The London School of Economics and Political Science

Department of Methodology

Challenging the Gold Standard: A Methodological Study of the Quality and Errors of Web Tracking Data

Oriol J. Bosch

A thesis submitted to the Department of Methodology of the London School of Economics and Political Science for the degree of Doctor of Philosophy

London
September 2023

Declaration

I certify that the thesis I have presented for examination for the PhD degree of the London School of Economics and Political Science is solely my own work other than where I have clearly indicated that it is the work of others (in which case the extent of any work carried out jointly by me and any other person is clearly identified in it).

I declare that my thesis consists of approximately **65,258** words.

Statement of co-authored work

I confirm that Chapter 3 was jointly co-authored with Melanie Revilla, and I contributed **80**% of this work; and that Chapter 4 was jointly co-authored with Patrick Sturgis, Jouni Kuha, and Melanie Revilla, and I contributed **70**% of this work.

Oriol Bosch Jover

As the candidate's primary supervisor I hereby confirm that the extent of the candidate's contribution to the joint-authored papers was as indicated above.

Prof. Patrick Sturgis

I belong to a culture. A culture of invisible people.

Let this work be a reminder: I was here. And so were they.

λ

Felt good, though, just being what I was

TOM SPANBAUER, The Man Who Fell in Love with the Moon

Acknowledgements

I am a measurement guy. It makes sense because I tend to obsess over things. That is why I am an academic. Life is terribly confusing; I only want a bit of clarity. Being at the end of this journey, I wonder: what is a PhD?

When I first embarked on this journey, I saw it as a chance to create a well-structured body of research, one that would showcase my intellectual capabilities. I aimed to compile a set of research papers that could help fellow social scientists better understand the digital world.

However, looking back, I've come to see that a PhD is much more than an academic pursuit. It's an act of love, not my own, but the love and support I've received from those who have stood by me, encouraged me, challenged me, and cared for me. This PhD has been made possible by the people in my life, and they truly deserve my heartfelt gratitude and recognition.

Patrick and Jouni, I want to express my utmost gratitude to both of you for being the best supervisors I could have hoped for. Right from the beginning, you understood me and provided the space and freedom I needed to excel. Your intellectual challenges have taught me what it means to be a good and rigorous academic. Patrick, in particular, I am deeply grateful for the numerous opportunities you've afforded me. Your support and uplifting words of encouragement have meant the world to me. Your ability to grasp the bigger picture has been a guiding light throughout this journey, and I have always felt exceptionally supported under your mentorship. Jouni, I am sincerely thankful for sharing your genius with me. Every meeting with you has left me in awe of your profound brilliance. If this PhD has any trace of rigor and sophistication, it is undeniably a result of your invaluable influence and guidance.

Melanie, I am writing these acknowledgement mainly thanks to you. I could never have foreseen that enrolling in your Survey Design class would have such a profound and lasting impact on my life. For reasons unknown to me, you recognized something within me and chose to support me in a way that no one else had done before. Your influence has completely transformed my life, and I'm indebted to you

beyond words. Working alongside you has ignited a deep passion for research within me. Your unwavering dedication, critical thinking, and sheer brilliance are what set you apart as the best in your field. I take immense pride in being not just your student and coauthor, but also your friend. You've tirelessly trashed every one of my papers, filling them with comments and suggestions, and this has truly made me into a serious academic. You've granted me the freedom to explore, provided unparalleled access to resources, and presented me with countless opportunities, all of which have placed me in the privileged position I find myself in today. Yet, beyond these professional aspects, you've welcomed me into your life, introduced me to your family, and offered invaluable life advice. It's because of these gestures that I consider you my friend. Your relentless critical intellect is matched only by your extraordinary generosity.

I am extremely grateful to the Department of Methodology. Finishing this PhD has broken my heart; there is no other place I would rather be. I am especially thankful to my fellow PhD students, past and present: Nancy, Yuanmo, Qi, Ross, Midanna, Ronni, Amal, Rosie, Poorvie, Thiago, and all others. We're not just colleagues; we are friends. Coming to the office every day has been a joy just because of you. I will forever miss our lunches, our little chats, the gossiping, the afterwork drinks. I am extremely proud of each one of you; I am in awe of how much you all have grown over the years. Always remember this: you are incredible, and we are not just a department, but a family that takes care of each other. I will deeply miss all of you, you have made me incredibly happy. You better keep inviting me to the pub crawls, or there will be consequences. I am also incredibly thankful to Camilya. I do not think there is anyone more competent than you. Thanks for being so nice with us, your positive energy has kept all of us warm. The department is lucky to have you. There are many more people that I would like to thank in the department: Jon, Milena, Sian, Marion, Anne, Flora, and everyone else; you are all deeply kind-hearted people.

Family is important, and I consider myself truly fortunate to have been blessed with not one, but three families. As everyone else, I have my biological family: Mama, Papa, Pol, Guillem, Palmi Palmi, avi Vicens, avia Margarita, and Vicky. We are far from conventional, but there's something beautiful in our unorthodoxy. I am who I am because of you. Mama, I owe you a debt of gratitude for raising us, even when the journey wasn't always easy. Thank you for embracing your quirks and eccentricities, for filling our childhood with countless joyful memories, and for being my best friend when I was a kid. Thanks for educating us in your own hippie way, teaching us that we have to be kind, respect the world and everything that lives on it, and to recognize a world of opportunities beyond our small town. You always told us that we should leave the country, explore what the world had to offer. I listened, and here I am. Papa, thanks for inspiring me. Thanks for those long conversation during car rides home. You instilled in me the unshakable belief that no future is possible without hard work. Thanks for cultivating our minds, for the movies and books, the music, and the talks about politics. You implanted your love for culture in my mind and my heart, and that made all the difference. Pol and Guillem, thanks for being my brothers. I would have lost my mind if I had had to navigate everything alone. There was joy in all the craziness. To my grandparents: Palmi Palmi, i avi Vicens (allà on siguis), gràcies per la vostra generositat infinita. Gràcies per donar-me una casa, per estar allà quan feia falta, per cuidar-me. Gràcies per donar-me amor incondicional. Gràcies, avi, per fer-me de taxi, recader i amic. Et trobo a faltar cada dia. Gràcies Palmi Palmi per ser el raig de bondat que ets, casa és on siguis tu.

I am fortunate enough to have a second family: Joan Carles and Eugeni. You found me in the lowest point of my life and provided me with unwavering support when I needed it most. You are the kindest, most generous people I know. Our connection is bound solely by love and a shared determination to create a family, and therein lies our uniqueness. Thank you for always lending me your ears, for aiding in my personal growth, for the boundless encouragement you've showered upon me throughout the years, for the delicious and ginormous meals you cook for me. I genuinely doubt I would have believed in myself if you hadn't believed in me first. No hi ha forma en què us pugui expressar com en sou d'importants per mi, gràcies per tot. Ah, y Eugeni, recuerda, te guardé un higo, pero como no te vi, me lo comí.

Lastly, my chosen family here in London: Nancy (again!), Edu, MJ (and Norman and Clo < 3), Camilo, Gab, Parker, and Roy. You make this damp, dark island feel like home. It's because of all of you that these past four years have been the happiest of my life. You are an incredibly brilliant, fun, and inspiring group of people. Nancy, thanks for being my work wife. I can't even fathom how different this PhD journey would have been without you by my side. You're unequivocally the coolest person I've ever had the privilege to meet, and words fall short to describe how much I admire you. Our friendship has made me a better person. Thanks for your big heart, and for all the fun we have had. Edu, you are the (third) brother I never knew I needed. You are the first person I go to when I need to talk. You get me like no one else does. Thanks for always being there, for being so supportive, for always having words of encouragement. Knowing that I will always have you by my side makes the future look way brighter. To all of you, thanks from the bottom of my heart. Queer joy, as we've experienced it together, truly is the most profound form of joy.

My life is graced by the presence of many incredible individuals, both in London, back in my hometown, and scattered across the globe. There are so many of you, and I hope you can pardon me if I don't mention everyone by name. To Marc and Laura, thanks for sharing a home with me for almost the entirety of this PhD. We navigated some wild adventures with landlords and housemates, and there's no one who has witnessed more of my tears over these crazy four years. Thanks for being my rock and supporting me all the way. You better invite me to your wedding. Alexis, thank you for injecting happiness and much-needed stability into the final year of my PhD. Life is uncertain, but what we have shared will always be beautiful. Ah, and thanks for that pasta with mushrooms. Asensio, thanks for 14 years of friendship. I cannot imagine a world without you and your absurdity. There is no one that sees me with such kind eyes, you really make me want to be as good as you think I am. Thanks for all the trips, dinners, conversations, and weird tweets. T'estimo molt, amic. Oriol, my Moroncho, thank you for consistently being there. You once said that being an academic was cooler than working in political communication and look where I am now. Thanks for always being there to chat about anything, for all the support you have given me over the years. It has been 5 years since I left Barcelona, but you still make it feel like home. Alba, you inspired me to want to do great things with my life. I do not know if I would have survived my first year in the UK if it was not for you. I miss you so much. How I wish you lived in London.

I love you all.

Abstract

Measuring what people consume and do online is crucial across the social sciences. In the last few years, web tracking data has gained popularity, being considered by most as the gold standard for measuring online behaviours. This thesis studies whether this prevailing notion holds true. Specifically, through a combination of traditional survey and computational methods, I assess the quality of web tracking data, its associated errors, and the consequences of these. The thesis is comprised of three distinct papers. In the first paper, inspired by the Total Survey Error, I present a Total Error framework for digital traces collected with Meters (TEM). The TEM framework describes the data generation and the analysis process for web tracking data and documents the sources of bias and variance that may arise in each step of this process. The framework suggests that metered data might indeed be affected by the error sources identified in our framework and, to some extent, biased. The second paper adopts an empirical approach to address a key error identified in the TEM framework: researchers' failure to capture data from all the devices and browsers that individuals utilize to go online. The paper shows that tracking undercoverage is highly prevalent when using commercial panels. Additionally, through a simulation study, it demonstrates that web tracking estimates, both univariate and multivariate, are often substantially biased due to tracking undercoverage. The third paper explores the validity and reliability of web tracking data when used to measure media exposure. Merging traditional psychometric and computational techniques, I conduct a multiverse analysis to assess the predictive validity and true-score reliability of thousands of web tracking measures of media exposure. The findings show that web tracking measures have an overall low validity but remarkably high reliability. Additionally, results suggest that the design decisions made by researchers when designing web tracking measurements can have a substantial impact on their measurement properties. Collectively, this thesis challenges the prevailing belief in web tracking data as the gold standard to measure online behaviours. Methodologically, it illustrates how computational methods can be used to adapt survey methodology techniques to assess the quality of digital trace data.

Contents

Page

Chapter 5

List of Figures

List of Tables

Chapter 1

Introduction

The landscape of data collection in the social sciences has undergone a transformative shift in recent years. Traditionally, researchers heavily relied on probability-based face-to-face surveys to understand human behaviours and attitudes. However, the advent of the Internet has ushered in a new era of data abundancy. In this era, when individuals engage with online platforms and digital technologies, they leave behind digital traces. These digital systems include telecommunication networks, websites, social media platforms, mobile apps, business transaction systems, sensors built in wearable devices, and digital devices (Stier et al., 2019). Through innovative data collection methods, such as data donations or web trackers, researchers can collect these traces for scientific research. The resulting data is broadly conceptualised as *digital trace data*.

From a substantive standpoint, this development is not trivial. During the last decade the time that people spend online has doubled. The ways in which people connect to the Internet, and how they use it, have also been altered, with mobile devices and social media platforms becoming ubiquitous and gaining more and more importance into people's lives. This change of paradigm has had, most likely, critical consequences on society, and how humans think, feel, and behave. Understanding the effects that the Internet has on society, and the extent to which these might be negative, is an important puzzle for social science and policy. This understanding, nonetheless, can only be produced if scientists can accurately measure people's behaviours on the Internet. The availability and use of digital trace data, hence, is key in the quest towards a better understanding of the role of the Internet and digital technologies on society. Indeed, to some extent, digital trace data has contributed to the emergence of computational social sciences (Edelmann et al., 2020;

Lazer et al., 2009).

Methodologically, the availability of digital trace data has heralded what some have termed a measurement revolution in the social sciences (Golder and Macy, 2014; Lazer et al., 2009). This enthusiasm is well-founded, given the immense potential that digital traces hold for social research. Digital trace data enables the observation of human behaviour on a scale, granularity, and depth previously inconceivable. These traces are produced in real-time, granting researchers the ability to explore fluctuations in behaviours rather than relying solely on discrete time-point observations. Researchers can now investigate trends in real-time, in topics such as mobility (Elevelt et al., 2019) or migration (Zagheni and Weber, 2012; Zagheni et al., 2014). These traces, furthermore, allow to study human behaviours within the content of critical events such as wars (Leasure et al., 2023) or natural disasters (Sutton et al., 2013).

These data are not only very granular, but also offer unique, unsolicited insights into the ways in which people behave and express themselves (Cesare et al., 2018). Traditional data collection methods often grapple with issues such as social desirability bias or memory errors when measuring both attitudes and behaviours. In contrast, digital traces offer direct access to the behaviours of individuals, and their public and sometimes private communications. This non-reactive nature of digital trace data might help circumvent some of the errors of self-reports. For instance, if behaviours can be directly observed, measures cannot be affected by recall bias. Additionally, observed behaviours might potentially reveal controversial or deviant behaviours, which individuals might conceal in other forms of data collection (Krumpal, 2011).

The enthusiasm surrounding digital trace data has led to a notable trend among researchers: a shift away from traditional data sources such as surveys, in favour of digital traces (Schoen et al., 2013). Paradoxically, however, this transition from surveys to digital trace data has often been accompanied by a somewhat naive approach to measurement theory (Jungherr, 2019). The allure of digital trace data has, in many cases, overshadowed the critical scrutiny typically applied to surveys. While it is widely acknowledged that digital trace data often lacks representativeness, it has been somewhat uncritically regarded as a source that inherently reveals the "objective" truth behind phenomena of interest. This perspective contradicts the well-established measurement theories in social sciences and statistics. In these fields, a nuanced understanding prevails (Jungherr, 2019): phenomena of interest are translated into theoretical concepts, which, in turn, are transformed into measures. These measures undergo rigorous assessment of their measurement quality, and statistical procedures are developed to draw meaningful inferences from the available data.

In stark contrast, digital trace data has been somewhat unwarrantedly placed on a pedestal as an absolute gold standard. This perception is somewhat surprising when considering that digital trace data arises as a byproduct of activities not originally intended for research purposes (Ang et al., 2013). Digital traces are shaped by the design of the digital services and devices that produce the data, users' motivations when using those, cultural usage norms, and the technologies used to capture these traces (Jungherr et al., 2016). Hence, these traces cannot be taken at face value. Instead, it is imperative to understand how these factors might deviate the variables created with these traces from the constructs of interest. To date, nonetheless, there is no theoretical or empirical evidence to assume that these byproducts are error-free, or inherently superior in quality than survey self-reports. Hence, researchers should focus not only on the opportunities that digital traces bring to the social science, but instead on testing the sources of biases and variance that might affect digital trace data. Similarly important is the development of guidelines, research designs, and statistical techniques specifically developed for the correct use of digital trace data in the social sciences (Salganik, 2019). As Jungherr (2019) succinctly argues, "the development of a sophisticated measurement theory is a precondition for digital trace data to be meaningfully integrated into the social sciences."

To move in this direction, it is crucial to recognize that the term "digital trace data" encompasses a broad spectrum of approaches, each with its unique advantages and limitations. These distinctions can arise from factors such as the nature of the traces collected (e.g., Tweets vs. URLs), the data collection methods employed (e.g., web trackers vs. APIs), or the level of engagement of the individuals generating the data (ranging from passive collection to complete control over the process). While it is possible to discuss digital trace data at a macro-level, it is imperative to conduct micro-level methodological investigations into each type of digital trace data source independently. The sources of error affecting digital trace data are likely to vary based on the specific traces of interest, the methods of collection, and the extent of user involvement. Therefore, while overarching discussions about digital trace data are valuable, the bulk of methodological research should be dedicated to establishing a comprehensive understanding of the challenges unique to each type of digital trace data. By doing so, researchers interested in utilizing the various digital trace data sources can benefit from specific guidelines and practical procedures tailored to their chosen data type.

A key type of digital trace data is **web tracking data**. Web tracking data stands as one of the most prevalent sources in the realm of digital trace data for measuring individual-level online behaviours. As the name suggests, this approach to gathering digital traces hinges on the utilization of web tracking technologies (Christner et al., 2021). These technologies, known as *meters* (Revilla et al., 2021), encompass a diverse array of solutions that participants can install or configure voluntarily onto their browsing devices. Once installed, these meters enable the tracking of various traces left by participants during their online interactions, including visited URLs, accessed apps, search engine queries, and the content participants have encountered (e.g., HTML information).

Historically, web tracking has been upheld as the de facto gold standard for measuring online behaviours, especially those involving media exposure. Some research has even used it as the benchmark to estimate the accuracy of survey self-reports (Araujo et al., 2017; Scharkow, 2016), with certain authors advocating for the replacement of survey self-reports with digital traces (Konitzer et al., 2021). Because of this, web tracking data has been widely used in the literature (see section 6 of the conceptual overview for more). Leveraging the data collected through meters, researchers have investigated significant topics, predominantly in the media and communication field. For example, existing studies have quantified the prevalence of dubious media exposure during elections (Guess et al., 2020), the overlap in political media diets between partisans (Guess et al., 2021), or the extent to which online news environments are segmented by age groups (Mangold et al., 2021). Another salient area of inquiry has been the degree to which social media and other sites serve as intermediaries for online media exposure (Cardenal et al., 2019; Jürgens and Stark, 2022; Stier et al., 2021; Scharkow et al., 2020).

However, it is imperative to underscore that the assertion of web tracking data enjoying gold standard status lacks substantial support. In reality, certain studies have already sounded the alarm regarding the potential susceptibility of this data to errors (Jürgens et al., 2019; Revilla et al., 2017). A recent report from the Pew Research Center (2020), for instance, concluded that "[web tracking data] does not, at present, seem well suited for high-level estimates of news consumption." Thus, it becomes apparent that web tracking data shares more characteristics with surveys than early enthusiasts had proclaimed, rendering it subject to many of the same limitations (Jungherr, 2019). Unlike surveys, however, systematic information about potential errors associated with web tracking data collection and their consequences remains scarce.

Why is this crucial? While it is widely recognized that surveys are susceptible to a plethora of errors (Groves et al., 2010), decades of research have equipped scholars and practitioners with a wealth of evidence to comprehend the limits of surveys and establish best practices in designing, collecting, and analysing survey data. For instance, researchers can employ frameworks like Total Survey Error (TSE) (Groves et al., 2009) to pinpoint and estimate potential errors, assess their impact on estimates, and develop strategies to mitigate them. Additionally, years of empirical research in survey methodology have fostered an in-depth understanding of the measurement quality of various survey questions and the influence of different design decisions on the final quality of survey measurements (DeCastellarnau, 2017).

