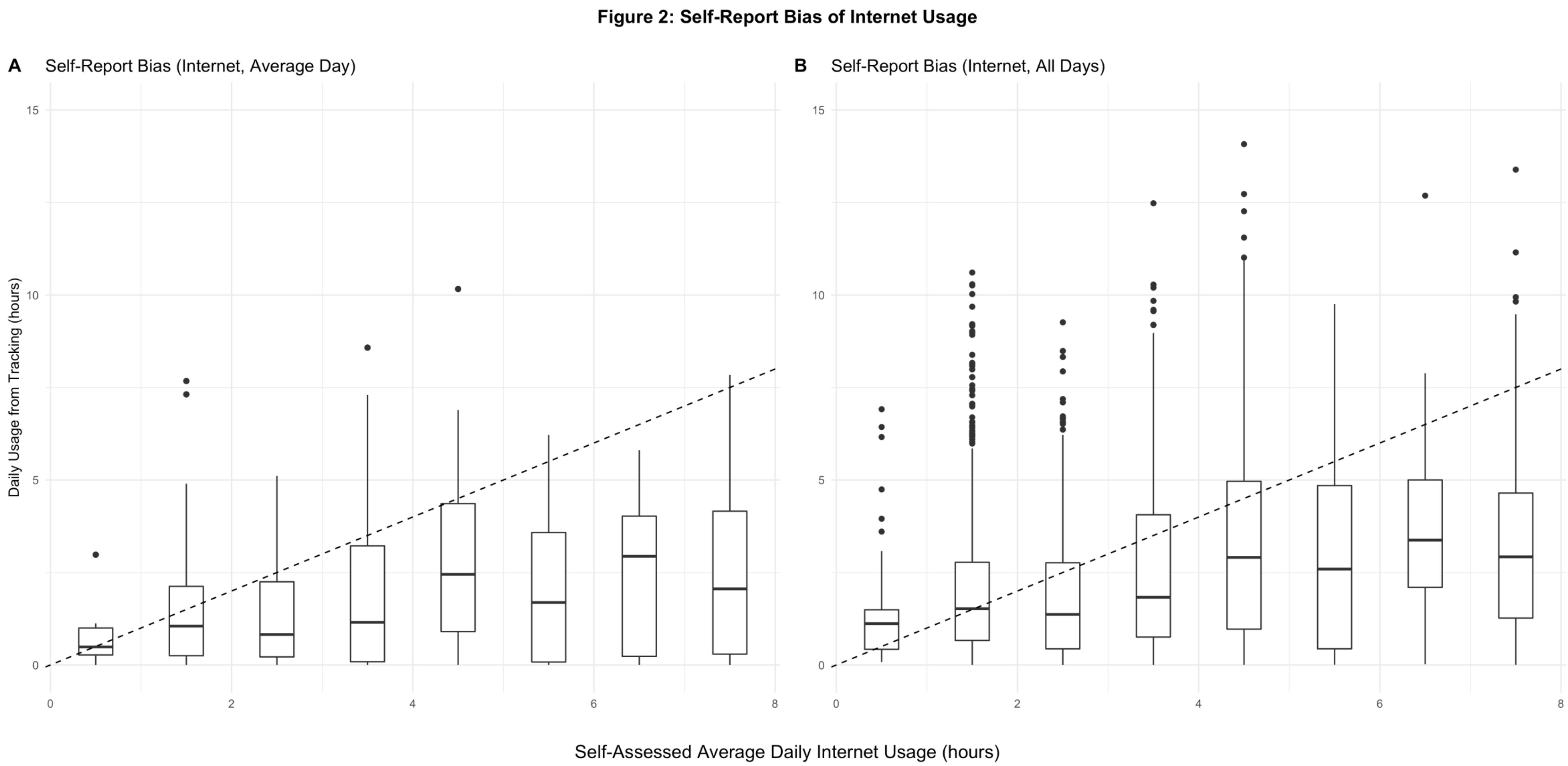


Figure 2: Self-Report Bias of Internet Usage

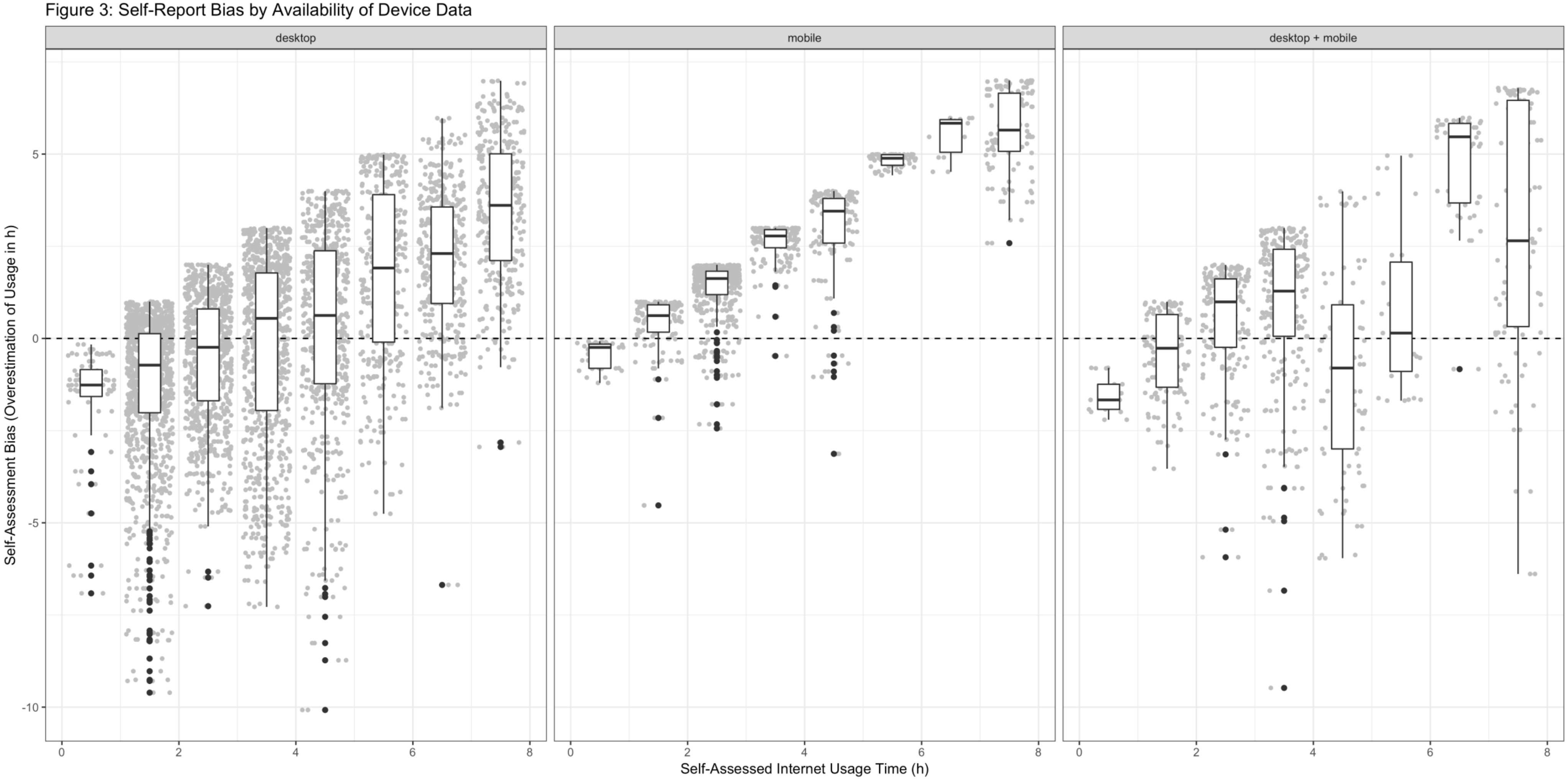


Figure 3: Self-Report Bias by Availability of Device Data

Appendix for Paper:

Two Half-Truths Make a Whole? On Bias in Self-reports and Tracking Data.

This appendix contains some additional information regarding the quantitative analysis present in the paper. Firstly, we provide a series of tables with descriptive information and question wordings, which could not be included in the manuscript for lack of space. The second part contains the code necessary to precisely reproduce our results.

Descriptive Statistics
Metric Variables

Variable	Minimum	Quartile 1	Median	Quartile 3	Maximum
Age (years)	15	33	47	55	74
Perceived Polarization (0-100)	0	16	38	64	100
Own Position (0-100)	0	30	59.26	88.64	100
Internet Usage: Tracking (daily, hours)	0	0.68	1.89	3.86	14.1

Variable: Internet Usage: Survey (typical hours)

Typical Internet Usage per Day (hours)	Cases
< 1 hour *	8
1-2 hours	78
2-3 hours	80
3-4 hours	69
4-5 hours	38
5-6 hours	25
6 hours and above	27
Basically online all the time*	29
Total	354

* This scale was transformed into a quasi-metric variable, for details see the paper. Introducing it into the bias calculation can result in censoring (data hitting an enforced limit), we identified those cases and modeled them in the BRMS calls (see code).

Variable: Sex
Female: 165, Male: 189

Variable: Education

Education Level	Description	Cases
Pupil	Pupil in regular school (up to age 18)	3
No Education Level	Did not finish any secondary school (elementary school is mandatory and can be assumed)	2
Level 1 “Hauptschulabschluss”	Lowest level of secondary education, 5 years	75
Level 2 “Realschulabschluss”	Intermediate level of secondary education, 6 years	134
Level 3 “Abitur” (Hochschulreife)	Highest level of secondary education, 8-9 years, prerequisite to university attendance	43
Level 4 University	Any university degree, i.e. BA, MA or PhD	97
Total		354

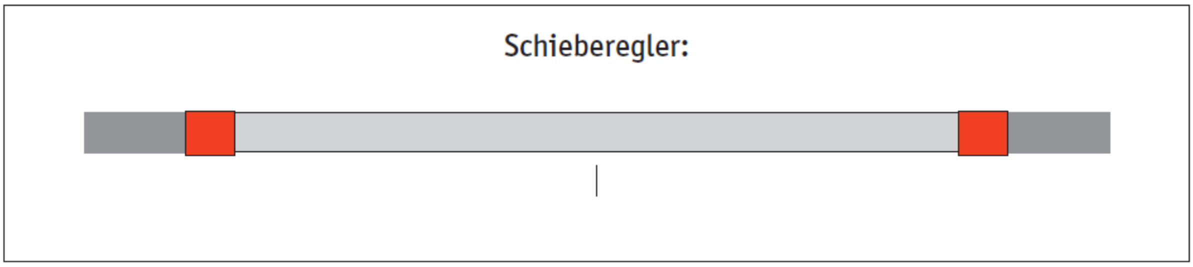
Variable: Household Income

Household Income (yearly, in Euro)	Cases
< 1000 €	35
1000 < 2000 €	97
2000 < 3000 €	88
3000 < 4000 €	59
4000 < 5000 €	31
5000 and more	14
No Answer	30
Total	354

Variable: Device

Data Presence from Device	Cases
No tracking data	73
Desktop devices (PC, macOS)	243
Mobile devices (android, iOS)	25
Both	13
Total	354

Wordings (German with translations where relevant):
Sex*: Sind Sie …* (Option: Männlich, Weiblich, pick one)
Age*: Geben Sie bitte Ihr Alter an:* ___________ *Jahre*
Education: *Was ist Ihr höchster Bildungsabschluss?* (battery, pick one)
Internet Usage: Survey: *An den Tagen, an denen Sie das Internet nutzen, wie lange sind Sie im Durchschnitt ungefähr online? Mit online sein meinen wir das Internet aktiv nutzen, z. B. surfen, etwas suchen, auf Facebook oder in einem anderen sozialen Netzwerk unterwegs sein.*
Translation: On those days when you use the Internet – how long on average are you online? Being online meaning actively using the net, for example for surfing, searching, using Facebook or other social networks.
Income: *Wie hoch ist etwa das monatliche Netto-Einkommen, dass Sie alle zusammen in Ihrem Haushalt haben, nach Abzug der Steuern und Sozialversicherungen?*
Translation: What is the monthly income of your household, after deducting taxes and social welfare contributions?
Note on perceived polarization and own position:
Perceived polarization and own position were measured using a sliding scale. First, participants were to drag one knob to the edge, while the other knob would symmetrically move into the same position, reflecting the perceived distance between opinions. They then subsequently placed their own position regarding the issue on the previously defined spectrum.