In stark contrast, our understanding of the constraints of web tracking data is nearly non-existent. This gap in research predominantly stems from the presumption that web tracking data is inherently unbiased, or that its potential errors are inconsequential. Recognizing that a data source may indeed be susceptible to systematic and random errors opens avenues for a more comprehensive understanding of these errors and strategies to mitigate them. Acknowledging that a data source is not infallible does not inherently imply its rejection; instead, it suggests that data collection, processing, and analysis should proceed cautiously, guided by informed decisions.

In this thesis, I assess the quality of web tracking data, its associated errors, and the consequences of these. The papers in my thesis contribute to the decades of methodological literature trying to improve the way in which social scientists collect and use data. They do so by adapting this vast knowledge to digital trace data, through the use of computational methods. This thesis, hence, serves as a guide for academics and practitioners to better understand the quality and errors of web tracking data, showcasing best practices that anyone can follow when collecting and using this type of data.

The papers comprising this thesis

The first paper of my dissertation is entitled "When survey science met web tracking: Presenting an error framework for metered data" and can be found in Chapter 3. Inspired by the Total Survey Error, in this paper I present a Total Error framework for digital traces collected with Meters (TEM). The TEM framework (1) describes the data generation and the analysis process for metered data and (2) documents the sources of bias and variance that may arise in each step of this process. Using a case study, the paper also shows how the TEM can be applied in real life to identify, quantify, and reduce metered data errors. The results of this paper suggest that web tracking data might indeed be affected by the error sources identified in the framework and, to some extent, bias. Hence, caution should be taken when using metered data for inferential statistic. In the context of this thesis, the framework works as the basis for the other papers. By clearly showing how web tracking data is collected and analysed, and identifying the errors of web tracking data, it allows to develop approaches to quantify those errors, and strategies to minimise them. The paper, co-authored with Dr Melanie Revilla, has already been published in the Journal of the Royal Statistical Society: Series A.

The second paper, "Uncovering digital trace data biases: tracking undercoverage in web tracking data", can be found in Chapter 4. The second paper adopts an empirical approach to address tracking undercoverage, a key error identified in the TEM framework: researchers' failure to capture data from all the devices and browsers that individuals utilize to go online. The paper shows that tracking undercoverage is highly prevalent when using commercial panels. Additionally, through a simulation study, it demonstrates that web tracking estimates, both univariate and multivariate, are often substantially biased due to tracking undercoverage. This represent the first empirical evidence demonstrating that web tracking data is, effectively, biased. Methodologically, the paper showcases how survey questions can be used as auxiliary information to identify errors in web tracking data. Additionally, it shows how the granularity of web tracking data can be leveraged, in combination with simulation techniques, to estimate the size of the web tracking errors. The paper, co-authored with Professor Patrick Sturgis, Professor Jouni Kuha, and Dr Melanie Revilla, has already been submitted to the journal Communication Methods and Measures.

The third and last paper, entitled "Validity and Reliability of Digital Trace Data in Media Exposure Measures: A Multiverse of Measurements Analysis", is found in Chapter 5. The last paper explores the validity and reliability of web tracking data when used to measure media exposure. Merging traditional psychometric and computational techniques, I conduct a multiverse analysis to assess the predictive validity and true-score reliability of thousands of web tracking measures of media exposure. The findings show that web tracking measures have an overall low validity but remarkably high reliability. Additionally, results suggest that the design decisions made by researchers when designing web tracking measurements can have a substantial impact on their measurement properties. In terms of methods, this paper demonstrates that the granularity of web tracking data can be used to explore the measurement properties of the entire multiverse of measurements, instead of a few ones, which is the norm in survey research. It also showcases how tools such as Random Forests algorithms can be used to find patterns within the multiverse, helping make sense of the high dimensionality of the results obtained through the multiverse approach. The paper, single authored, will soon be submitted to the journal Political Analysis.

The contributions of this thesis

The contributions of this thesis are varied. First, this thesis challenges the prevailing belief in web tracking data as the gold standard to measure online behaviours. The thesis comprehensively shows that web tracking data is affected by a plethora of different errors. Not only do I prove this theoretically, but I also provide some of the first evidence, if not the first, that web tracking data is indeed biased. Although the thesis mostly points to problems of web tracking data, it also shows that overall web tracking data has a remarkably high reliability.

Second, this thesis includes a number of methodological advancements. I illustrate how the frameworks and methods drawn from the extensive literature on survey methodology and psychometrics, commonly used to assess survey quality, can be adapted to evaluate web tracking data. To achieve this, the thesis demonstrates that computational methods can be harnessed to aid in this endeavour, taking advantage of the granularity and flexibility of web tracking data. Specifically, Paper 2 showcases how survey questions can be used as auxiliary information to identify errors in web tracking data. Furthermore, it shows how simulation techniques can be used to leverage the granularity of web tracking data to estimate the size of the web tracking errors. Paper 3, additionally, demonstrates that the granularity of web tracking data can be used to explore the measurement properties of the entire multiverse of measurements, instead of a few ones, which is the norm in survey research. It also showcases how tools such as Random Forests algorithms can be used to find patterns within the multiverse, helping make sense of the high dimensionality of the results obtained through the multiverse approach.

Next, this thesis helps advance the study of media exposure. According to my own literature review, most research using web tracking data has focused on media exposure. This field of knowledge, hence, has become heavily dependent on web tracking data. By specifically focusing on media exposure when assessing the quality of web tracking data, this thesis provides some cautionary evidence to communication scholars. Paper 2 shows that many of the most commonly computed statistics in the media exposure literature are significantly biased when using web tracking data, due to tracking undercoverage. Additionally, Paper 3 shows that web tracking measures of media exposure present, overall, the same lack of association with political knowledge that led many researchers to consider that surveys had a worrying lack of predictive validity. All in all, the evidence presented by this thesis should incentivise media scholars to re-assess some of their findings, as well as rethink the way in which they used web tracking data.

Fourth, this thesis has partially led to the creation of the TRI-POL database (Torcal et al., 2023). This represents the *first-of-its-kind* open-access dataset merging cross-national longitudinal survey data with individual-level web tracking information. Thanks to the work done in this PhD, the TRI-POL database is the first to be designed acknowledging the errors of web tracking data, with strategies in place to minimize, quantify and report those errors. Hence, the thesis has contributed to making web tracking data more transparent, as well as accessible to researchers with limited access to resources.

Finally, the findings of these studies have practical implications and can be applied by researchers and practitioners alike. Beyond critiquing the quality of web tracking data, this thesis puts much focus on identifying best practices when collecting and analysis web tracking data. The TEM framework presented in Paper 1 can be used by researchers to improve the way in which they collect and analyse web tracking data, as well as how they report the processes followed and their limitations. Paper 2 presents and approach that any researcher can apply to identify tracking undercoverage and simulate its impact on the statistics of interest. Additionally, it provides much needed recommendations to online fieldwork companies offering web tracking web tracking panels, to improve their practices. Paper 3 helps understand the key design choices that significantly influence the validity and reliability of web tracking measurements. This information can be used by substantive researchers to make informed decisions when translating constructs into web tracking measurements. In addition, the paper presents a multiverse of measurements approach that can be used to better convey the uncertainty of their results produced with web tracking data.

Bibliography

Ang, Chee Siang, Ania Bobrowicz, Diane J. Schiano, and Bonnie Nardi. 2013. “Data in the wild.” *Interactions* 20:39–43.

Araujo, Theo, Anke Wonneberger, Peter Neijens, and Claes de Vreese. 2017. “How Much

Time Do You Spend Online? Understanding and Improving the Accuracy of Self-Reported Measures of Internet Use." *Communication Methods and Measures* 11:173–190.

Cardenal, Ana S., Carlos Aguilar-Paredes, Carol Galais, and Mario Pérez-Montoro. 2019. "Digital Technologies and Selective Exposure: How Choice and Filter Bubbles Shape News Media Exposure." *The International Journal of Press/Politics* 24:465–486.

Cesare, Nina, Hedwig Lee, Tyler McCormick, Emma Spiro, and Emilio Zagheni. 2018. "Promises and pitfalls of using digital traces for demographic research." *Demography* 55:1979–1999.

Christner, Clara, Aleksandra Urman, Silke Adam, and Michaela Maier. 2021. "Automated Tracking Approaches for studying online media use: A critical review and recommendations." *Communication Methods and Measures* 16:79–95.

DeCastellarnau, Anna. 2017. "A classification of response scale characteristics that affect data quality: a literature review." *Quality & Quantity* 52:1523–1559.

Edelmann, Achim, Tom Wolff, Danielle Montagne, and Christopher A. Bail. 2020. "Computational Social Science and Sociology." *Annual Review of Sociology* 46:61–81.

Elevelt, A., W. Bernasco, P. Lugtig, S. Ruiter, and V. Toepoel. 2019. "Where you at? using GPS locations in an electronic time use diary study to derive functional locations." *Social Science Computer Review* p. 089443931987787.

Golder, Scott A. and Michael W. Macy. 2014. "Digital Footprints: Opportunities and challenges for online social research." *Annual Review of Sociology* 40:129–152.

Groves, Robert M., Floyd J. Fowler, Mick P. Couper, James M. Lepkowski, Eleanor Singer, and Roger Tourangeau. 2009. "Survey Methodology, 2nd Edition." *Wiley series in survey methodology* .

Groves, Robert M., Eleanor Singer, James M. Lepkowski, Steven G. Heeringa, and Duane F. Alwin. 2010. "Survey methodology." In *A Telescope on Society: Survey Research and Social Science at the University of Michigan and Beyond*.

Guess, Andrew M., Pablo Barberá, Simon Munzert, and JungHwan Yang. 2021. "The consequences of online partisan media." *Proceedings of the National Academy of Sciences* 118:e2013464118.

Guess, Andrew M., Dominique Lockett, Benjamin Lyons, Jacob M. Montgomery, Brendan Nyhan, and Jason Reifler. 2020. ""Fake news" may have limited effects on political participation beyond increasing beliefs in false claims." *Harvard Kennedy School Misinformation Review* 1.

Jungherr, Andreas. 2019. "Normalizing Digital Trace Data." In *Digital Discussions*.

Jungherr, Andreas, Harald Schoen, and Pascal Jürgens. 2016. "The Mediation of Politics through Twitter: An Analysis of Messages posted during the Campaign for the German Federal Election 2013." *Journal of Computer-Mediated Communication* 21:50–68.

Jürgens, Pascal, Birgit Stark, and Melanie Magin. 2019. "Two Half-Truths Make a Whole? On Bias in Self-Reports and Tracking Data." *Social Science Computer Review* p. 089443931983164.

Jürgens, Pascal and Birgit Stark. 2022. "Mapping exposure diversity: The divergent effects of algorithmic curation on news consumption." *Journal of Communication* 72:322–344.

Konitzer, Tobias, Jennifer Allen, Stephanie Eckman, Baird Howland, Markus Mobius, David Rothschild, and Duncan J Watts. 2021. "Comparing estimates of news consumption from survey and passively collected behavioral data." *Public Opinion Quarterly* 85:347–370.

Krumpal, Ivar. 2011. "Determinants of social desirability bias in sensitive surveys: A literature review." *Quality amp; Quantity* 47:2025–2047.

Lazer, David, Alex Pentland, Lada Adamic, Sinan Aral, Albert-László Barabási, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, and et al. 2009. "Computational social science." *Science* 323:721–723.

Leasure, Douglas R., Ridhi Kashyap, Francesco Rampazzo, Claire A. Dooley, Benjamin Elbers, Maksym Bondarenko, Mark Verhagen, Arun Frey, Jiani Yan, Evelina T. Akimova, and et al. 2023. "Nowcasting daily population displacement in Ukraine through social media advertising data." *Population and Development Review* 49:231–254.

Mangold, Frank, Sebastian Stier, Johannes Breuer, and Michael Scharkow. 2021. "The overstated generational gap in online news use? A consolidated infrastructural perspective." *New Media amp;amp; Society* 24:2207–2226.

Pew Research Center. 2020. "Measuring News Consumption in a Digital Era." Technical report.

Revilla, Melanie, Mick P. Couper, Ezequiel Paura, and Carlos Ochoa. 2021. "Willingness to Participate in a Metered Online Panel." *Field Methods* 33:202–216.

Revilla, Melanie, Carlos Ochoa, and Germán Loewe. 2017. "Using Passive Data From a Meter to Complement Survey Data in Order to Study Online Behavior." *Social Science Computer Review* 35:521–536.

Salganik, Matthew J. 2019. *Bit by bit: Social research in the Digital age*. Princeton University Press.

Scharkow, Michael. 2016. "The Accuracy of Self-Reported Internet Use—A Validation Study Using Client Log Data." *Communication Methods and Measures* 10:13–27.

Scharkow, Michael, Frank Mangold, Sebastian Stier, and Johannes Breuer. 2020. "How social network sites and other online intermediaries increase exposure to news." *Proceedings of the National Academy of Sciences* 117:2761–2763.

Schoen, Harald, Daniel Gayo-Avello, Panagiotis Takis Metaxas, Eni Mustafaraj, Markus Strohmaier, and Peter Gloor. 2013. "The power of prediction with social media." *Internet Research* 23:528–543.

Stier, Sebastian, Johannes Breuer, Pascal Siegers, and Kjerstin Thorson. 2019. "Integrating Survey Data and Digital Trace Data: Key Issues in Developing an Emerging Field." *Social Science Computer Review* p. 089443931984366.

Stier, Sebastian, Frank Mangold, Michael Scharkow, and Johannes Breuer. 2021. "Post post-broadcast democracy? news exposure in the age of online intermediaries." *American Political Science Review* 116:768–774.

Sutton, Jeannette, Emma S. Spiro, Britta Johnson, Sean Fitzhugh, Ben Gibson, and Carter T. Butts. 2013. “Warning tweets: Serial transmission of messages during the warning phase of a disaster event.” *Information, Communication amp;amp; Society* 17:765–787.

Torcal, Mariano, Emily Carty, Josep Maria Comellas, Oriol J. Bosch, Zoe Thomson, and Danilo Serani. 2023. “The dynamics of political and affective polarisation: Datasets for Spain, Portugal, Italy, Argentina, and Chile (2019-2022).” *Data in Brief* 48:109219.

Zagheni, Emilio, Venkata Rama Garimella, Ingmar Weber, and Bogdan State. 2014. “Inferring international and internal migration patterns from Twitter data.” *Proceedings of the 23rd International Conference on World Wide Web* .

Zagheni, Emilio and Ingmar Weber. 2012. “You are where you e-mail.” *Proceedings of the 4th Annual ACM Web Science Conference* .

Chapter 2

Literature Review

Digital devices and the Internet might have a lasting influence on society, affecting how humans develop and socialize. During the last decade the time that people spend online has doubled. The ways in which people connect to the Internet, and how they use it, have also been altered, with mobile devices and social media platforms becoming ubiquitous and gaining more and more importance into people's lives. This change of paradigm has had, most likely, critical consequences on society, and how humans think, feel, and behave. Digital technologies are altering what it means to communicate, work (Garrote Sanchez et al., 2021) and learn (Shortt et al., 2021), and have changed how people consume news and information (Newman et al., 2021), buy products and services (Terzi, 2011), and connect with family, friends, and potential partners (Rosenfeld et al., 2019).

Understanding the effects that the Internet has on society, and the extent to which these might be negative, is an important puzzle for social science and policy. This understanding, nonetheless, can only be produced if scientists can accurately measure people's behaviours on the Internet. While surveys have been one of the main tools available for researchers across disciplines to understand people's opinions and attitudes, as well as their offline behaviours, measuring behaviours in the online realm has proven challenging. This, combined with the explosion in alternative data sources generated as by-products of people's interactions with digital devices and services, has led social scientists to increasingly rely on digital trace data to investigate the impact of the Internet and digital devices on people's lives and, as an extension, on society as a whole. On of the most commonly used types of individual-level digital trace data is web tracking data, which is the focus of this thesis.

In this chapter, I present a broad review of the relevant literature on the importance of understanding what people do online, with a specific focus on digital media, and the different challenges and opportunities of measuring online behaviours with both surveys and digital trace data. I also describe and contextualise web tracking data, showing its main characteristics, and how it compares to other sources of digital trace data. Additionally, I set the foundations to understand why, across the thesis, I propose that the way forward for social scientists interested in understanding the quality of digital trace data is to leverage new computational methods to adapt the vast literature on survey methods and psychometrics to these new sources of data.

The importance of understanding digital media consumption

The measurement of media exposure, which refers to the extent to which individuals encounter specific media messages or content online (Slater, 2004), is of paramount importance for studying the uses and effects of online media. In this era defined by the prevalence of digital and mobile technologies, with Britons on average spending around four hours on the Internet (Ofcom, Online Nation 22), digital media already accounts for more than half of the overall time that people spend consuming media. Within digital media consumption, over fifty percent of this consumption is done through mobile devices (Nielsen and Fletcher, 2020).

This shift in the way that media is created and consumed has sparked substantial interest from both the general public and the academic community, each seeking to comprehend the multifaceted applications and implications that they bring to the table. Specifically, much attention has been devoted to two key topics: the impact that social media might have on people's mental wellbeing; and the role that digital media might have on current concerning political trends such as polarization and misinformation. Considering this, below I review some of the hypotheses and findings regarding these two topics, to showcase the importance of having high-quality measures of digital media exposure.

The relationship between social media exposure and mental well-being

Over the past decade, there has been a marked increase in depression, anxiety, and suicidality among adolescents. Coinciding with this rise is a substantial increase in the time young individuals spend online, particularly on social media platforms. Consequently, both the public and various stakeholders, including policymakers and researchers, have grown increasingly concerned about a potential connection between these two trends. The question that arises is: what might link the amount of time exposed to social media with mental well-being? Among a multitude of hypotheses, researchers have explored the notion that passive engagement with social media (e.g., scrolling through newsfeeds, profiles, or images and videos) could foster heightened upward social comparisons and envy (Verduyn et al., 2017). Additionally, scholars have theorized that individuals who present inauthentic versions of themselves online may experience diminished self-esteem and heightened social anxiety as a result (Twomey and O'Reilly, 2017). Beyond individual behaviours and predispositions, some online content may inherently pose risks, particularly content that promotes unhealthy habits (e.g., anorexia) (Boero and Pascoe, 2012) or is directly abusive in nature.

While research findings have been somewhat conflicting, a consensus emerges that there exists a modest negative correlation between the time spent on social media and various well-being measures (Orben and Przybylski, 2019a,b). Moreover, experimental evidence has demonstrated that participants who abstained from using social media reported increased life satisfaction (Vanman et al., 2018).

However, the overall quality of much of the research conducted to date remains a subject of debate. A key factor contributing to this debate is that a substantial portion of the evidence relies on self-reported measures of individuals' use and consumption of digital media (Orben, 2020). Furthermore, critics have pointed out the overly simplistic focus on the quantity of time spent using digital technologies and media. They argue that the impact of digital media on mental well-being may vary based on the type of digital diet individuals have, the balance of this diet, or the utility of specific online behaviours (Orben, 2021).

Utilizing digital trace data could potentially enhance our understanding of the possible link between digital media exposure and mental well-being. Specifically, the granularity of digital trace data should allow to get more nuanced insights into the digital diets of young people, helping understand not only how long media exposure lasts, but to what media people are exposed to, and how they interact with it.

Is digital media responsible for the rise in polarization?