Perceived Polarization: *Sicher existieren zum Thema 1/2 unterschiedliche Meinungen. Wenn Sie einmal an die am weitesten voneinander entfernten Meinungen denken: Wie weit liegen diese Ihrer Meinung nach auseinander? Bitte visualisieren Sie dies mit Hilfe des Schiebereglers. Der Mittelpunkt der Linie beschreibt zum Beispiel ein Thema, bei dem alle Beteiligten derselben Meinung sind. Im Gegensatz dazu stehen die beiden Punkte am Ende der Linie für ein Thema mit sehr starken Meinungsunterschieden.*
Translation: Surely there are differing opinions regarding issue 1/2 (insert current issue). If you think of those opinions that are furthest away from each other: How distant are those? Please visualize this using the slider. The center of the slider represents an issue for which everyone has the same opinion. The endpoints, on the other hand, represent a state where people have extremely differing opinions.
Own Position*: Bitte stellen Sie sich erneut die beiden existierenden Extremmeinungen zum Thema 1/2 vor, die Sie soeben mit Hilfe des Schiebereglers eingetragen haben. Angenommen Sie beteiligen sich nun an der Diskussion über das Thema 1/2. Wo würden Sie Ihre eigene Meinung innerhalb der beiden festgesetzten Extrempositionen auf der Linie verorten?*
Translation: Please remember the two extreme opinions from issue 1/2, which you just visualized via the sliding scale. Assuming you were to partake in the discussion about this issue. Where would you place your opinion on this scale?