In recent years, digital media has emerged as the favoured source of news across advanced democratic societies, surpassing television, and print media by a significant margin. Additionally, two-thirds of online media consumers prefer utilizing alternative avenues, such as social media platforms or news aggregators, to access news outlets (Newman et al., 2021). These evolving consumption patterns have precipitated an accelerated diversification and fragmentation of the media landscape (Van Aelst et al., 2017), alongside a transformation in how media is curated and disseminated (Bakshy et al., 2015; Flaxman et al., 2016a). This changing landscape has led some to worry about the implications for our political systems. Can the news media landscape be responsible to some extent for the increased polarization of western democracies? Are digital media driving individuals to isolate themselves in their own ideological bubbles, eroding the common ground in the public arena?

Some suggest so. For years, researchers have hypothesised that the influx of digital technologies could contribute to the emergence of echo chambers, defined as "a bounded, enclosed media space that has the potential to both magnify the messages delivered within it and insulate them from rebuttal" (Jamieson and Cappella 2010:76). How can digital media lead to echo chambers? Digital technologies grant individuals' greater control over their content consumption. At the same time, people are inclined to consume congenial information, especially in the realm of political news (Stroud, 2007; Iyengar et al., 2008; Iyengar and Hahn, 2009). This increased control might lead individuals to choose to consume only information that agrees with their views and opinions, leading to the formation of echo chambers. Empirical research employing both surveys and digital trace data, nonetheless, contradicts this hypothesis. Across various countries, only a limited portion of the population appears to inhabit politically partisan online news echo chambers (for an extensive review, refer to Arguedas 2022). For instance, in the UK, approximately 2% are estimated to engage with a left-leaning echo chamber, while around 5% align with a right-leaning echo chamber (Fletcher et al., 2021). Indeed, some research has shown a remarkable degree of balance in people's overall media diets regardless of partisan affiliation, with audiences tending to converge onto mainstream news outlets (Guess et al., 2020a; Nelson and Webster, 2017; Barberá, 2015). In reality, hence, digital media might have the opposite effect, by enhancing the likelihood of individuals encountering non-congenial information unexpectedly (Barberá, 2015; Dubois and Blank, 2018; Dvir-Gvirsman et al., 2014a).

Others have contended, furthermore, that due to the prevalence of intermediaries like Google and Facebook, which curate and personalize news delivery, individuals might predominantly encounter content aligning with their preexisting beliefs. This phenomenon is known as filter bubbles. However, contrary to this notion, gatekeepers such as search engines and social media platforms are, in most cases, associated with an expansion in the diversity of news consumption (Flaxman et al., 2016b; Cardenal et al., 2019b; Fletcher et al., 2021). As observed by Stier et al. (2021) across six nations, intermediaries such as Facebook and Google foster exposure to a broad range of political and nonpolitical news sources, enhancing the breadth of news types consumed.

The concern underlying individuals' media consumption patterns, and consequently the notions of echo chambers and filter bubbles, revolves around the apprehension that people's media choices might correlate with their levels of ideological and affective polarization. If a surge in partisan media consumption coincides with a decline in the diversity of people's media diets, there is a potential for ideological factions to drift further apart, eroding common ground. Presently, evidence on this matter remains limited and inconclusive. Some studies suggest that a more diverse media diet tends to mitigate polarization among individuals (Padró-Solanet and Balcells, 2022). Complementary, research has also found that predominantly engaging with like-minded partisan media can intensify affective polarization (Hasell and Weeks, 2016). Nevertheless, certain evidence indicates a weak or negligible association between individuals' media diversity and their polarization levels (Guess et al., 2021; Trilling et al., 2016).

Although recent research suggests that early worries about the pernicious effects of social media and digital media on the political system might have been overstated, this does not mean that these have no negative effect. If anything, this body of research shows that the effect of digital media might be more complex and nuanced. Evidence seems to show that a small but relevant proportion of the population engages with misinformation and highly partisan media, sometimes making their own positions more extreme. These small groups, generally with right-wing and conservative tendencies (Guess et al., 2019), are significant enough to have a relevant impact on the political system, and society as a whole. More refined research, hence, is still needed. Furthermore, given that most of this research has been done using digital trace data sources such as web tracking data, the uncertainty about their results is real. As Paper 2 will show, many of the statistics computed with web tracking data backing some of these claims are most certainly significantly biased. It is highly likely that estimates of, for example, people falling into echo chambers or consuming misinformation might be to some extent underestimated.

The challenges of measuring online behaviours with self-reports

Numerous prior studies have employed a diverse array of methodologies to investigate the when, why, and how of individuals' online behaviours. Among these methods, the prevailing approach has been reliant on self-reports collected through survey responses (de Vreese and Neijens, 2016). However, a plethora of research has shown that modifications in wording, formatting, or order can have an effect on the answers that participants give and, subsequently, their measurement quality (DeCastellarnau, 2017). Hence, if not properly designed, questions asking about behaviours can lead to noisy and invalid measures. Specifically in the case of behavioural constructs, the inherent limitation of studies reliant on self-reports lies in the requirement for participants to recollect their past actions and accurately report them. A substantial body of psychometric research has highlighted that self-reported behavioural measurements are susceptible to a multitude of errors, encompassing recall errors and social desirability bias (Sudman, 2010; Schwarz, 2001). Accurately recollection behaviours is impeded by various cognitive constraints inherent in autobiographical memory (Schwarz, 2001; Neisser, 1986). These constraints are particularly salient for behaviours deeply integrated into respondents' daily routines, making them challenging to accurately discern and retrieve (Jobe, 2003; Schwarz, 2001; Neisser, 1986). These challenges are magnified when studying online behaviours, given their increasing fragmentation across various situations, devices, and platforms (de Vreese and Neijens, 2016). Online behaviours typically coexist with offline activities, complicating efforts to disentangle specific behaviours. Moreover, individuals' online actions are composed of micro-interactions (Andrews et al., 2015) that often become intertwined with browsing or app sessions primarily serving a different purpose (e.g., visiting Instagram to check something and inadvertently engaging with an advertisement). At a cognitive level, recollecting these fleeting exposure episodes becomes daunting. This can be even more challenging when recalling specific information such as the content they have been exposed to (e.g., content of articles) or interacted with (e.g., ads clicked) in specific websites or services.

In the past decade, the proliferation of alternative data sources, generated as by-products of people's interactions with digital devices and services, has enabled social scientists to juxtapose self-reports with ”objective” log data. When comparing self-reported and log measures of the same concepts, a consistent trend emerges: there is no agreement between measures. For instance, Revilla et al. (2017)

found that fewer than 4% of participants accurately reported their five most visited webpages over the past week. Specifically, participants tend to overstate their internet consumption (Araujo et al., 2017; Guess, 2015; Pew Research Center, 2020; Scharkow, 2016), with this tendency being more pronounced among heavy users than light users (Araujo et al., 2017; Jürgens and Stark, 2022). Given these discrepancies, self-reports of digital media usage exhibit only moderate correlations with different measurements derived from digital trace data (e.g., call records, screen time) (Parry et al., 2021).

The overarching inference drawn from the divergence between these two sources of data is that self-reports are systematically subject to biases (Araujo et al., 2017; Guess, 2015; Revilla et al., 2017; Scharkow, 2016), thereby challenging their appropriateness for measuring online behaviours whenever alternative options are feasible. While this might not be true, and it will be the focus of much discussion during this thesis, this consensus in the literature has led to an explosion in the interest and use of digital trace data to measure people's online behaviours and, specifically, their exposure to digital media.

Digital trace data to the rescue

In response to the challenges of surveys, social scientists have made significant efforts to develop alternative approaches that do not rely on participants' memory. One increasingly popular method involves collecting digital trace data. This type of data records the interactions of users with specific digital systems (Howison et al., 2011), such as business transaction systems, telecommunication networks, websites, social media platforms, mobile apps, sensors built in wearable devices, and digital devices (Stier et al., 2019). As discussed in Chapter 1, the appeal of digital trace data primarily emanates from its "objective" and highly granular nature, enabling the direct observation of individuals' online behaviours in real-time, at a frequency that surveys cannot match. Moreover, the measurement of digital trace data is nonreactive and non-invasive, obviating the need for individuals to self-report their behaviours. Consequently, human cognition is removed from the data collection process (Keusch and Kreuter, 2021). This capability has the potential to mitigate many of the aforementioned errors associated with self-reports, ultimately enhancing the accuracy of the resulting measurements. Given these advantages, recent literature has increasingly positioned digital trace data as the de facto gold standard for measuring online behaviours (e.g., Araujo et al. 2017; Scharkow 2016), with some authors directly advocating for the substitution of survey self-reports with digital traces when assessing individuals' online behaviours (Konitzer et al., 2021).

However, it is important to note that despite being commonly treated as a homogeneous data source, the term ”digital trace data” encompasses a wide array of approaches, each with its own set of advantages and limitations. These approaches can vary based on factors such as the type of traces collected (e.g., Tweets vs. URLs), the methods employed for data collection (e.g., web trackers vs. APIs), or the degree of involvement of the individuals generating the data (ranging from no involvement to total control over the process). Additionally, owing to the rapid evolution of the field, the types of digital trace data available are in constant flux. For instance, while the original definition of digital trace data was limited to found data created incidentally during activities unrelated to a deliberate research instrument, digital trace data now can also be produced in a designed way. This is sometimes known as designed digital data.

Although providing a taxonomy of the different types of digital trace data in this context risks quickly becoming outdated (Keusch and Kreuter, 2021), the following subsections aim to elucidate some of the sources of digital trace data most commonly employed for measuring general online behaviours, aside from web tracking data. The objective of these subsections is to offer context and situate web tracking data within the broader category of digital trace data, illustrating how similar research questions could be addressed with alternative types of digital trace data and when such choices might be advantageous or not.

Commercial audience measurement data

A method employed by some researchers to measure individuals' online behaviour involves leveraging third-party audience measurement data (e.g., Wu and Taneja 2020). This type of data encompasses aggregated metrics at the level of online entities (e.g., websites) and is derived from a panel of individuals tracked using specific, albeit typically opaque, tracking technologies (Taneja et al., 2017). These audience metrics are generally provided by companies such as comScore, Nielsen, or GfK.

While this data has been regarded as more precise than self-reports at the aggregated level (Taneja, 2016), it does present several limitations. Firstly, this data is sold at the aggregated level, rendering it unsuitable for individual-level analyses. Secondly, researchers lack control over the composition of the sample of individuals. Hence, compared with opt-in panels, there is no option of applying quotas or weights to the samples. Lastly, this data is gathered for commercial purposes, with limited consideration for issues that might hold significance for academic researchers. For example, these panels typically do not treat all audience segments equally, often over-representing demographics that advertisers find more economically appealing (Taneja, 2016).

Platform trace data

An alternative involves acquiring data from specific online services and platforms (e.g., Facebook, Burke et al. 2010; Binge Toolbar, Flaxman et al. 2016a). With this approach, data is extracted directly from the platform, with users not involved in the data collection process. This can be accomplished through various means, with varying degrees of cooperation from the company that owns the data. At one end of the spectrum, researchers can employ web scraping techniques to extract content and data from websites without any collaboration from the platform. Another common approach is to utilize platform-provided *Application Programming Interfaces* (APIs), which allow for the extraction of retrospective user data from the platform. Lastly, researchers can establish direct partnerships with private companies, granting them with controlled access to some of the data that these companies collect for their internal needs.

Platform trace data presents some benefits (Ohme et al., 2023). First, users do not have to be burdened. Therefore, compared with other approaches, researchers do not need to worry about people's willingness to share their data or install technologies. Second, in most cases data can be collected for free or at a reduced cost. Third, when using APIs or collecting data through direct collaborations, the data extraction tends to be relatively easy: either documentation and tutorials are available, or the company directly collects the data and shares it with researchers. All in all, hence, platform trace data can be very convenient and less burdensome to collect for both user and researchers.

Nonetheless, all these approaches present problems. In terms of web scrapping, this approach is affected by its own set of legal, ethical (Krotov et al., 2020), and data quality challenges (see Freelon 2014; Lazer et al. 2021; Tufekci 2014). Moving to APIs and direct collaboration, both approaches hinge on contractual agreements with the data-generating companies, which typically impose specific restrictions on researchers' access to and usage of the data. This, in one way or another, limits researchers' autonomy. Specifically in the case of direct collaborations, companies may exert control over what data can be accessed, and how it might be utilized (Breuer et al., 2020). Additionally, the need for corporate cooperation means that the access to this data can be altered at the discretion of private companies, at any time. A recent example of this is Twitter's announcement regarding the monetization of its API access. Even without such changes, in many cases, few researchers are granted access to this type of data: either because only a few elite researchers can sign agreements with companies, or because APIs are becoming more and more restrictive. This division is contributing to disparities in data access within the computational social science community. Ultimately, as (Wagner, 2023:391) succinctly puts it, ”independence by permission is not independent at all. Rather, it is a sign of things to come in the academy: incredible data and research opportunities offered to a select few researchers at the expense of true independence. Scholarship is not wholly independent when the data are held by for-profit corporations, nor is it independent when those same corporations can limit the nature of what it studied.”

Beyond these considerations, from a data quality perspective, platform trace data typically does not originate from a designed sample of participants but rather from individuals who choose to become users of a specific app or service, making it more susceptible to self-selection biases. Additionally, these approaches usually do not permit the collection of data from the entire platform, but only from specific subsets of users or types of traces. Additionally, since users are not directly involved in the collection of data, in most cases they cannot be contacted for follow-up surveys, barring the collection of auxiliary survey data. Furthermore, the type of data that can be collected and its quality are constrained by what data-generating companies track for their internal purposes. Therefore, even if researchers are granted extensive access to a company's data, they will still be bound by the data infrastructure and limitations set by that company, which to some extent dictate what they can and cannot investigate (Wagner, 2023).

Data donations

An increasingly popular method for gathering digital trace data is through data donations. With this approach, users directly provide researchers with data that has already been collected by their devices or platforms and to which they have access. Various methods are employed to collect data donations in practice, such as requesting participants to capture screenshots of information stored on their devices (e.g., screen time recorded by iOS devices, Ohme et al. 2020a) or downloading the Data Download Package (DDP) of a specific service like Instagram and sharing it with researchers (van Driel et al., 2022). These methods vary along three dimensions (Baumgartner et al., 2022): a) how participants access the relevant traces, b) how they capture these traces, and c) how they share the captured information with researchers. Hence, researchers should aim to make design choices across these three dimensions that minimize the effort required from participants to share data, thereby maximizing compliance, and reducing representation bias.

Data donations offer numerous advantages (Ohme et al., 2023). First, their user-centric nature enables researchers to collect auxiliary survey data into their designs. Second, participants have a higher agency over the data they share, allowing them to provide proper informed consent. Third, compared with the other approaches, data donations allow gathering information of a more private or sensitive nature, such as individual's private messages. Finally, data donations do not require the direct collaboration of data-generating companies, nor rely on expensive technologies normally controlled by a few private companies (more on this in section 4.2).

However, data donation also has its limitations. Because it requires active participation from individuals, it is generally more burdensome for respondents compared to more passive alternatives. In most cases, data donations necessitate participants to take additional actions to collect and share the data with researchers. For instance, with DDPs, participants need to ask companies for their data, wait until this data is available, download it, and then locate it in their device in order to upload it onto the specific data donation system. This can pose challenges for individuals with limited technical proficiency and create a significant burden for those unwilling to go beyond traditional survey tasks. Consequently, this results in a higher perceived cost for participants and potentially lower participation and compliance rates (Silber et al., 2022), which can introduce biases into the final sample of donors.

Indeed, prior research has often found data donation rates to be below 20% (Hodes and Thomas, 2021; Ohme et al., 2020b; Gower and Moreno, 2018). Second, although data donations do not require the direct collaboration of companies, this approach can only yield data in the structure determined by companies. Hence, if companies do not make some data available in their DDPs, or do not allow users to access it in any other way, researchers will not be able to collect this data. This is the case, for instance, with iOS devices, which are significantly more restrictive in the data they make available to users and third parties, complicating the inclusion of iOS users to many data donation projects. This severely limits researchers' agency compared to, for example, web trackers. Finally, data collected via data donation tends to be retrospective in nature. Compared with passive approaches, the granularity of data collected through donations might be limited by the amount of data stored in people's devices and digital services.

Understanding web tracking data

Having considered some of the main types of digital trace data used in the literature, now I turn to the specific data source that this thesis investigates in detail: web tracking data. This approach of collecting digital traces relies on the use of digital tracking solutions (Christner et al., 2021). These solutions, called meters (Revilla et al., 2021), are a heterogeneous group of tracking technologies that can be installed, upon agreement, by participants on their browsing devices. Meters then allow to track a variety of traces left by participants when interacting with their devices online.

Depending on the characteristics of the tracking technologies used, different traces can be collected. For instance, the URLs or apps visited, the terms used in search engines or the content that participants have been exposed to (e.g., HTML information). A variety of terms have been used in the literature to refer to this resulting data, for example, 'metered data', 'web tracking data', 'web log data' and 'digital trace data' (Bach et al., 2019; Cardenal et al., 2019b; Revilla et al., 2017; Dvir-Gvirsman et al., 2016; Cid, 2018). Although data coming from meters might fall under the umbrella of these broad terms, during this thesis I mainly use the terms 'metered data' and 'web tracking data' to refer to data coming from web trackers/meters.

Web tracking data, as a type of design digital data source, differs from other types of digital trace data in two fundamental aspects. First, it is collected from a deliberately designed sample of participants. Web tracking samples can be designed in a similar manner to surveys, enabling researchers to make inferences about specific target populations, similar to survey-based studies. As for surveys, these samples can be built using both probability and non-probability sampling approaches. As such, errors do not come from issues regarding how representative online platforms are, or the way in which to sample traces or users from those platforms, but rather from traditional sampling problems as well as the challenges introduced by asking participants to install tracking technologies on their devices. Together with data donations, the use of web trackers is one of the very few approaches available to collect designed digital data. This is one of the main advantages of web tracking data over other digital trace data sources.

Second, compared to the aforementioned digital trace data sources, including data donations, the nature and quality of the collected data are not heavily constrained by the original purpose of the traces, or the methods available to extract them. Metered data is generated by tracking solutions that capture the traces participants generate when interacting with their devices and online services. While the feasibility of collecting these traces is somewhat limited by the technological capabilities and the cooperation of different operating systems and online platforms with tracking solutions (e.g., iOS terms and conditions do not allow installing apps that can track users' online behaviours), the meters themselves are the primary factor shaping the data collection process, including what data can be collected and its characteristics. As such, the choice of what tracking technologies to use is potentially one of the main, if not the main, design feature of any web tracking project. This choice will determine 1) which participants can be tracked, 2) the types of traces that can be gathered, and 3) the quality and granularity of these traces.

A wide array of tracking technologies has been employed to collect web tracking data (see Christner et al. 2021, and Breuer et al. 2020, for comprehensive reviews). Below, I describe some of the most popular tracking technologies:

1. **VPNs:** Virtual Private Networks can be installed on participants' devices through apps or plugins. They work by routing a device's internet connection through a specially configured remote server network managed by researchers or the company offering this technology. Although not originally designed for

this purpose, VPN servers can be configured to log user activity, including visited websites, data transfers, and timestamps. Individuals typically do not need to perform any additional configuration once they install tracking software based on a VPN.

2. **Automatic Proxies:** Automatic proxies, also known as transparent proxies, can be installed as apps or plugins. Once installed, these proxy tools allow researchers to intercept devices' requests to the internet and save a comprehensive record of the content (Menchen-Trevino and Karr, 2012). Transparent proxies act as "an 'invisible' link in the chain of computers between a user and a website, through which all the traffic of all the participants flow through" (Bodo et al. 2018, p. 147). Individuals do not need to configure the proxy after installation.

3. **Manual Proxies:** Manual proxies, sometimes referred as non-transparent proxies, do not require the installation of any piece of software. Instead, they must be manually configured. Hence, individuals need to access their device settings and manually set up a proxy. Similar to transparent proxies, non-transparent proxies act as an intermediary between a participant's computer and the networks they use, automatically storing all traffic produced.

4. **Browsing History Downloaders:** Downloaders are pieces of software, normally in the form of browser plug-ins, that can access and download the browsing history of participants. This data is retrospective, given that these technologies do not passively track people's behaviours (Guess, 2015; Menchen-Trevino, 2016). Although some researchers consider this approach as a data donation (Ohme et al., 2023), its use of a technology which collects browsing information makes it qualify as a web tracker, according to this thesis definition of the concept.

5. **Screen Scraping:** Screen scrapers, often available as plugins, collect information from the websites that participants visit, by reading and extracting the HTML or XML information of those pages (Marres and Weltevrede, 2013). Data that can be collected is varied, such as URLs, the content of the web pages, or whether participants interact with the content.

6. **Smartphone Loggers:** Loggers can be installed as apps and are used to directly capture log data stored on participants' devices. Depending on the

programming of the logger, it can “monitor a wide range of user activities, including call and SMS histories, GPS data, and information about visited URLs and app usage” (Christner et al. 2021:85).

7. **Screen Recorders:** Screen recorders, generally available as apps, collect data by capturing screenshots of participants’ devices at high frequencies (e.g., every 5 seconds; Reeves et al. 2019). They can also capture this data by recording videos of users’ screens. Information from these images and videos is normally extracted using computer vision algorithms (Krieter, 2019; Bosch et al., 2019).

As Paper 1 and 2 explore in more detail, these different tracking technologies vary in many aspects, all presenting different benefits and drawbacks. Two aspects are particularly important from a data quality perspective. First, not all the different tracking technologies can collect the same type of information nor with the same frequency, granularity, and precision. For instance, if data must be collected after participants install the technology, using a Browsing History Downloader would not be ideal. Second, tracking solutions differ according to the devices (PC or mobile), Operating Systems (e.g., Android or iOS for mobile devices) and browsers (e.g., Chrome or Firefox) on which they can be installed. Hence, tracking solutions impact both who is tracked and how well they are tracked.

Regardless of the technologies employed, these systems need to be set up in order to collect, store, and extract the data (Harari et al., 2016). Researchers typically follow three main strategies, each with its unique characteristics, advantages, and drawbacks. First, **building custom solutions**. Researchers can opt to create tracking technology from scratch, as exemplified by the custom-built plug-in ”Robin” by Bodo et al. (2018). This approach offers the highest degree of control but comes with significant challenges. It demands advanced programming expertise and involves maintaining the software and potentially the hardware infrastructure (e.g., a proxy server). Researchers must also handle participant recruitment, management, and incentives. While it provides full research independence in terms of data collection, it can be resource intensive. Hence, in most cases, these approaches have only been used for smaller studies or pilot projects rather than larger-scale endeavours (Breuer et al., 2020).

Second, **using open-source technologies**. Researchers can leverage existing open-source tracking technologies developed by individuals and available for reuse. For instance, the ”FBforschung” plug-in created by Haim and Nienierza (2019) is designed for collecting browser information and is open for others to adopt. This approach offers the advantage of using technology already developed by others, typically at no cost. However, researchers may still need technical expertise to adapt and run the code and will need to set up a server. The downside is that open-source options are limited to the specifications of existing technologies, which can constrain researchers' capabilities due to the scarcity of suitable options.

Finally, **employing commercial trackers**. Some companies have developed their proprietary tracking technologies, such as Wakoopa and RealityMine. When using commercial trackers, researchers can either purchase data from opt-in online panels of metered individuals (e.g., Araujo et al. 2017) or secure the rights to install the commercial tracker within their own participant samples (e.g., Pew Research Center 2020). Commercial trackers do not require researchers to set up a data collection infrastructure, Nonetheless, they offer limited control over the data collection process, with many aspects of the process remaining undisclosed. The costs associated with purchasing metered data or using these trackers for researcher-owned samples can be high. Consequently, some researchers may be hesitant to rely on technologies not designed, configured, or processed by themselves, as this is the only way they can ensure data quality. As the next subsection will show, this option is the most used in the literature.

Web tracking data in the literature

Due to the well-documented limitations of self-reports and the growing recognition of the advantages of web tracking data, its utilization has gained significant traction in recent years. An exhaustive review of the literature has uncovered 80 published and unpublished papers that leverage web tracking data, as defined in this thesis, for substantive and methodological applications. Table 1 encapsulates some of the key technical characteristics found in these papers. Specifically, it highlights the countries under investigation, the samples employed, the devices subject to tracking, the tracking technologies employed, the focus of the papers, and the year of publication. Focusing on the latter, it seems clear that there has been an upward tendency, with an increased popularity of this approach from 2020 onwards.

Among the 80 papers identified, a notable concentration is observed in terms of the countries studied. The majority of projects have centred their attention on a

Table 2.1: Technical Characteristics of the Corpora of Papers Using Web Tracking Data

Categories	Types	% Papers
Focus	Substantive	74.4
	Methodological	25.6
Sample Provider	YouGov	30.8
	Netquest	20.5
	Unreported	9.0
	GFK	6.4
	LISSS	5.1
	Respondi	3.8
	Lucid	3.8
	Others	25.7
Tracking Technology	Wakoopa	47.4
	Unreported	16.7
	RealityMine	14.1
	WebHistorian	5.1
	Robin	5.1
	FBforschung / Eule	3.8
	Custom built	3.8
	Ethica	2.6
	WebTrack	1.3
	URL Historian	1.3
Devices Tracked	Desktop only	52.6
	Desktop and mobile	44.9
	Mobile only	2.6
	Unreported	2.6
Country	USA	42.3
	Germany	32.1
	Spain	17.9
	UK	11.5
	The Netherlands	9.0
	France	9.0
	Italy	6.4
	Switzerland	3.8
	Other	6.5
Year	2015 - 2019	16.4
	2020 - 2023	83.6
Total Number of Papers		80

Note: Percentages may not add up to exactly 100% due to rounding.

limited selection of countries, with the United States (42.3%) and Germany (32.1%) emerging as the most frequently examined. Additionally, a substantial portion of research has targeted other European nations, including Spain (17.9%), the United Kingdom (11.5%), the Netherlands (9.0%), and France (9.0%). These findings underscore a significant gap in the adoption of web trackers for academic research in regions often categorized as the Global South. This discrepancy may stem from the absence of robust private and public web tracking infrastructures in these markets.

In terms of sample selection, a majority of papers have relied on opt-in online panels, with YouGov (30.8%) and Netquest (20.5%) emerging as the dominant providers. This dependence on a limited number of web-tracking panels, largely controlled by two companies, underscores the substantial influence these companies wield over the technologies and procedures employed for data collection and data quality assurance. Notably, only four studies to date have employed probability-based samples. Consequently, despite the research's overarching goal to make inferences about the general population, minimal emphasis has been placed on ensuring the highest possible sample quality.

The majority of papers have utilized third-party tracking technologies offered by private companies, with Wakoopa being the most prevalent (47.4%). Although RealityMine has gained traction in recent years, largely due to its collaboration with YouGov, its adoption remains limited and primarily concentrated among a few American scholars. Among tracking technologies developed by academics, WebHistorian and Robin are the most widely used (both at 5.1% of papers). Notably, Robin is not open-access, and WebHistorian lacks passive data collection capabilities. An issue of concern is that 16.7% of the papers do not disclose the tracking technologies employed, rendering it impossible for readers to comprehend the data collection approach.

Regarding the devices tracked, most research has exclusively monitored desktop devices (52.6%). However, this approach may not be ideal, as internet consumption is increasingly prevalent on mobile devices compared to desktops (StatCounter, 2017). Consequently, much of the research conducted using web tracking data may exhibit inherent bias due to designed data gaps. Nevertheless, in more recent papers, especially those utilizing opt-in panels such as Netquest and YouGov, there is a growing trend towards simultaneously tracking both desktop and mobile devices. This thesis will further delve into these findings, addressing the potential implications of these trends and patterns in the use of web tracking data for academic research.

Moreover, as highlighted in Table 1, a significant proportion of research conducted using web tracking data has primarily centred on substantive topics (74.4%). In contrast, only a quarter of the 80 papers have delved into the methodological intricacies and challenges associated with this emerging data source. This distribution is unsurprising since the majority of research endeavours are inherently substantive in nature. However, it is worth noting that the prevalent trend has been for research to swiftly adopt web tracking data to address substantive questions, often overlooking the methodological uncertainties surrounding its use and the potential biases it may introduce into substantive findings.

Focusing more on the substantive applications of web tracking data, fields such as political science, communication, and digital journalism research have embraced web tracking data as their gold standard. In the realm of political science, web tracking data has been instrumental in investigating various topics, including the identification of visits to untrustworthy websites (Guess et al., 2019, 2020b), the prevalence of echo chambers (Cardenal et al., 2019b; Dvir-Gvirsman et al., 2014b; Peterson and Damm, 2019), and filter bubbles (Cardenal et al., 2019b). Beyond merely describing these phenomena, research has also delved into exploring the connections between ideology and the use of intermediaries like search engines, examining how these factors influence individuals' susceptibility to echo chambers or exposure to fake news (Cardenal et al., 2019a; Guess et al., 2020a; Peterson and Damm, 2019). Moreover, a substantial body of research has explored how media consumption, particularly partisan media, may be linked to offline behaviours such as voting intention and vaccine uptake (Bach et al., 2019; Cardenal et al., 2019c; Guess et al., 2021).

Beyond political science, web tracking data has found applications in communication and digital journalism research, shedding light on topics like generational disparities in online news consumption (Mangold et al., 2021), the mechanisms behind news discovery and consumption (Kalogeropoulos et al., 2019; Vermeer et al., 2020), the impact of social networks on news consumption (Scharkow et al., 2020), and individuals' receptivity to various forms of branded content (Bol et al., 2020). This multidisciplinary adoption underscores the versatility and relevance of web tracking data as a valuable tool for investigating various facets of contemporary society.

The drawbacks of web tracking data

As highlighted in the preceding sections, web tracking data has been regarded as the gold standard for measuring online behaviours. However, this assumption primarily rests on two key pillars: 1) the acknowledgment that self-reports exhibit some degree of bias, and 2) the presumption that web tracking data measurements are, if not entirely bias-free, at least less biased. While there is ample evidence to support the first pillar of this assumption, the second pillar remains largely unsubstantiated. Not only has no research definitively demonstrated that web tracking data is devoid of errors or of higher quality than self-reports, but certain evidence even suggests the opposite.

Like any novel data collection approach, web tracking data poses methodological challenges that can lead to errors. To liken web tracking data to surveys, and for it to be considered unbiased, two conditions must be met:

1. The utilization of the meter should not introduce selection bias, causing the final sample to systematically deviate from the target population of interest.
2. Throughout the processes of data collection, processing, and measurement creation, no biases should be introduced that deviate the value of the final processed and adjusted measurement from the true behaviour.

Below I present some of the few literature available showing that these conditions, in most cases, might not be realistic.

Representation problems

The installation of a meter can be perceived as intrusive and burdensome by some individuals, potentially leading to a reluctance to participate. When participants who are unwilling to install the meter systematically differ from those who are willing to participate, it can introduce bias into samples of metered individuals.

Previous research has consistently demonstrated low willingness among individuals to install web tracking technologies for scientific research. For instance, Revilla et al. (2019) found that only 16.6% of participants in Spain were willing to install a meter. Keusch et al. (2019) reported that across various experimental settings in Germany, only 35.2% of respondents were willing to install an unspecified app for passive data collection from their devices. Wenz et al. (2019) found that a quarter of participants in the probability-based Innovation Panel were willing to install an app to track their smartphone usage. Actual participation rates, which involve both installing the technology and providing data, align with willingness studies. Gil-López et al. (2023) found a participation rate of 30.4% in a German opt-in panel, while Keusch et al. (2020b) reported a participation rate of 12.8% in a probability-based survey, also in Germany.

These low participation rates may or may not pose problems depending on whether those who accept participation differ significantly from the target population of interest. Previous research has indicated that in nonprobability samples, participant characteristics such as age, gender, education, income, and awareness and use of digital platforms are associated with their willingness to share data, including passive data collection (Kreuter et al., 2018; Revilla et al., 2019; Wenz et al., 2019). Older age has been related to lower willingness to participate in passive data collection via smartphone or smartwatch, while men have shown higher willingness to participate in studies requiring tracking device installation (Mulder and de Bruijne, 2019; Revilla et al., 2021). Additionally, a negative relationship has been observed between education and income and the willingness to participate in online tracking studies (Mulder and de Bruijne, 2019; Revilla et al., 2021).

These differences in willingness to participate have also been evident in studies asking a probability-based sample of participants to install tracking technologies on their devices. Specifically, those who accept tend to be older, have higher income, are more likely to be male, possess greater educational attainment, and exhibit greater awareness and use of digital platforms (Pew Research Center, 2020). These differences persist even when weighted, suggesting that survey and metered data estimates derived from samples of metered individuals could significantly differ from those representing the general Internet population.

Moreover, a review of the existing literature reveals that most papers have predominantly tracked individuals on their desktop devices. This implies that much of the research has excluded mobile-only individuals and, at times, even more specific subgroups such as iOS-only individuals. From a representation standpoint, this can be problematic, especially considering that in some countries, a growing proportion of the population comprises ”mobile-only” Internet users, meaning they exclusively use mobile devices to access the Internet (USA – 10%, Europe – 7%, Smith, 2015). Research exploring the differences between mobile and PC survey participants has found significant disparities in terms of age, gender, income, ethnicity, household type, and ideology (Cook, 2014; de Bruijne and Wijnant, 2014; Lambert and Miller, 2015; Toepoel and Lugtig, 2014; Wells et al., 2014). Consequently, excluding individuals based on device type and operating system can introduce coverage errors, which sociodemographic weighting may not fully rectify (Keusch et al., 2020a).

Measurement problems

While representation issues have been acknowledged as a primary limitation of metered data compared to high-quality surveys, web tracking data, in general, has been perceived as unbiased or less biased than self-reports in terms of measurement (Araujo et al., 2017; Parry et al., 2021). However, this assumption has yet to be substantiated. As Jungherr (2019) points out, for metered data to be perfectly unbiased, every process involved in generating this kind of data must be free of errors. Thus, no errors should occur during 1) the installation or configuration of the meter on individuals' devices, 2) the passive recording of individuals' online behaviour, and 3) the processing and adjustment of the data to calculate final measures. While Paper 1 delves into the various sources of error that can affect web tracking data and presents available evidence on the prevalence and impact of these errors, I will briefly summarize some of the existing evidence regarding potential errors in metered data:

1. **Technology Problems:** Metered data relies on highly complex technologies that are often not designed for research purposes. The complexity and novelty of these technologies make them susceptible to errors. Although there is no direct evidence of technologies introducing errors, Jürgens et al. (2019) found that when using metered data as a gold standard to assess survey data, self-reports of mobile internet use exhibited a higher average bias than self-reports of desktop use. Individuals tended to overestimate the time they spent online on their mobile devices to a greater extent than for desktop data. The authors hypothesize that different tracking technologies used in PCs and mobile devices might result in differential measurement errors.

2. **Device Undercoverage:** To obtain a comprehensive understanding of participants' online activity, meters should be installed on all devices individuals use. Failure to do so will result in missing some of their behaviors, potentially biasing the estimates derived from this data. Several factors can contribute

to undercoverage (see Paper 2). Available information suggests that undercoverage may be prevalent across countries and technologies. For example, combining paradata and survey data, Revilla et al. (2017) found that while almost 57% of respondents had the meter installed on one device, only 4% of those participants reported using only one device to access the Internet. Similarly, only 28% of metered panellists in an Ipsos' Knowledge sample reported having all their internet-accessing devices tracked (Pew Research Center, 2020).

3. **Shared Devices:** While researchers aim to measure individual behavior, they are, in reality, tracking device information. It is typically assumed that all behaviors observed on a device can be attributed to the individual of interest. However, this assumption may be problematic. According to Revilla et al. (2017), more than 60% of desktops, 40% of laptops and tablets, and 9% of smartphones among a sample of metered data participants were shared. This means that some of the observed online activities collected should not be attributed solely to the surveyed individuals.

4. **By-Design Missing Data:** Researchers may decide to measure a concept even when they are aware that some data is missing by design. For example, during the design stage, researchers might know that their tracking approach will not allow them to track mobile devices. Nevertheless, they might still choose to make statistical inferences about individuals' entire online behavior, including their mobile activities, based on incomplete data. This can introduce specification errors because what is measured differs from the actual concept of interest. For instance, Reiss (2022) demonstrates that when measuring the proportion of individuals avoiding news, omitting information about news exposure through mobile apps leads to problematic outcomes. Specifically, disregarding app-based exposure results in an overestimation of the proportion of individuals identified as news avoiders by 8.9 percentage points.

Moving forward

As demonstrated in the previous section, there is evidence to suggest that web tracking data can be susceptible to both representation and measurement errors. The assumption that a data source is inherently unbiased or that its potential errors are negligible can be problematic, as it implies a lack of need to comprehend, quantify, and rectify these potential issues. This is a critical concern because representation and measurement errors can significantly impact statistical analyses. Conversely, acknowledging that a data source may indeed be affected by systematic and random errors opens the door to a more comprehensive understanding of these errors and strategies to minimize them.

Accepting that a data source is not perfect does not necessarily imply that it should be dismissed altogether. Rather, it suggests that data collection, processing, and analysis should proceed with caution, driven by informed decisions. Treating web tracking data as an imperfect source enables researchers to determine when and under what circumstances it is most appropriate for use. It also provides a framework for reducing the overall error associated with web tracking data measures through improved design choices or statistical methods. For instance, although self-reports can be biased, when evaluating their statistics, researchers can use frameworks such as the TSE (Groves and Lyberg, 2010) to identify and estimate potential errors, the effects of those on estimates, and how to minimize them. Extensive empirical research in the survey methodology field has yielded a deep understanding of the measurement quality of different survey questions and how various design decisions can influence the final quality of survey measurements (see DeCastellarnau 2017).

To this date, researchers do not have systematic information about the potential errors associated with collecting web tracking data, or the consequence of these. To help guiding the quality assessment of web tracking data and design appropriate data collection approaches, I assess the quality of web tracking data from a conceptual and an empirical perspective. To achieve this, I suggest adapting frameworks and methods drawn from the extensive literature on survey methodology and psychometrics, commonly used to assess survey quality, to the evaluation of web tracking data. I argue that digital trace data sources, despite their distinct data collection processes and sources of error, share similarities with surveys in terms of representation and measurement errors. Consequently, errors in web tracking data can be identified and categorized in a manner analogous to surveys.