Code

The following code can be used to replicate the statistical analysis and re-create the graphs and tables in the manuscript. Comments are included where we felt necessary. A data file containing the study's data is necessary to run it – contact the authors for access.
A copy of this code was also placed in a publicly available repository on github, see https://github.com/trifle/two-half-truths

```
# Replication Code for "Two Half-Truths Make a Whole?"
# Author: Pascal Jürgens (p@atlasnovus.net)

# The following code enables replication of statistical analyses and
graphics contained in the paper.
# It requires a data file that can be obtained from the authors, see
the corresponding section in the text.

# IMPORTANT NOTE ON REPLICABILITY
# The bias simulation step is only replicable if the seed value was
set properly.
# Special care must be taken to include that step when running parts
of the script selectively.
# Software versions at publication time:
# R version 3.5.1 (2018-07-02) -- "Feather Spray", Platform: x86_64-
apple-darwin15.6.0 (64-bit)
# brms 2.6.0
# cowplot 0.9.3
# dplyr 0.7.8
# ggplot2 3.1.0
# sjPlot 2.6.1
# rstan 2.18.2

###############
# Prerequisites:
# Libraries: ggplot2, brms, sjPlot, cowplot, dplyr
# Data file: replication.data.csv
###############

require(brms)
require(ggplot2)
require(cowplot)
require(sjPlot)
require(dplyr)

###############
# Setup
###############

# Set seed for the PRNG so that results are replicable.
# Given this seed, numerical results will match the published
manuscript
# as long as there are no changes to the employed libraries.
```

```
# Note that brms/rstan models don't use R's seed and still need to
have the value as an explicit argument.

set.seed(112358)

# Read data
source("replication.data.Rdmpd")

###############
# Data Overview
###############

###############
# Figures
###############

###############
# Figure 2: Self-Report Bias of Internet usage
# Note: This section uses the library "cowplot" to combine two plots
###############

# Output to file
png("f2_bias_daily_and_avg.png", width=5000, height=2500, res=300)

average_bias <- ggplot(
    # Data
    replication.data,
    aes(x=online_time_survey, y=online_time_tracking_average,
group=online_time_survey)) +
    geom_boxplot() +
    # Fix scales across both plots
    scale_y_continuous(limits = c(0, 15)) +
    # Plot dashed line for zero bias
    geom_abline(intercept = 0, slope = 1, linetype="dashed") +
    theme_minimal() +
    # Hide label for this x axis, having one on the other plot is
enough
    xlab("") +
    ylab("Daily Usage from Tracking (hours)") +
    ggtitle("Self-Report Bias (Internet, Average Day)") +
    theme(axis.line.x=element_blank(),axis.line.y=element_blank())

daily_bias <- ggplot(
    # Data
    replication.data,
    aes(x=online_time_survey, y=online_time_tracking_daily,
group=online_time_survey)) +
    geom_boxplot() +
    # Fix scales across both plots
    scale_y_continuous(limits = c(0, 15)) +
```

```
    # Plot dashed line for zero bias
    geom_abline(intercept = 0, slope = 1, linetype="dashed") +
    theme_minimal() +
    # Hide label for x and y axis
    xlab("") +
    ylab("") +
    ggtitle("Self-Report Bias (Internet, All Days)") +
    theme(axis.line.x=element_blank(),axis.line.y=element_blank())

# Combine both plots side by side, adding a common x axis label and
title
combined_plot <- cowplot::plot_grid(average_bias, daily_bias, labels
= "AUTO")
title <- ggdraw() + draw_label("Figure 2: Self-Report Bias of
Internet Usage", fontface = 'bold')
combined_plot <- add_sub(combined_plot, "Self-Assessed Average Daily
Internet Usage (hours)")
cowplot::plot_grid(title, combined_plot, ncol = 1, rel_heights =
c(0.1, 1))
dev.off()


###############
# Figure 3: Self-Report Bias by Availability of Device Data
###############

# Output to file
png("f3_bias_by_device.png", width=5000, height=2500, res=300)
ggplot(
    # Data to use. Omit group with no data.
    subset(replication.data, device_factor!='none'),
    aes(x=online_time_survey, y=bias_daily,
group=online_time_survey)) +
    # Use jitter to make individual data points discernible+
    geom_jitter(color="grey", size=1) +
    # Show dotted line at zero bias
    geom_abline(linetype="dashed", slope=0) +
    facet_grid(~device_factor) +
    geom_boxplot() +
    theme_bw() +
    theme(axis.line.x=element_blank(),axis.line.y=element_blank()) +
    scale_fill_grey(start = 0.7, end = 1) +
    ggtitle("Figure 3: Self-Report Bias by Availability of Device
Data") +
    ylab("Self-Assessment Bias (Overestimation of Usage in h)") +
    xlab("Self-Assessed Internet Usage Time (h)")
dev.off()

###############
# Statistical Models
```

```
###############

###############
# Model 1 (Table 1): Predictors of Bias in Self-Reports (Bayesian
Mixed Model)
###############

# Data includes up to two issues per day. Since this model does not
include
# the DV on a per-issue basis, we need to pool the data into daily
data points