However, it is essential not to overlook the inherent differences between surveys and web tracking data. Surveys and web trackers produce vastly different data, both in terms of their structure and dimensionality. The specific limits of survey data have shaped, to some extent, the methods and approaches used to assess its own quality. While methods can be adapted, directly transplanting survey assessment approaches would not be prudent. The unique characteristics of web tracking data present novel possibilities for methodologists to explore. Specifically, web tracking data is more granular and flexible than survey data. In this dissertation, I illustrate how, with the aid of computational techniques, these distinctive characteristics can be harnessed by researchers to enhance traditional psychometric approaches and evaluate the quality of web tracking data. It is worth noting that this concept builds upon prior work in the field. Several researchers have already expanded the TSE framework to assess the quality of specific digital trace data sources like Twitter (Hsieh and Murphy, 2017), online platforms (Sen et al., 2021), and Big Data in general (Amaya et al., 2020). Furthermore, Oberski et al. (2017) have demonstrated the feasibility of extending the use of MTMMs (Multi-Trait Multi-Method) to simultaneously estimate the measurement quality of administrative and survey data. My work builds on the foundation laid by these and other researchers in this evolving field of study.

Bibliography

Amaya, Ashley, Paul P Biemer, and David Kinyon. 2020. “Total Error in a Big Data World: Adapting the TSE Framework to Big Data.” *Journal of Survey Statistics and Methodology* 8:89–119.

Andrews, Sally, David A. Ellis, Heather Shaw, and Lukasz Piwek. 2015. “Beyond Self-Report: Tools to Compare Estimated and Real-World Smartphone Use.” *PLOS ONE* 10:e0139004.

Araujo, Theo, Anke Wonneberger, Peter Neijens, and Claes de Vreese. 2017. “How Much Time Do You Spend Online? Understanding and Improving the Accuracy of Self-Reported Measures of Internet Use.” *Communication Methods and Measures* 11:173–190.

Arguedas, Ross. 2022. *Echo chambers, filter bubbles, and polarisation: A literature review*.

Bach, Ruben L., Christoph Kern, Ashley Amaya, Florian Keusch, Frauke Kreuter, Jan Hecht, and Jonathan Heinemann. 2019. “Predicting Voting Behavior Using Digital Trace Data.” *Social Science Computer Review* p. 089443931988289.

Bakshy, Eytan, Solomon Messing, and Lada A. Adamic. 2015. “Exposure to ideologically diverse news and opinion on Facebook.” *Science* 348:1130–1132.

Barberá, Pablo. 2015. “Birds of the Same Feather Tweet Together: Bayesian Ideal Point Estimation Using Twitter Data.” *Political Analysis* 23:76–91.

Baumgartner, Susanne E., Sindy R. Sumter, Vladislav Petkevič, and Wisnu Wiradhany.

2022. "A novel IOS data donation approach: Automatic processing, compliance, and reactivity in a longitudinal study." *Social Science Computer Review* 41:1456–1472.

Bodo, B., N Helberger, K Irion, K Zuiderveen Borgesius, J Moller, B van de Velde, N Bol, B van Es, and C de Vreese. 2018. "Tackling the Algorithmic Control Crisis -the Technical, Legal, and Ethical Challenges of Research into Algorithmic Agents." *Yale Journal of Law and Technology* 19.

Boero, Natalie and C.J. Pascoe. 2012. "Pro-anorexia communities and online interaction: Bringing the pro-ana body online." *Body amp; Society* 18:27–57.

Bol, Nadine, Joanna Strycharz, Natali Helberger, Bob van de Velde, and Claes H de Vreese. 2020. "Vulnerability in a tracked society: Combining tracking and survey data to understand who gets targeted with what content." *New Media & Society* 22:1996–2017.

Bosch, Oriol J., Melanie Revilla, and Ezequiel Paura. 2019. "Answering mobile surveys with images: an exploration using a computer vision API." *Social Science Computer Review* 37:669–683.

Breuer, Johannes, Libby Bishop, and Katharina Kinder-Kurlanda. 2020. "The practical and ethical challenges in acquiring and sharing digital trace data: Negotiating public-private partnerships." *New Media & Society* 22:2058–2080.

Burke, Moira, Cameron Marlow, and Thomas Lento. 2010. "Social network activity and social well-being." In *Proceedings of the 28th international conference on Human factors in computing systems - CHI '10*. ACM Press.

Cardenal, Ana S, Carlos Aguilar-Paredes, Camilo Cristancho, and Sílvia Majó-Vázquez. 2019a. "Echo-Chambers in online news consumption: Evidence from survey and Navigation Data in Spain." *European Journal of Communication* 34:360–376.

Cardenal, Ana S., Carlos Aguilar-Paredes, Carol Galais, and Mario Pérez-Montoro. 2019b. "Digital Technologies and Selective Exposure: How Choice and Filter Bubbles Shape News Media Exposure." *The International Journal of Press/Politics* 24:465–486.

Cardenal, Ana S, Carol Galais, and Silvia Majó-Vázquez. 2019c. "Is Facebook Eroding the Public Agenda? Evidence From Survey and Web-Tracking Data." *International Journal of Public Opinion Research* 31:589–608.

Christner, Clara, Aleksandra Urman, Silke Adam, and Michaela Maier. 2021. "Automated Tracking Approaches for Studying Online Media Use: A Critical Review and Recommendations." *Communication Methods and Measures* 0:1–17.

Cid, Enric. 2018. "3 steps to adopt online behavioral data."

Cook, William A. 2014. "Is mobile a reliable platform for survey taking?" *Journal of Advertising Research* 54:141–148.

de Bruijne, Marika and Arnaud Wijnant. 2014. "Mobile Response in Web Panels." *Social Science Computer Review* 32:728–742.

de Vreese, Claes H. and Peter Neijens. 2016. "Measuring Media Exposure in a Changing Communications Environment." *Communication Methods and Measures* 10:69–80.

DeCastellarnau, Anna. 2017. "A classification of response scale characteristics that affect data quality: a literature review." *Quality & Quantity* 52:1523–1559.

Dubois, Elizabeth and Grant Blank. 2018. "The Echo Chamber is overstated: The moderating effect of political interest and diverse media." *Information, Communication amp; Society* 21:729–745.

Dvir-Gvirsman, Shira, Yariv Tsfati, and Ericka Menchen-Trevino. 2014a. "The extent and nature of ideological selective exposure online: Combining survey responses with actual web log data from the 2013 Israeli elections." *New Media amp; Society* 18:857–877.

Dvir-Gvirsman, Shira, Yariv Tsfati, and Ericka Menchen-Trevino. 2014b. "The extent and nature of ideological selective exposure online: Combining survey responses with actual web log data from the 2013 Israeli Elections." *New Media & Society* 18:857–877.

Dvir-Gvirsman, Shira, Yariv Tsfati, and Ericka Menchen-Trevino. 2016. "The extent and nature of ideological selective exposure online: Combining survey responses with actual web log data from the 2013 Israeli Elections." *New Media & Society* 18:857–877.

Flaxman, Seth, Sharad Goel, and Justin M. Rao. 2016a. "Filter bubbles, Echo Chambers, and online news consumption." *Public Opinion Quarterly* 80:298–320.

Flaxman, Seth, Sharad Goel, and Justin M. Rao. 2016b. "Filter Bubbles Echo Chambers, and Online News Consumption." *Public Opinion Quarterly* 80:298–320.

Fletcher, Richard, Craig T. Robertson, and Rasmus Kleis Nielsen. 2021. "How many people live in politically partisan online news echo chambers in different countries?" *Journal of Quantitative Description: Digital Media* 1.

Freelon, Deen. 2014. "On the interpretation of digital trace data in communication and Social Computing Research." *Journal of Broadcasting amp; Electronic Media* 58:59–75.

Garrote Sanchez, Daniel, Nicolas Gomez Parra, Caglar Ozden, Bob Rijkers, Mariana Viollaz, and Hernan Winkler. 2021. "Who on earth can work from home?" *The World Bank Research Observer* 36:67–100.

Gil-López, Teresa, Clara Christner, Ernesto de León, Mykola Makhortykh, Aleksandra Urman, Michaela Maier, and Silke Adam. 2023. "Do (not!) Track me: Relationship between willingness to participate and sample composition in online information behavior tracking research." *Social Science Computer Review* p. 089443932311566.

Gower, Aubrey D and Megan A Moreno. 2018. "A novel approach to evaluating mobile smartphone screen time for iphones: Feasibility and preliminary findings." *JMIR mHealth and uHealth* 6.

Groves, R. M. and L. Lyberg. 2010. "Total Survey Error: Past Present, and Future." *Public Opinion Quarterly* 74:849–879.

Guess, Andrew, Jonathan Nagler, and Joshua Tucker. 2019. "Less than you think: Prevalence and predictors of fake news dissemination on Facebook." *Science Advances* 5.

Guess, Andrew M. 2015. "Measure for Measure: An Experimental Test of Online Political Media Exposure." *Political Analysis* 23:59–75.

Guess, Andrew M., Pablo Barberá, Simon Munzert, and JungHwan Yang. 2021. "The consequences of online Partisan Media." *Proceedings of the National Academy of Sciences*

118.

Guess, Andrew M., Dominique Lockett, Benjamin Lyons, Jacob M. Montgomery, Brendan Nyhan, and Jason Reifler. 2020a. "“Fake news” may have limited effects on political participation beyond increasing beliefs in false claims." *Harvard Kennedy School Misinformation Review* 1.

Guess, Andrew M., Brendan Nyhan, and Jason Reifler. 2020b. "Exposure to untrustworthy websites in the 2016 US election." *Nature Human Behaviour* 4:472–480.

Haim, Mario and Angela Nienierza. 2019. "Computational observation : Challenges and opportunities of automated observation within algorithmically curated media environments using a browser plug-in." *Computational Communication Research* 1:79–102.

Harari, Gabriella M., Nicholas D. Lane, Rui Wang, Benjamin S. Crosier, Andrew T. Campbell, and Samuel D. Gosling. 2016. "Using Smartphones to Collect Behavioral Data in Psychological Science: Opportunities, Practical Considerations, and Challenges." *Perspectives on Psychological Science* 11:838–854. PMID: 27899727.

Hasell, A. and Brian E. Weeks. 2016. "Partisan provocation: The role of partisan news use and emotional responses in political information sharing in social media." *Human Communication Research* 42:641–661.

Hodes, Leora N. and Kevin G.F. Thomas. 2021. "Smartphone screen time: Inaccuracy of self-reports and influence of psychological and contextual factors." *Computers in Human Behavior* 115:106616.

Howison, James, Andrea Wiggins, and Kevin Crowston. 2011. "Validity issues in the use of social network analysis with Digital Trace Data." *Journal of the Association for Information Systems* 12:767–797.

Hsieh, Yuli Patrick and Joe Murphy. 2017. "Total Twitter Error." In *Total Survey Error in Practice*, pp. 23–46. John Wiley & Sons Inc.

Iyengar, Shanto and Kyu S Hahn. 2009. "Red Media, Blue Media: Evidence of ideological selectivity in media use." *Journal of Communication* 59:19–39.

Iyengar, Shanto, Kyu S. Hahn, Jon A. Krosnick, and John Walker. 2008. "Selective exposure to campaign communication: The role of anticipated agreement and issue public membership." *The Journal of Politics* 70:186–200.

Jamieson, Kathleen Hall and Joseph N. Cappella. 2010. *Echo chamber rush limbaugh and the Conservative Media Establishment*. Oxford University Press.

Jobe, Jared B. 2003. *Quality of Life Research* 12:219–227.

Jungherr, Andreas. 2019. "Normalizing Digital Trace Data." In *Digital Discussions*.

Jürgens, Pascal, Birgit Stark, and Melanie Magin. 2019. "Two Half-Truths Make a Whole? On Bias in Self-Reports and Tracking Data." *Social Science Computer Review* p. 089443931983164.

Jürgens, Pascal and Birgit Stark. 2022. "Mapping exposure diversity: The divergent effects of algorithmic curation on news consumption." *Journal of Communication* 72:322–344.

Kalogeropoulos, Antonis, Richard Fletcher, and Rasmus Kleis Nielsen. 2019. "News brand attribution in distributed environments: Do people know where they get their news?" *New Media Society* 21:583–601.

Keusch, Florian, Sebastian Bähr, Georg-Christoph Haas, Frauke Kreuter, and Mark Trappmann. 2020a. "Coverage Error in Data Collection Combining Mobile Surveys With Passive Measurement Using Apps: Data From a German National Survey." *Sociological Methods Research* p. 004912412091492.

Keusch, Florian, Sebastian Bähr, Georg-Christoph Haas, Frauke Kreuter, and Mark Trappmann. 2020b. "Coverage error in data collection combining mobile surveys with passive measurement using apps: Data from a German national survey." *Sociological Methods amp;amp; Research* 52:841–878.

Keusch, Florian and Frauke Kreuter. 2021. "Digital trace data: Modes of data collection, applications, and errors at a glance." *Handbook of Computational Social Science* 1.

Keusch, Florian, Bella Struminskaya, Christopher Antoun, Mick P Couper, and Frauke Kreuter. 2019. "Willingness to Participate in Passive Mobile Data Collection." *Public Opinion Quarterly* 83:210–235.

Konitzer, Tobias, Jennifer Allen, Stephanie Eckman, Baird Howland, Markus Mobius, David Rothschild, and Duncan J Watts. 2021. "Comparing estimates of news consumption from survey and passively collected behavioral data." *Public Opinion Quarterly* 85:347–370.

Kreuter, Frauke, Georg-Christoph Haas, Florian Keusch, Sebastian Bähr, and Mark Trappmann. 2018. "Collecting survey and smartphone sensor data with an app: Opportunities and challenges around privacy and informed consent." *Social Science Computer Review* 38:533–549.

Krieter, Philipp. 2019. "Can I record your screen?" *Proceedings of the 18th International Conference on Mobile and Ubiquitous Multimedia* .

Krotov, Vlad, Leigh Johnson, and Leiser Silva. 2020. "Legality and ethics of web scraping." *Communications of the Association for Information Systems* 47:539–563.

Lambert, Amber D. and Angie L. Miller. 2015. "Living with Smartphones: Does Completion Device Affect Survey Responses?" *Research in Higher Education* 56:166–177.

Lazer, David, Eszter Hargittai, Deen Freelon, Sandra Gonzalez-Bailon, Kevin Munger, Katherine Ognyanova, and Jason Radford. 2021. "Meaningful measures of human society in the twenty-First Century." *Nature* 595:189–196.

Mangold, Frank, Sebastian Stier, Johannes Breuer, and Michael Scharkow. 2021. "The overstated generational gap in online news use? A consolidated infrastructural perspective." *New Media & Society* p. 146144482198997.

Marres, Noortje and Esther Weltevrede. 2013. "Scraping the social?" *Journal of Cultural Economy* 6:313–335.

Menchen-Trevino, Ericka. 2016. "Web historian: Enabling multi-method and independent research with real-world web browsing history data." *iConference 2016 Proceedings* .

Menchen-Trevino, Ericka and Chris Karr. 2012. "Researching real-world web use with Roxy: Collecting observational web data with informed consent." *Journal of Information*

Technology amp;amp; Politics 9:254–268.

Mulder, Joris and Marika de Bruijne. 2019. "Willingness of online respondents to participate in alternative modes of data collection." *Survey Practice* 12:1–11.

Neisser, Ulric. 1986. "Nested structure in autobiographical memory." *Autobiographical Memory* p. 71–81.

Nelson, Jacob L. and James G. Webster. 2017. "The myth of Partisan Selective Exposure: A portrait of the online political news audience." *Social Media + Society* 3:205630511772931.

Newman, N, R Fletcher, A Schulz, S Andi, C T Robertson, and R K Nielsen. 2021. *Reuters Institute digital news report 2021. Reuters Institute for the study of Journalism*.

Nielsen, Rasmus Kleis and Richard Fletcher. 2020. "Democratic creative destruction? The effect of a changing media landscape on democracy." In *Social Media and Democracy*, pp. 139–162. Cambridge University Press.

Oberski, D. L., A. Kirchner, S. Eckman, and F. Kreuter. 2017. "Evaluating the Quality of Survey and Administrative Data with Generalized Multitrait-Multimethod Models." *Journal of the American Statistical Association* .

Ohme, Jakob, Theo Araujo, Laura Boeschoten, Deen Freelon, Nilam Ram, Byron B. Reeves, and Thomas N. Robinson. 2023. "Digital Trace data collection for social media effects research: Apis, Data Donation, and (screen) tracking." *Communication Methods and Measures* p. 1–18.

Ohme, Jakob, Theo Araujo, Claes H. de Vreese, and Jessica Taylor Piotrowski. 2020a. "Mobile data donations: Assessing self-report accuracy and sample biases with the iOS Screen Time function." *Mobile Media and Communication* 9:293–313.

Ohme, Jakob, Theo Araujo, Claes H. de Vreese, and Jessica Taylor Piotrowski. 2020b. "Mobile data donations: Assessing self-report accuracy and sample biases with the iOS Screen Time function." *Mobile Media & Communication* 9:293–313.

Orben, Amy. 2020. "Teenagers, screens and social media: A narrative review of reviews and key studies." *Social Psychiatry and Psychiatric Epidemiology* 55:407–414.

Orben, Amy. 2021. "Digital Diet: A 21st century approach to understanding digital technologies and development." *Infant and Child Development* 31.

Orben, Amy and Andrew K. Przybylski. 2019a. "The association between adolescent well-being and digital technology use." *Nature Human Behaviour* 3:173–182.

Orben, Amy and Andrew K. Przybylski. 2019b. "Screens, teens, and psychological well-being: Evidence from three time-use-diary studies." *Psychological Science* 30:682–696.

Padró-Solanet, Albert and Joan Balcells. 2022. "Media diet and polarisation: Evidence from Spain." *South European Society and Politics* 27:75–95.

Parry, Douglas A., Brittany I. Davidson, Craig J. Sewall, Jacob T. Fisher, Hannah Mieczkowski, and Daniel S. Quintana. 2021. "A systematic review and meta-analysis of discrepancies between logged and self-reported digital media use." *Nature Human Behaviour* 5:1535–1547.

Peterson, Erik and Emily Damm. 2019. "A Window To the Worlds: Americans' Exposure to Political News from Foreign Media Outlets."

Pew Research Center. 2020. "Measuring News Consumption in a Digital Era." Technical report.

Reeves, Byron, Nilam Ram, Thomas N. Robinson, James J. Cummings, C. Lee Giles, Jennifer Pan, Agnese Chiatti, Mj Cho, Katie Roehrick, Xiao Yang, and et al. 2019. "Screenomics: A framework to capture and analyze personal life experiences and the ways that technology shapes them." *Human–Computer Interaction* 36:150–201.

Reiss, Michael V. 2022. "Dissecting non-use of online news – systematic evidence from combining tracking and automated text classification." *Digital Journalism* 11:363–383.

Revilla, Melanie, Mick P. Couper, and Carlos Ochoa. 2019. "Willingness of online panelists to perform additional tasks."

Revilla, Melanie, Carlos Ochoa, and Germán Loewe. 2017. "Using Passive Data From a Meter to Complement Survey Data in Order to Study Online Behavior." *Social Science Computer Review* 35:521–536.

Revilla, Melanie, Ezequiel Paura, and Carlos Ochoa. 2021. "Use of a research app in an online opt-in panel: The Netquest case." *Methodological Innovations* 14:205979912098537.

Rosenfeld, Michael J., Reuben J. Thomas, and Sonia Hausen. 2019. "Disintermediating your friends: How online dating in the United States displaces other ways of meeting." *Proceedings of the National Academy of Sciences* 116:17753–17758.

Scharkow, Michael. 2016. "The Accuracy of Self-Reported Internet Use—A Validation Study Using Client Log Data." *Communication Methods and Measures* 10:13–27.

Scharkow, Michael, Frank Mangold, Sebastian Stier, and Johannes Breuer. 2020. "How social network sites and other online intermediaries increase exposure to news." *Proceedings of the National Academy of Sciences* 117:2761–2763.

Schwarz, N. 2001. "Asking questions about behavior: Cognition, Communication, and questionnaire construction." *The American Journal of Evaluation* 22:127–160.

Sen, Indira, Fabian Flöck, Katrin Weller, Bernd Weiß, and Claudia Wagner. 2021. "A Total Error Framework for Digital Traces of Human Behavior on Online Platforms." *Public Opinion Quarterly* 85:399–422.

Shortt, Mitchell, Shantanu Tilak, Irina Kuznetcova, Bethany Martens, and Babatunde Akinkuolie. 2021. "Gamification in mobile-assisted language learning: A systematic review of Duolingo Literature from public release of 2012 to early 2020." *Computer Assisted Language Learning* 36:517–554.

Silber, Henning, Johannes Breuer, Christoph Beuthner, Tobias Gummer, Florian Keusch, Pascal Siegers, Sebastian Stier, and Bernd Weiß. 2022. "Linking surveys and digital trace data: Insights from two studies on determinants of data sharing behaviour." *Journal of the Royal Statistical Society Series A: Statistics in Society* 185.

Slater, Michael D. 2004. "Operationalizing and analyzing exposure: The foundation of Media Effects Research." *Journalism amp;amp; Mass Communication Quarterly* 81:168–183.

StatCounter. 2017. “Desktop vs mobile vs tablet market share united kingdom.”

Stier, Sebastian, Johannes Breuer, Pascal Siegers, and Kjerstin Thorson. 2019. “Integrating Survey Data and Digital Trace Data: Key Issues in Developing an Emerging Field.” *Social Science Computer Review* p. 089443931984366.

Stier, Sebastian, Frank Mangold, Michael Scharkow, and Johannes Breuer. 2021. “Post post-broadcast democracy? news exposure in the age of online intermediaries.” *American Political Science Review* 116:768–774.

Stroud, Natalie Jomini. 2007. “Media use and political predispositions: Revisiting the concept of selective exposure.” *Political Behavior* 30:341–366.

Sudman, Seymour. 2010. *Thinking about answers: The application of cognitive processes to survey methodology*. Jossey-Bass (Pod).

Taneja, Harsh. 2016. “Using Commercial Audience Measurement Data in Academic Research.” *Communication Methods and Measures* 10:176–178.

Taneja, Harsh, Angela Xiao Wu, and Stephanie Edgerly. 2017. “Rethinking the generational gap in online news use: An infrastructural perspective.” *New Media & Society* 20:1792–1812.

Terzi, Nuray. 2011. “The impact of e-commerce on International Trade and Employment.” *Procedia - Social and Behavioral Sciences* 24:745–753.

Toepoel, Vera and Peter Lugtig. 2014. “What Happens if You Offer a Mobile Option to Your Web Panel? Evidence From a Probability-Based Panel of Internet Users.” *Social Science Computer Review* 32:544–560.

Trilling, Damian, Marijn van Klingeren, and Yariv Tsfati. 2016. “Selective exposure, political polarization, and possible mediators: Evidence from the Netherlands.” *International Journal of Public Opinion Research* .

Tufekci, Zeynep. 2014. “Big questions for social media big data:nbsp; representativeness, validity and other methodological pitfalls.” *Proceedings of the International AAAI Conference on Web and Social Media* 8:505–514.

Twomey, Conal and Gary O’Reilly. 2017. “Associations of self-presentation on Facebook with mental health and personality variables: A systematic review.” *Cyberpsychology, Behavior, and Social Networking* 20:587–595.

Van Aelst, Peter, Jesper Strömbäck, Toril Aalberg, Frank Esser, Claes de Vreese, Jörg Matthes, David Hopmann, Susana Salgado, Nicolas Hubé, Agnieszka Stepińska, and et al. 2017. “Political Communication in a high-choice media environment: A challenge for democracy?” *Annals of the International Communication Association* 41:3–27.

van Driel, Irene I., Anastasia Giachanou, J. Loes Pouwels, Laura Boeschoten, Ine Beyens, and Patti M. Valkenburg. 2022. “Promises and pitfalls of Social Media Data donations.” *Communication Methods and Measures* 16:266–282.

Vanman, Eric J., Rosemary Baker, and Stephanie J. Tobin. 2018. “The burden of online friends: The effects of giving up facebook on stress and well-being.” *The Journal of Social Psychology* 158:496–508.

Verduyn, Philippe, Oscar Ybarra, Maxime Résibois, John Jonides, and Ethan Kross. 2017. "Do social network sites enhance or undermine subjective well-being? A critical review." *Social Issues and Policy Review* 11:274–302.

Vermeer, Susan, Damian Trilling, Sanne Kruikemeier, and Claes de Vreese. 2020. "Online News User Journeys: The Role of Social Media News Websites, and Topics." *Digital Journalism* 8:1114–1141.

Wagner, Michael W. 2023. "Independence by permission." *Science* 381:388–391.

Wells, Tom, Justin T. Bailey, and Michael W. Link. 2014. "Comparison of Smartphone and Online Computer Survey Administration." *Social Science Computer Review* 32:238–255.

Wenz, Alexander, Annette Jäckle, and Mick P. Couper. 2019. "Willingness to use mobile technologies for data collection in a probability household panel." *Survey Research Methods* .

Wu, Angela Xiao and Harsh Taneja. 2020. "Platform enclosure of human behavior and its measurement: Using behavioral trace data against platform episteme." *New Media & Society* p. 146144482093354.

Chapter 3

When survey science met web tracking: Presenting an error framework for metered data

Oriol J. Bosch and Melanie Revilla

Status: This paper has been published in the *Journal of the Royal Statistical Society: Series A.*

Abstract

Metered data, also called web-tracking data, are generally collected from a sample of participants who willingly install or configure, onto their devices, technologies that track digital traces left when people go online (e.g., URLs visited). Since metered data allow for the observation of online behaviours unobtrusively, it has been proposed as a useful tool to understand what people do online and what impacts this might have on online and offline phenomena. It is crucial, nevertheless, to understand its limitations. Although some research have explored the potential errors of metered data, a systematic categorisation and conceptualisation of these errors are missing. Inspired by the Total Survey Error, we present a Total Error framework for digital traces collected with Meters (TEM). The TEM framework (1) describes the data generation and the analysis process for metered data and (2) documents the sources of bias and variance that may arise in each step of this process. Using a case study we also show how the TEM can be applied in real life to identify, quantify and reduce metered data errors. Results suggest that metered data might indeed be affected by the error sources identified in our framework and, to some extent, biased. This framework can help improve the quality of both stand-alone metered data research projects, as well as foster the understanding of how and when survey and metered data can be combined.

Keywords:
Digital trace data · Error framework · Metered data · Web tracking data · Passive data · Total survey error

1. Introduction

1.1. Definitions and main issues

Given the widespread adoption of the Internet, it is becoming vital to better understand what people do online and what impact this has on online and offline phenomena. This requires high-quality data regarding people's online behaviours. Although surveys are one of the most used methods for collecting data in the social sciences (Sturgis and Luff, 2021), it can be complex for participants to accurately remember and report their behaviours through different devices and contexts. Besides, the type of data that is collectable with surveys, as well as its granularity, is inherently limited (e.g., we cannot ask thousands of questions on a single questionnaire, every day). Therefore, in recent years there has been an increase in the use of digital trace data to directly observe what people do online (Breuer, Bishop, and Kinder-Kurlanda, 2020).

A prominent strategy to collect traces about the web browsing and app behaviour of individuals has been to use digital tracking solutions (Christner et al., 2021). These solutions, called meters (Revilla et al., 2021), are a heterogeneous group of tracking technologies that can be installed, upon agreement, by participants on their browsing devices. Meters then allow for a variety of traces left by participants when interacting with their devices online to be tracked. Depending on the characteristics of the technology, different traces can be collected. For instance, the URLs or apps visited, the terms used in search engines or the content that participants have been exposed to (e.g., HTML information). A variety of terms have been used in the literature to refer to this resulting data, for example, "web-tracking data", "web log data" and "digital trace data" (e.g., Dvir-Gvirsman et al., 2014; Bach et al., 2019; Cid, 2018; Cardenal et al., 2019). Although data coming from meters might fall under the umbrella of these broad terms, following Revilla, Ochoa, and Loewe (2017), we use the term **metered data**, which describes the exact data collection procedure.

By directly capturing the digital traces created by participants when interacting with their devices online, data free of recall errors and memory limitations can be captured, with a granularity not achievable by surveys (Revilla, 2022). This data can be used to measure behavioural concepts of interest, potentially bypassing some of the challenges faced by self-reports when measuring online behaviours to make inferences about theoretical concepts for finite populations. Non-behavioural concepts like attitudes might also be measurable with metered data, although there may be fewer benefits.

Albeit limited attention has been paid to metered data errors when used to draw statistical inferences for finite populations, some research has warned about potential errors (Jürgens et al., 2019; Revilla et al., 2017). Indeed, a recent report from the Pew Research Center(2020: 9-10) concluded that "there are still too many pitfalls to rely on [metered data] (. . .) [metered data] does not, at present, seem well suited for high-level estimates of news consumptions". However, a systematic categorisation and conceptualisation of these errors have yet to be developed. Thus, in this paper, we propose a Total Error framework for digital traces collected with Meters (TEM). Total Error is a paradigm used to refer to all the sources of bias and variance that may affect the accuracy and efficiency of data (Lavrakas, 2008). When operationalised as a framework, the Total Error conceptualises and categorises the different sources of error allowing for the understanding of the data collection and analysis process, as well as the identification and estimation of potential errors, their effects on estimates and how to minimise them.

1.2. Goals and contribution

The two main goals of the TEM framework are to (1) describe the data generation and analysis process of metered data, and (2) document all error sources that can affect metered data when they are used to conduct inferential statistics (both univariate and multivariate). To this end, we adapt the Total Survey Error framework (TSE, Groves et al., 2010) for metered data, assuming that the error components presented in the TSE framework can also be found in metered data. Hence, instead of creating a completely new framework for metered data, we start from the TSE framework and modify it to the specific error generating processes and error causes of metered data. Consequently, the TEM can be used by researchers from different backgrounds. This framework provides a common understanding of how to choose the best design options for metered data projects and how to catalogue the potential errors affecting them. For projects integrating both metered and survey data collection (sometimes referred to as *Smart Surveys*; see Ricciato et al., 2020), the TEM can also help to make better-informed decisions while planning when and how to supplement or replace survey data with metered data.

The TEM framework enriches the current landscape of Total Error frameworks, which have been developed for different types of digital trace data sources. Indeed, previous frameworks (see Section 2.2) have focused on secondary data, which are usually called “found” or “organic” digital trace data and which are not designed for research, like the data coming from online platforms (e.g., Sen et al., 2021). Instead, we propose a framework adapted to design-based digital data, mapping in detail the specific error sources produced when tracking individuals’ online behaviours using the heterogeneous group of technologies identified as meters. Furthermore, we illustrate how the TEM framework can help plan metered data collection while minimising errors, using a case study: the Triangle of Polarisation, Political Trust and Political Communication (TRI-POL) project (`https://www.upf.edu/web/tri-pol`). This project combined a cross-national longitudinal survey and metered data collection. We show how the TEM was implemented in the design stage to document, quantify and minimise (when possible) potential error sources affecting the metered data. We also present empirical evidence about the prevalence of some of the error sources in the TRI-POL datasets, and how these might affect the quality of metered data.

2. Background

2.1. Distinctive aspects of metered data

As a design-based digital data source, metered data differ from found data sources in two key design aspects: the *data collection* and *sampling* approaches.

Regarding found data, the nature and quality of the data are heavily limited by the original purpose of the traces (e.g., data from Twitter can only be obtained as Twitter intends and allows to), and the approaches available to download them (APIs, web scrapping, partnering with companies). Researchers have little control over this. Metered data, conversely, are produced by specific tracking solutions that capture the traces that participants generate when interacting with their devices and online services. Although the feasibility of collecting these traces is limited to some extent by the technological capabilities and the “friendliness” of the different operating systems and online platforms towards tracking solutions (e.g., iOS terms and conditions do not allow tracking apps), meters can be considered as the main factor shaping what data can be collected, as well as their characteristics. Many different technological approaches have been used to collect metered data (see Christner et al., 2021 and Breuer et al., 2020 for in-depth reviews), which can be broadly grouped into four categories:

1. Apps that passively and continuously track information from a device and its browser(s).
2. Plug-ins that passively and continuously collect web browsing history and other device and browsing information.
3. Plug-ins that collect the available web browsing history at a given point in time, but without continuously tracking the device/browser activity.
4. Manually configured proxies that send all internet connections made by a device through a network (e.g., WIFI at home) to a server set by the researcher. This information is automatically stored.

These different tracking solutions vary in many aspects, but two are particularly important from a data quality perspective. First, not all the different tracking technologies can collect the same type of information nor with the same frequency, granularity and precision. For instance, if data must be collected after participants install the technology, using the second category (a plug-in that collects available browsing history) would not be ideal. Second, tracking solutions differ according to the devices (PC or mobile), Operating Systems (OSs; e.g., Android or iOS for mobile devices) and browsers (e.g., Chrome or Firefox) on which they can be installed. Hence, tracking solutions impact both who is tracked and how well they are tracked. The Supplementary Online Material (SOM) 1 summarises the capabilities and limitations of the tracking solutions offered by Wakoopa (https://www.wakoopa.com/), which is currently the leading company providing these services. This is also the solution used in our case study. Unlike found data sources (e.g., online platforms), which require selecting existing data sources or platforms (e.g., Twitter) and then extracting available traces (e.g., tweets), metered data are provided using samples of individuals who install tracking solutions on their devices. As for surveys, these samples can be built using both probability and non-probability sampling approaches. As such, errors do not come from issues regarding how representative online platforms are, or the way in which to sample traces or users from those platforms, but rather from traditional sampling problems as well as the challenges introduced by asking participants to install tracking technologies on their devices.

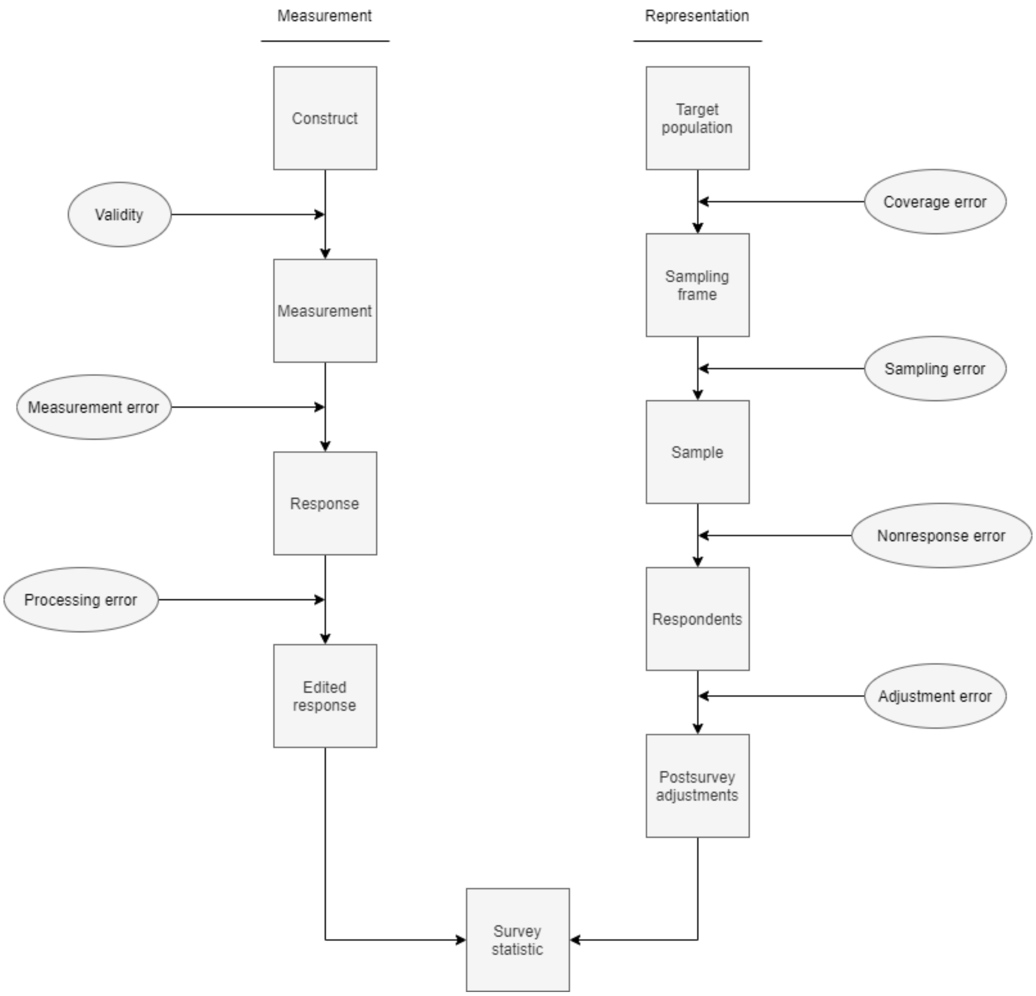

Figure 3.1: Reproduction of the TSE framework by Groves et al.(2009: 48).

2.2. Classification of error sources

Classifying error sources is a good way of thinking about data quality. Although data quality can be conceptualised in many ways (e.g., credibility, comparability, usability, relevance, accessibility), for the last 80 years, most error classification frameworks have explored those sources affecting data accuracy (Groves and Lyberg, 2010). Focusing on accuracy, Groves et al. (2010) built a highly influential error framework for cross-sectional probability-based surveys, the TSE, which links the steps of survey design, collection and estimation into the error sources and separates these into two different dimensions: representation and measurement (see Figure 1). Errors of representation refer to failures to measure eligible members of the population of interest. They include coverage, sampling, non-response and adjustment errors. Errors of measurement refer to deviations between the concepts of interest for researchers and the processed measures collected, and include validity, measurement errors and processing errors. All these errors can affect the estimates' variance or bias, contributing to the overall mean square error of a statistic.

Although the TSE framework was initially conceived for probability-based cross-sectional surveys, in recent years, given the emergence of new types of digital data, researchers have expanded the TSE framework for some found data: Twitter (Hsieh and Murphy, 2017), online platforms (Sen et al., 2021) and Big Data in general (Amaya, Biemer, and Kinyon, 2020). These frameworks assume that the found digital traces, although presenting their own specific data collection processes and error sources, suffer from a number of representation and measurement errors that is comparable to surveys. Therefore, their errors can be identified and classified in a similar way to surveys. Despite these frameworks showing that the TSE framework can be expanded to digital trace data sources, their applicability to design-based digital data sources such as metered data is limited, given the differences in data collection and sampling approaches exposed in Section 2.1. As such, this paper builds on this cumulative knowledge to propose a new framework for metered data.