# Create new variable for day-id pairs
replication.data$day_id <- apply(replication.data[c("id", "day")],
1, paste, collapse="_")

# Group by day_id
replication.data.grouped <- group_by(replication.data, day_id)

# Create new pooled data frame
replication.data.pooled <- summarize(
  replication.data.grouped,
  id=first(id),
  day=first(day),
  age=first(age),
  sex=first(sex),
  education=first(education),
  income=first(income),
  online_time_survey=first(online_time_survey),
  device_factor=first(device_factor),
  bias_daily=first(bias_daily),
  bias_censored=first(bias_censored)
)

bias_predictor_model <- brm(
    # DV is the daily bias, which can be censored
    # (has a capped maximum). We therefore use a binary variable to
flag censored
    # values for the model.

    bias_daily|cens(bias_censored)

    # Predictors

    ~ device_factor +
    scale(age) +
    sex +
    scale(education) +
    scale(income) +
    scale(online_time_survey) +
```

```
    # Levels, fixed slope, random intercept
    (1|day) +
    (1|id)

    # parameters
    # Note that the model run will take some time due to
    # increased number of chains and iterations.

    ,data=replication.data.pooled
    ,chains=8
    ,cores=8
    ,iter=8000
    ,seed=112358
)

###############
# Example Models (Table 2): Bayesian Models for Bias Simulation
###############

###############
# Authentic Models
# Note that the first model (without device factor) is not included
in the paper.
###############

two.mlm <- brm(
  p.polarization
    ~ scale(age) +
    sex +
    scale(education) +
    scale(income) +
    scale(own_position) +
    scale(online_time_survey) +
    scale(online_time_tracking_daily) +
    (1|day) +
    (1|issue)
    ,data=replication.data
    ,cores=8
    ,seed=112358
)

three.mlm <- brm(
  p.polarization
    ~ scale(age) +
    sex +
    scale(education) +
    scale(income) +
    scale(own_position) +
    scale(online_time_survey) +
```

```
    scale(online_time_tracking_daily) +
    device_factor +
    (1|day) +
    (1|issue)
    ,data=replication.data
    ,cores=8
    ,seed=112358
)

###############
# Simulated Models
###############

###############
# Generate Bias
###############

# Set seed again to enable replicability of this part even when run
in isolation.
set.seed(112358)

# Make copy of the data
d.sim <- replication.data

# Calculate means to have a rough indicator of the impact of our
simulated bias
pre.mean.mobile <-
mean(d.sim$online_time_survey[d.sim$device_factor=='mobile'])
pre.mean.other <-
mean(d.sim$online_time_survey[d.sim$device_factor!='mobile'])
pre.mean <- mean(d.sim$online_time_survey)

# Get number of cases for each device group
n.mobile <-
length(d.sim$online_time_survey[d.sim$device_factor=='mobile'])
n.other <-
length(d.sim$online_time_survey[d.sim$device_factor!='mobile'])

# Create normally distributed error
# Add higher bias to mobile, and balance this out by
# subtracting the appropriate counterpart from the other devices.
mobile.add <- rnorm(n.mobile, 15, 1)
other.add <- rnorm(n.other, 10, 1)
other.factor <- sum(other.add) / sum(mobile.add)
other.add <- other.add / other.factor

# Add bias to self-assessments
d.sim$online_time_survey[d.sim$device_factor=='mobile'] <-
d.sim$online_time_survey[d.sim$device_factor=='mobile'] + mobile.add
```

```
d.sim$online_time_survey[d.sim$device_factor!='mobile'] <-
d.sim$online_time_survey[d.sim$device_factor!='mobile'] - other.add

# Calculate impact by comparing to previous means
post.mean.mobile <-
mean(d.sim$online_time_survey[d.sim$device_factor=='mobile'])
post.mean.other <-
mean(d.sim$online_time_survey[d.sim$device_factor!='mobile'])
post.mean <- mean(d.sim$online_time_survey)

###############
# Models with Biased Data
###############

two.sim <- brm(p.polarization
    ~ scale(age) +
    sex +
    scale(education) +
    scale(income) +
    scale(own_position) +
    scale(online_time_survey) +
    scale(online_time_tracking_daily) +
    (1|day) +
    (1|issue)
    ,data=d.sim
    ,cores=8
    ,seed=112358
)
```

```
three.sim <- brm(p.polarization
    ~ scale(age) +
    sex +
    scale(education) +
    scale(income) +
    scale(own_position) +
    scale(online_time_survey) +
    scale(online_time_tracking_daily) +
    device_factor +
    (1|day) +
    (1|issue)
    ,data=d.sim
    ,cores=8
    ,seed=112358
)

# Create table with all models in one.
# The function tab_model is from sjPlot, tweak parameters to show
desired data.
# Note: To get additional model diagnostics, such as R hat, use
summary(model)

# All models
tab_model(two.sim, three.sim, two.mlm, three.mlm, emph.p = T,
show.hdi50 = F)

# Table2 from paper
tab_model(three.mlm, two.sim, three.sim, emph.p = T, show.hdi50 = F)

# Model analysis. Provided for completeness, since the comparison is
not meaningful
# across simulation and authentic data.
waic.three.mlm <- waic(three.mlm)
waic.two.sim <- waic(two.sim)
waic.three.sim <- waic(three.sim)

compare_ic(waic.three.mlm, waic.two.sim, waic.three.sim)
```