3. Building a Total Error framework for Metered data

The TEM framework is designed to be flexible and applicable across different types of projects. To achieve this, first, we conceptualise the data collection and analysis process of metered data, and the error components, in a comparable way to surveys. Hence, we use the seven error components of the TSE presented by Groves et al. (2009) as a starting point (see Figure 1). Nonetheless, given the Big Data nature of metered data, we borrow the terminology used by Amaya, Biemer, and Kinyon (2020) to refer to some error components, when it is better suited (i.e., we use "specification errors" instead of "validity" and "missing data errors" instead of "non-response errors").

Moreover, although metered data are longitudinal by nature, most research has used it in a cross-sectional way, aggregating data points to create a measure for a given period (e.g., the time spent visiting online news during a week). As for surveys, aspects of the metered data errors and their interaction might be different in longitudinal contexts. Considering that most past research has used metered data in a cross-sectional way, and for the sake of simplicity, we developed the framework for cross-sectional applications of metered data. Inspired by a previous adaptation of the TSE for longitudinal settings (Lynn and Lugtig, 2017), nonetheless, we highlight processes and error causes that could differ when researchers use metered data in a longitudinal way.

Finally, as for surveys, probability and non-probability-based sampling strategies can be used. Although substantial differences exist between both (Unangst et al., 2019), these are not meter-specific and have already been discussed in previous research (Unangst et al., 2019; Pew Research Center, 2016). Thus, we present the data collection and analysis process when using a probabilistic approach and consider the error causes which would happen for a probability-based approach. However, since most research to date has used metered online opt-in panels, using the work by Unangst et al. (2019), we also highlight the steps and errors that are different or non-existent in that case. It should be noted, nonetheless, that large variations can exist within online opt-in panels (e.g., because of the methods used to recruit participants or select the samples).

4. The TEM: Metered data from a process perspective

Figure 2 presents an ideal workflow of the data collection and analysis process of metered data. For concision purposes, we focus on metered data only, not in combination with survey data.

Researchers conducting metered data research need to make decisions related to two main aspects: the sample and the measurement. On the measurement side (left set of boxes), the first decision is to **define the concept(s) of interest**. This means defining what the researchers want to measure (elements of information that researchers want to collect). To obtain data about these constructs, researchers need to subsequently **design the measurements**, i.e., the specific instrument(s) to be used to gather information about the concept(s) of interest. In the case of surveys, measurements are survey questions (Groves et al., 2009). For metered data, measurements are the defined pieces of information from the participants' tracked online behaviour that are combined, and sometimes transformed, to compute a specific variable (e.g., all URLs that are considered as political articles).

Next, researchers need to **develop or choose the tracking technology** (or technologies) that will be used to obtain the information needed to create the measurement(s). The Supplementary Material (SM) 1 provides a summary of the ones available through the most used provider of tracking solutions: Wakoopa. When using an opt-in metered panel, participants have already installed some tracking technologies on at least one of their devices. Thus, researchers only have the possibility to choose the panel with the best-suited technology (or technologies) for their project.

On the representation side (right set of boxes), the first step is to **Define the Target Inferential Population**, i.e., who the researchers aim to draw conclusions about. The second step is to **Construct the Frame**. A frame is a list (e.g., emails of university students) or a procedure (e.g., a map of houses) that is intended to identify the elements of the target population. When using a metered panel, the panel acts as the frame (Unangst et al., 2019). Consequently, it acts as a list of individuals with a meter either installed or configured on at least one of their devices with an e-mail associated. It does not aim to provide full coverage, but rather looks to include individuals with enough diversity to cover the panel's needs (Groves et al., 2009). The next step is to **Draw the Sample**, which means selecting a fraction of the frame from which measurement will be obtained. Ideally, this should be done using a probability-based sampling approach. In practice, for metered panels, normally non-probability sampling approaches are used, especially quota sampling (Ochoa and Paura, 2018).

Once the sample has been drawn and the technology chosen, sampled individuals can be asked to **Install the Meter** onto their devices. This usually involves asking participants to install or configure several tracking solutions across different devices (e.g., download an app onto an Android smartphone to track the behaviour on the device's browsers and download a plug-in onto a Chrome browser on a Windows PC). This depends on (1) the traces needed by researchers and (2) the capabilities of the chosen technologies. Traces of interest are produced when participants connect to the Internet through specific web browsers and apps installed on a device that is connected to a specific network (i.e., home Wi-Fi or 4G data plan). From now on, we will call these combinations of device/app/web browser/networks the *targets* to track. In some cases, more than one technology might be needed to track different targets on one device (e.g., a plug-in for each web browser used, or a proxy for each

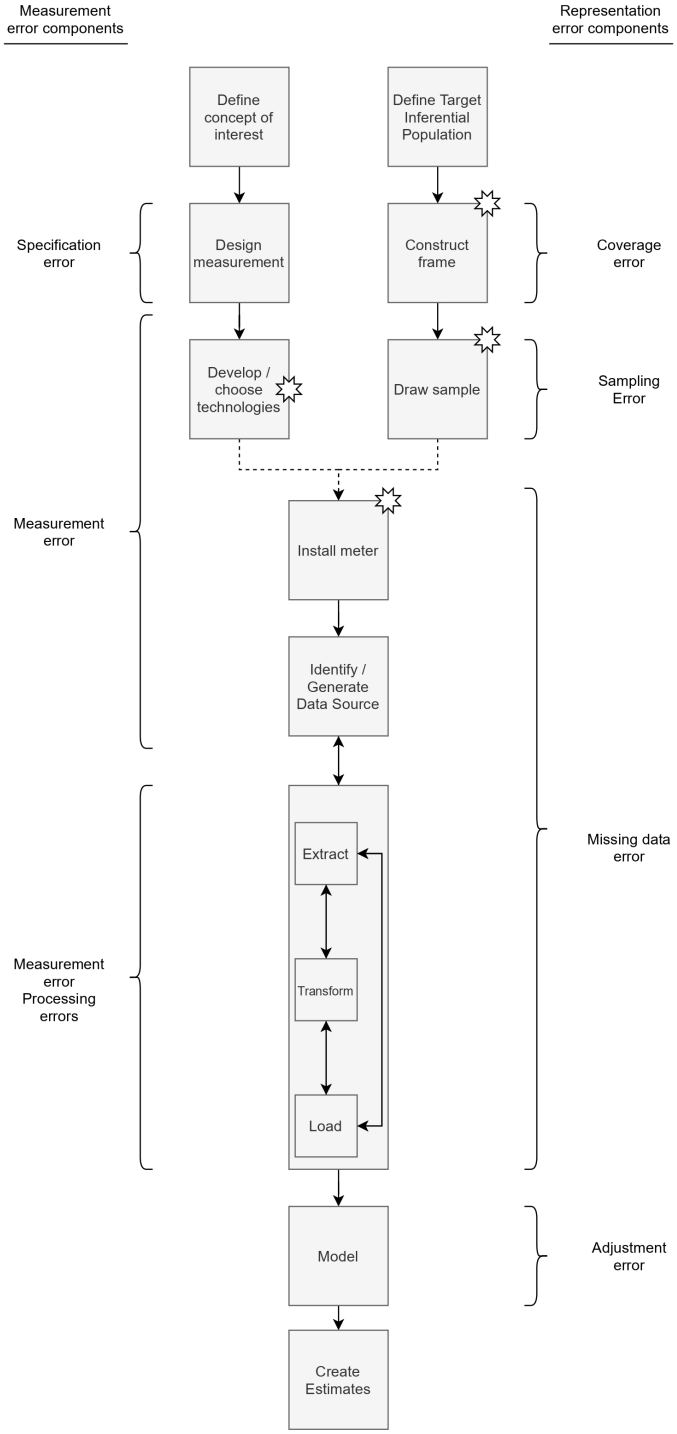
Measurement error components
Representation error components
Define concept of interest
Define Target Inferential Population
Specification error
Design measurement
Construct frame
Coverage error
Develop / choose technologies
Draw sample
Sampling Error
Measurement error
Install meter
Identify / Generate Data Source
Missing data error
Extract
Transform
Load
Measurement error Processing errors
Model
Adjustment error
Create Estimates

Figure 3.2: Data collection and analysis process for metered data. The stars indicate those processes that are different or non-existent for opt-in metered online panels.

network). Consequently, the process of inviting participants can be complex and may include various phases, with there being no standard way of doing it yet. Once correctly installed, the meter starts collecting data from the device and browser logs. When using an opt-in metered panel, the process of installing the technology is beyond the researchers' control since researchers sample from a pool of already tracked individuals. Nonetheless, these individuals can still be asked to further install tracking solutions on top of the ones that they have, only for the specific project (Haim et al., 2021).

In the next step, the information collected by the meter is uploaded to a server that **Generates the Data Source**. Systems to collect and store data can be set in different ways depending on the technology. For example, for smartphone apps, Harari et al. (2016) propose to do the following: a portal server receives the data collected by the meter and checks them against the participant manager, which provides the unique user ID. The portal server subsequently stores the data collected in the data storage, which is normally a database that can handle large datasets (e.g., MySQL). These datasets allow for the data to be queried, extracted and, when necessary, allow for transformations to be applied for the construction of the final dataset for the analyses. When using a metered panel, apart from the information generated after individuals have been sampled, the panel can also provide data already collected from when the participants joined the panel.

Once the dataset of interest has been identified and/or generated, then comes the **Extraction, Transformation and Loading (ETL)** of the metered data. These steps follow a similar process as the one described by Amaya et al. (2020) for found data. Usually, they involve converting the raw and unstructured data into "structured" variables. The steps can be done simultaneously or iteratively (e.g., extracting information, transforming it, loading it back, and extracting it again). First, the process of *extracting* traces can involve (1) selecting subsets from the raw dataset to perform further transformations or (2) extracting information and performing calculations to create "structured" variables (e.g., counts of visits to specific URLs). After extracting traces from the whole dataset, information might need to be further *transformed* to fit the defined measurement. Both simple transformations (e.g., from seconds to minutes) and complex codification procedures that may require using machine learning (ML) applications (see Grimmer et al., 2021, for a discussion of ML application for the social sciences) might be needed. For example, in order to create the variable "time spent visiting pro-conservative news articles", researchers need to code whether the content of the visited news articles can be considered as pro-conservative or not. This can be done manually (but might be too time-consuming) or through a supervised ML algorithm. Once this information is added to the extracted dataset, researchers can create the desired variables. Finally, the extracted and transformed datasets are *loaded* and stored on the researchers' devices or servers. When using a metered panel, these steps might be done by the company or the researchers, depending on whether it is the latter who acquire the raw dataset (i.e., information from all URLs visited by participants in addition to auxiliary information like timestamps) and perform the *ETL* steps by themselves; or if they buy a structured dataset created by the fieldwork company following the researchers' guidelines (i.e., panellists in the rows, variables in the columns; variables based on the information from the raw dataset).

Once a final dataset is loaded, researchers can proceed to **Model**. This involves adjusting the data to better reflect the target inferential population. As such, it can include weighting for missing data, non-response or coverage deficiencies and/or imputation for missing data. Finally, with the adjusted and modelled data, an estimate can be created (e.g., the mean hours of media consumption).

5. The TEM: Metered data from a quality perspective

Each step of the process from constructing the frame to creating the estimates contains some risk of errors. We consider that metered data are affected by the same error components as the ones presented in the TSE (Figure 1). They differ, however, in some of their characteristics and the causes behind them. In the following subsections, we conceptualise those error components for metered data and present their specific error sources.

Metered data and surveys share a similar process when it comes to drawing the sample from the frame, contacting sampled units and adjusting the estimates, and consequently, some error causes are similar or shared with surveys. Since those error causes have been explored extensively (Biemer, 2010; Groves et al., 2009), here we mainly discuss those specific to metered data. Table 1 summarises all meter-specific error causes, by component.

Table 3.1: Specific Error Causes for Metered Data by Error Component

Error components	**Specific error causes**
Specification errors	- Defining what qualifies as valid information
	- Measuring concepts with by-design missing data
	- Inferring attitudes and opinions from behaviours
Measurement errors	- Tracking undercoverage
	- Technology limitations
	- Technology errors
	- Hidden behaviours
	- Social desirability
	- Extraction errors
	- Misclassifying non-observations
	- Shared devices
Processing errors	- Coding error
	- Aggregation at the domain level
	- Data anonymisation
Coverage errors	- Non-trackable individuals
Sampling errors	- Same error causes as for surveys
Missing data error	- Non-contact
	- Non-consent
	- Tracking undercoverage
	- Technology limitations
	- Technology errors
	- Hidden behaviours
	- Social desirability
	- Extraction errors
	- Misclassifying non-observations
Adjustment errors	- Same error causes as for surveys

5.1. Specification errors

A specification error (also known as (in)validity) arises when the concept being measured differs from the concept of interest (Biemer, 2010). For surveys, this arises when the question and scales do not properly measure the defined concept (e.g., the wording refers to something else). For metered data, errors might occur when the traces used to build the variables do not properly match the concept of interest. For instance, to construct the measurement "average time spent consuming political news", the list of visits to URLs considered as "political news" must be defined, and then the time spent on them must be added. If this defined measurement instrument deviates from the concept of interest, for instance by defining some non-political content as political, specification errors appear.

5.1.1. Defining what qualifies as valid information

For surveys, the words used in the request for an answer and the scale categories can affect how valid a measurement is (Saris and Gallhofer, 2014). Equally, when constructing a measurement for metered data, researchers must define which pieces of tracked information should be used and which not to measure behavioural or attitudinal concepts; for instance, whether to consider URLs as fake news or not (Guess et al., 2020). If, due to these specifications, the defined measurement instrument deviates from the concept of interest (e.g., including non-fake news), specification errors are introduced.

5.1.2. Measuring concepts with by-design missing data

Researchers might decide to measure a concept even when they are aware that part of the data is missing by design. For instance, Guess et al. (2018) intended to measure the total fake news consumption of a sample of Americans during the 2016 presidential election. However, the authors collected data only from metered PCs. Thus, by design, the total fake news consumption could not be measured, but only the fake news consumption from the PCs. However, the authors used the collected data to make inferences about the total fake news consumption, making the (strong) assumption that total fake news consumption can be inferred from the fake news consumption found on PCs. If this is not the case, specification errors occur.

5.1.3. Inferring attitudes and opinions from behaviours

Metered data collect behavioural information which can be used as a proxy to measure online behaviours. Other types of digital trace data have been used to measure attitudes and opinions. For instance, Barberá (2015) inferred individuals' left/right position based on the Twitter accounts they followed. If a behavioural indicator (e.g., URLs visited) is used to infer about attitudes or opinions (e.g., left/right position) without a solid theory behind it, it might produce weaker relationships, affecting the validity of the measurement.

5.2. Measurement errors / Missing data errors

When using metered data, measurement and missing data errors can be confounded. As such, we discuss them together. For surveys, the measurement consists in (at least) one question. Participants can either answer or not answer (for whatever reason). Those not providing an answer are considered as missing. Since no information is available from them, they are excluded from the specific analyses. Missingness might happen at the unit level (i.e., no information is available for any measure for a given unit) or at the item level (i.e., information is not available for an item for a given unit). When data are missing, estimates are drawn on a subset of the sample. This can produce *missing data errors* if missing data differ systematically from the available data. For those answering, their answer might deviate from their true values (e.g., their self-reported income does not match their real income), introducing *measurement errors*. This can happen, among other reasons, due to human memory limitations, interviewer influence, deliberate falsification or comprehension errors.

For metered data, a measurement is understood as the defined traces to use and how to transform them (see section 5.1). Tracking solutions are used to collect these traces, which are then transformed into usable variables. This process can fail in at least two ways. First, undefined traces might be wrongly collected and/or classified as correct (e.g., behaviours done by third non-tracked individuals). Thus, more traces are observed than those needed, provoking a measurement errors (e.g., for univariate analyses, it normally leads to a similar phenomenon as with over-reporting). Second, defined traces might not be collected and/or wrongly excluded when creating the variables. This can lead to observing *part* or observing *none* of a participant's behaviour. On the one hand, if we observe part of the behaviour, the observed values are smaller than the participant's true behaviour, leading to measurement errors (e.g., similar to under-reporting in surveys). On the other hand, not observing any of the defined traces leads to a lack of observation, which (as for surveys) should lead to the participants being excluded from the analyses (the lack of behaviour cannot be considered real, so the real value is unknown), introducing missing data errors. Nonetheless, given the nature of metered data, a lack of data (e.g., no adult website URLs recorded) might mean a true absence of behaviour (the individual has not visited any URL of interest) or a failure to capture data (e.g., the participant deactivated the meter to visit such URLs). Therefore, deciding whether the lack of information is considered as missing requires additional information and often depends on the researchers' judgement (see subsection 5.2.9 for a more in-depth discussion of the misclassification problem). This might not be the case, nonetheless, when measuring non-behavioural concepts that require observations of specific behaviours. For instance, to compute the participants' left/right orientation using their visits to political news media websites as a proxy, for participants who do not visit any news media website, no left/right value can be computed.

In short, missing data and measurement errors might be confounded since the same sources of error can lead to each of them depending on (1) how many traces are missed, (2) the behavioural or non-behavioural nature of the concept and (3) how researchers deal with the observed absence of behaviours.

5.2.1. Non-contacts

In order to collect metered data, sampled individuals need to be contacted and asked to install the meter. Regarding surveys, researchers might fail to contact some of the sampled units. For instance, mail or e-mail invitations might never arrive or be seen by a sampled unit. In this scenario, the sampled individual does not become a participant, producing a missing data error. No measurement errors are introduced by non-contacts.

5.2.2. Non-consents

Once contacted, individuals are asked to install the meter onto at least one of their targets. Some sampled units might not be willing to do so because of a variety of reasons. For instance, if the project requires long tracking periods, some participants might not be willing to participate in the study. Hence, no information is collected for those sampled units that do not become participants. Exploring the Netquest opt-in metered panels in nine countries, Revilla et al. (2021) found that between 28.1% (United States) and 53% (Colombia) of those invited accepted to install the meter. No measurement errors are introduced by non-consents.

5.2.3. Tracking undercoverage

Although researchers are normally interested in measuring individuals' behaviours, meters measure the traces left by individuals when using specific targets. To measure the complete behaviour for a specific concept, meters must be installed on all specific targets used to engage in the given behaviour. Nonetheless, several reasons might prevent targets from being tracked:

1. *Non-trackable targets.* Some of the targets used by participants might not be trackable with the chosen technology (or technologies). Hence, information cannot be collected from them. For instance, tracking apps cannot be installed onto iOS devices due to Apple's terms of service.

2. *Meter not installed.* Individuals who consent to being tracked must install or configure the meter into their targets. Even if participants agree to install the meter, they might finally decide not to do so (e.g., after reading the instructions they realise it is too much of a burden) or might fail to successfully do it in some targets (e.g., low IT skills, technical problems).

3. *Uninstalling the meter.* Participants who installed the meter at the beginning of the study can change their mind over time (e.g., lack of memory in the device, change on privacy concerns) and decide to uninstall the meter. Moreover, some participants may uninstall the meter accidentally. Both can happen for some or all the tracked targets. Data are unavailable from the moment the meter is uninstalled.

4. *New non-tracked targets.* During the study, participants might purchase new devices or substitute old ones, switch to new browsers or start to use new networks. If these new targets are not tracked, their information is lost.

When any of these problems occur, not all participant's targets are tracked, leading to tracking undercoverage. The evidence so far suggests that 53% (Spain, Revilla et al., 2017) to 68% (USA, Pew Research Center, 2020) of metered participants do not have the meter installed on some devices. As a result, for some specific concepts of interest, part or all of the data from some participants might not be observed. When the loss of information is partial, this induces measurement errors (see section 5.2). When it is absolute, researchers observe a lack of behaviour. In this scenario, and for behavioural concepts, whether missing data or measurement errors are introduced depends on the researchers' decisions (see section 5.2.9). For non-behavioural measures (e.g., left/right position), an absolute loss of information provokes missing data errors.

For longitudinal research, if the level of undercoverage fluctuates over time, measures of change can be affected by the potential fluctuations in measurement error sizes.

5.2.4. Technology limitations

Tracking technologies are subject to limitations. These prevent them from capturing some data types, in particular, currently: 1) not all available tracking technologies allow for behaviours happening in incognito mode to be captured. 2) Although most technologies can capture domain-level information (i.e., theguardian.com) for all webpage types, some approaches cannot capture subdomain-level information (i.e., theguardian.com/sport/...) for https sites. 3) Behaviours happening inside apps (e.g., profiles visited within the Twitter app), to the best of our knowledge, cannot be captured with any current technology. 4) HTML content cannot be obtained from all tracking technologies. Therefore, depending on the technologies used, some information might not be trackable. Technology limitations can lead to both measurement and missing data errors depending on whether all traces are unobserved or not, the researchers' handling of the absence of behaviours, and the behavioural or non-behavioural nature of the concepts. For longitudinal uses, if technology limitations vary over time (e.g., a new version solving some of the limitations or introducing new ones), measures of change can be affected by the variations in measurement error sizes.

5.2.5. Technology errors

The meters, like any technology, can suffer from technological errors. If the meter stops recording or fails to correctly record information, information is lost. Several reasons can lead the technology behind the meter to fail: 1) the devices or third-party apps might shut down the ability to collect data when devices are low on battery, to reduce the devices' energy consumption. 2) If the meter is working through a proxy, the proxy generates raw data that must be processed to identify which part of the tracked traffic was done passively by the device (e.g., downloading Facebook information) or actively by the participant. This is normally done by trained algorithms. However, this is not completely accurate. 3) Since tracking technologies are built on top of OSs and browsers when new versions of the software are released, they can prevent the technologies from working, causing a loss of information until the technology is adapted to the new version. These errors can provoke an incorrect collection or a loss of information.

Technology errors can lead to both missing data errors and measurement errors. In addition, for proxies, if there is an incorrect collection of information (e.g., the algorithm incorrectly categorises a passive behaviour done by the device as an active behaviour done by the participant), this produces a measurement error similar to over-reporting. For longitudinal uses, technology errors have a similar impact as technology limitations.

5.2.6. Hidden behaviours

Some technological approaches allow participants to disconnect the meter or to configure blacklists of domains not to be tracked. For instance, this can be used to avoid sharing information when dealing with online banking or visiting sensitive web pages.

Hidden behaviours can lead to both measurement and missing data errors depending on whether all traces are unobserved or not, researchers handling of the absence of behaviours, and the behavioural or non-behavioural nature of the concepts. For longitudinal studies, measures of change can be affected by modifications of the blacklisted web pages or how the meter disconnection is allowed and/or used.

5.2.7. Social desirability/Hawthorne effect

Participants might change their behaviours if they know that they are being observed (Jürgens et al., 2019). Consequently, their observed behaviours could deviate from their habitual (non-observed) behaviours. This can especially affect sensitive behaviours, with participants behaving in a more socially desirable way when observed. Although no experimental research has been conducted yet, preliminary evidence using quasi-experimental data suggests that individuals might not change their behaviour when observed (see Toth and Trifonova, 2021).

Changes in behaviours produce measurement errors unless they produce a complete loss of the information needed to compute a non-behavioural measurement, in which case it should be considered as a missing data error. For longitudinal uses, if participants start behaving differently, measures of change can be biased.

5.2.8. Extraction/query errors

Often researchers do not extract all the data, but select specific domains, periods of time or individuals from which/whom to extract information. When specifying the domains or the time frame, incomplete or erroneous specifications can generate measurement errors. For instance, in the case of URLs, if a fake news domain is not specified in the query to extract data, this would underestimate the total fake news consumption. This is not a specification error since the error is produced not from the conceptualisation phase but rather as a mistake when creating and executing the queries used to extract the specified data. Extraction errors can lead to both missing data and measurement errors. Besides, problems with the query can leave sampled participants out of the final database if their information is not extracted.

5.2.9. Misclassifying non-observations

When extracting data from the entire dataset to create metered data variables, only available tracked traces can be used. When no traces are observed for a specific defined measurement, the only reportable information is that no observation exists for that individual. For instance, a query can specify that a variable should be created reporting every time that the URL "theguardian.com" has been observed during a determined time frame. If an individual does not present any observation containing this URL, the query can report that there is no observation. By default, this can be set to missing, or to 0. This lack of data can be due to a true absence of behaviour or a failure to capture data. In a perfect scenario, non-observations should be treated differently depending on whether they are real or the result of some of the previously-mentioned error sources (e.g., tracking undercoverage). Nonetheless, metered data alone do not provide enough information to make this decision. Therefore, researchers might misclassify an individual when deciding. If an individual is misclassified as presenting a lack of behaviour instead of being considered as missing, this inflates measurement errors and deflates missing data errors. Conversely, misclassifying a true lack of behaviour as an error-induced one inflates missing data errors. Hence, misclassifications can introduce missing data or measurement errors, depending on their nature.

5.2.10. Shared devices

Metered data are produced by individuals using specific devices. Devices, nonetheless, can be shared between different individuals. Revilla et al. (2017) found, for the Netquest opt-in metered panel in Spain, that more than 60% of desktops, 40% of laptops and tablets, and 9% of smartphones used to go online by the participants were shared to some degree. Let us assume that researchers want to measure partisan news consumption. Participant 1 shares a metered PC with their father. During the metered period, Participant 1 does not visit any news media website. However, Participant 1's father consumes an average of 1 hour of liberal news media outlets from the shared PC. Participant 1 is considered to present a liberal consumption pattern, although they did not visit any news media website. Now, let us assume that Participant 1 does in fact visit 1 hour of conservative media outlets. Then, Participant 1 is considered to engage with both conservative and liberal media outlets equally, not being polarised. However, their true behaviour would be exclusively conservative.

Shared devices, hence, introduce measurement errors but no missing data errors. For longitudinal uses, if the shared devices patterns vary across time, measures of change can be affected by variations in the sizes of measurement errors.

5.3. Processing errors

Processing errors can be introduced after the data have been collected and before the estimation process. These errors create deviations between the variables used for estimation and the observed ones. For survey data, processing errors can be produced during data entry, coding, editing, disclosure limitation and variable conversions or transformations (Amaya et al., 2020). In the case of metered data, tracked traces might need to be extracted and transformed in specific ways before building the variables, to fit the researchers' needs. Thus, the processed variables can differ from the desired measures.

5.3.1. Coding/categorisation errors

Metered data can take an unstructured form like URLs, text, images or videos. Unstructured data often need to be processed and transformed to be useful. This process might involve coding or categorising the unstructured data into classes, labels, sentiment, etc. This can either be done before extracting the data (e.g., coding domains and subdomains as "political", and then using queries to group their URLs in those defined categories), or after (e.g., extracting the raw dataset, coding each URL in it, and then building the variables). Categorisation, as a process, is related but different to the one presented in section 5.1.1. In this step, the definitions are used to classify the specific information, for instance with coders looking at each URL and judging by using the given definitions. This can be done manually (e.g., using MTurk coders, Peterson et al., 2018), using ad hoc ML algorithms (e.g., supervised ML to categorise the topic of news articles, Peterson and Damm, 2019), or using already available third-party ML algorithms (e.g., Google's Vision AI, Bosch et al., 2019). Manual coding can prompt the same errors as for survey data, i.e., that different individuals coding the same raw data have different judgments or that coders systematically misinterpret and misclassify some information. ML solutions might also present problems. Indeed, recent work has found that label errors are ubiquitous in the test sets of most of the popular benchmarks used in computer vision algorithms (Northcutt, Athalye, and Mueller, 2021). Even if classification algorithms correctly classify the information, labels can still be biased. For instance, for many commercially available systems, images of female U.S. members of congress receive three times more annotation about their physical appearance than about their profession, something which is not observed for their male counterparts (Schwemmer et al., 2020). If these algorithms were to be used to understand the type of images that individuals are exposed to when reading news articles, results would be systematically different depending on the proportion of articles about female/male politicians consumed.For longitudinal analyses, changes to the underlying characteristics of the ground truth data or the ML algorithms used could affect the size of the errors and have an impact on the obtained measures of change.

5.3.2. Data aggregation

In some cases, the final analyses cannot be carried out using the data at the URL level due to vendors, privacy regulations or researchers' decisions. Then, data are aggregated at the domain level (e.g., the domain for theguardian.com/uk/sport/ is theguardian.com). Hence, information is lost. On the one hand, this can lead to some concepts not being measurable. For instance, it is not possible to measure the time spent visiting the sports section of The Guardian if all the URLs with theguardian.com/uk/sport/ are converted into theguardian.com. On the other hand, some concepts might be measured less accurately. For instance, if interested in the total time spent visiting sports articles, the time spent on sports outlets (Eurosport) can be measured accurately but not the one spent on generalist outlets (e.g., theguardian.com). Thus, the final measure underestimates the total time spent visiting sports articles.

5.3.3. Data anonymisation

Data can be anonymised, i.e., all the pieces of information that could lead to identifying participants are obscured. This can be done manually or using ML algorithms. Both approaches, however, can cause errors. Thus, relevant information that was not intended to be hidden can be lost (Ochoa and Paura, 2018).

5.4. Coverage errors

Coverage errors occur when the sampling frame from which the sample is drawn differs from the target population, either because units are excluded, wrongly included or duplicated. If researchers use a metered panel, coverage errors occur when the full panel differs from the target population (Groves et al., 2009). Although unquantifiable per se when using a metered panel, errors are linked to one or more panel practices: for instance, their refreshment strategies or if they blend samples from different sources. Researchers can qualitatively assess panels beforehand to potentially reduce these errors (Unangst et al., 2019).

5.4.1. Non-trackable individuals

Individuals might only use non-trackable targets to access the Internet. Although these individuals might appear in the sampling frame, once contacted, they do not have the possibility of participating. Specific coverage errors related to trackable devices can often not be assessed until sampled units are contacted and the use of trackable devices is assessed. This type of coverage error can be solved if the sampled units using the Internet with non-trackable devices are provided with trackable devices.

5.5. Sampling errors

Sampling errors arise due to the analysis of a subset rather than the entire population of interest. The causes behind sampling errors do not differ between survey and metered data. When units in the sampling frame have a zero chance of selection, these units are excluded from every potential sample drawn. If the excluded units differ from the non-excluded ones in the frame, a bias is introduced. Sampling also introduces variance into estimates. For a given sampling design, many different samples can be drawn. Each sample, by chance, produces different values for the statistics of interest (e.g., average time spent visiting online political media outlets). Several factors can increase sampling variance, such as small sample sizes or the use of clustering.

When using an opt-in metered panel, non-probability sampling is used. Therefore, units are included with unknown probabilities and the sampling error size is unknown.

5.6. Adjustment errors

When modelling and creating estimates, researchers can make use of weighting or imputation strategies with the objective of improving the representativeness of statistical estimates. Since metered data can be based on probability and non-probability samples drawn in a similar fashion as done for surveys, similar weighting and imputation strategies can be used, with similar risks of producing errors. Hence, deficiencies in missing data and coverage error weighting adjustments, as well as imputation for an item missing data, can introduce adjustment errors (see Mercer et al., 2017: 256-257 for an example).

In recent years, ML approaches have been considered to improve weighting and imputation adjustments, with some promising results (Dagdoug, Goga, and Haziza, 2021). The volume and richness of metered data might allow adjusting strategies to be more accurate when dealing with missing data. Nonetheless, it is still too early to know whether we should expect different sources of error when applying these approaches.

6. Case study: applying the TEM to the TRI-POL project

The objective of the TRI-POL project is to understand whether and how online behaviours are related to affective polarisation across Southern European and Latin American countries (`https://www.upf.edu/web/tri-pol`). To this end, a three-wave survey was conducted in Argentina, Chile, Italy, Portugal and Spain between September 2021 and March 2022, and matched at the individual level with metered data. The TRI-POL web tracking strategy was designed using the TEM, to maximise the quality of its data.

Data were collected through the opt-in metered panels of Netquest (`www.netquest.com`) in the five countries of interest. Cross quotas for age, gender, educational level and region were used in each country to guarantee a sample similar to these variables to the general Internet population. Survey questions were used to measure the participants' affective polarization and other attitudinal and demographic variables, while metered data were used to measure variables related to the participants' general Internet use as well as their consumption of news media outlets, political news and social media.

TRI-POL represents a good case study for four main reasons: (1) it focuses on the most frequently measured concepts so far using metered data (social media and news media consumption). (2) As with most past research, it uses a metered panel, as well as the most common tracking solutions provider (Wakoopa). (3) Its cross-national nature allows for the testing of how the TEM framework can be used to create comparable cross-national measures, which is key when comparing standardised relationships across nations (Bosch and Revilla, 2021). (4) Once completed, TRI-POL data will be available with open access (check here: Torcal et al. 2023). The TEM allows for the transparent documentation and communication of the development of the dataset, its limits and best practice when using it.

Below, we describe how the TEM was used to minimise the size of the error sources and/or quantify them. For concision purposes, we limit our consideration to the error sources that could realistically cause bias in our estimates of interest and be measured and/or improved within the TRI-POL context. For those error sources, we also present empirical evidence about their prevalence and/or potential to introduce bias within the TRI-POL datasets of Italy, Portugal and Spain. SM 3 to 8 give more in-depth explanations of the data used and the analyses performed to reach those results. A discussion of the other error sources is provided in SM 2.

6.1. Specification errors in TRI-POL

6.1.1. Defining what qualifies as valid information

The TRI-POL project aimed to measure more than 5,000 concepts across five countries. Due to this big volume, it was key to develop a standardised strategy to create valid measurement instruments across topics and countries. Below, we briefly summarise our strategies to minimise and quantify the specification errors. A more in-depth discussion is proposed in Bosch and Revilla (2022).

Strategy to minimise errors: To operationalise each concept of interest into valid metered data measurements, we developed the following three-step procedure:

1. *Definition of the lists of URLs/apps to be used.* Regardless of the concept (e.g., time in specific sites, type of content exposed), traces of interest are created when an individual accesses to a specific URL or app. Hence, we created comprehensive lists of URLs/apps where traces of interest would be produced for each concept. For instance, for the concept “online news media exposure”, we defined a list of all the URLs/apps considered as ”news media articles”. This involved defining: (1) the news media outlets to be considered in each country and (2) the URLs to consider within each outlet.

2. *Definition of what a visit of interest is.* Even if an individual visited one of the defined URLs/apps, the generated traces could still not be relevant to the concept of interest. For instance, if what was being measured was whether a person read an article or not, only visits complying with the requirements for being considered as read should be used (e.g., a threshold of 120 seconds).

3. *Establishment of the time frame of interest.* Most of the TRI-POL concepts of interest involved measuring average behaviours. To this end, the total sum of behaviours of an individual during the tracked period was divided by a given number of days. This had to be grounded in theory because the chosen time frame could affect the likelihood and prevalence of outliers and the skewness of the data, which can ultimately impact the estimates.

Following these steps not only helped to properly define the traces to use for each concept of interest but also highlighted the design choices for which not enough evidence was available to make informed decisions.

Strategy to quantify errors: When different design choices could be used (e.g., using different lists of the main news media domains in a country), and not a particular one was identified as being better, we computed one variable for each potential design choice. For instance, for the concept “online news media exposure”, we had to make different decisions, such as which list of news media domains to consider or how many days of tracking to use when computing the variables (see SM 4 for the complete list of design choices that we considered). Nonetheless, given the available literature, we thought not possible to make an informed decision about which design choices to use (e.g., is it preferable to use 2 weeks of tracking or one week?). Thus, we created a variable for each potential combination, which resulted in 3,573 variables to measure the concept of “online news media exposure”.

Mirroring what is normally done in the survey literature (see e.g., Smith et al., 2020), for those concepts for which we created more than one variable, we then computed the convergent, discriminant and predictive validity. Following the example of “online news media exposure”, in terms of convergent validity, we computed one correlation for each potential pair of variables. Therefore, we obtained 6,349,266 unique Pearson’s correlation coefficients for each country. On average we found that, across countries, the correlation between the different computed variables was between .40 (Spain) and .51 (Italy), which can be considered as a sign of medium to low convergent validity. This indicates that, depending on the design choices made, variables could not be considered to measure the same latent concept of “online news media exposure.”

Finally, to make sense of this abundant amount of data, we applied random forests of regression trees to estimate the influence of the different design choices on the validity of measurements, drawing inspiration from the Survey Quality Predictor (SQP) software (Saris et al., 2011). This informed us about (1) which measures to use in the final analyses and (2) the robustness of TRI-POL results. (see SM 4 for an in-depth explanation about this). For example, for “online news media exposure”, we used this approach to investigate which design choices helped maximise the predictive validity of the measurements. We found that: 1) tracking participants in both PCs and mobile devices should be preferred over using only PCs or mobile devices (this contrasts with what most past literature has done); 2) when deciding which news media outlets to track, the top 50 most visited news media outlets of a country should be tracked, with little additional predictive power gained with extra tracked outlets; and 3) 10 to 15 days of tracking data should be used, with longer tracking periods not necessarily performing better. A more in-depth and up to date discussion of these results can be found in Bosch (2023).

6.2. Missing data/measurement errors in TRI-POL

6.2.1. Tracking undercoverage

For TRI-POL, we used a panel of self-selected individuals who were already being tracked with specific tracking solutions. Undercoverage was expected (Revilla et al., 2017). However, asking panellists to install new tracking technologies was not possible due to time and budget constraints. Therefore, we assessed the prevalence of undercoverage and estimated the bias it introduced, as summarised below. For a more in-depth explanation, we refer to Bosch and Revilla (2022).

Strategy to quantify errors: To assess the prevalence of undercoverage, two pieces of information were needed for each participant; which targets were tracked, and which targets they used to go online. The first piece of information was obtained using paradata and the second by asking participants questions about which devices and browsers they used to access the internet during the 15 days before the first survey wave (see SM 5 for the exact formulation used). Combining both sources of information, we estimated the proportion of undercovered individuals, as well as the number and types of non-tracked devices and browsers. Across the different countries, we found that the between 80.5% (Spain) and 85.7% (Portugal) of TRI-POL participants had at least one device or browser not tracked. Hence, tracking undercoverage is highly prevalent in the TRI-POL datasets.

The combination of survey and tracking paradata also allowed us to develop an approach to estimate the bias introduced by undercoverage. Specifically, using metered data from the subsample of fully covered individuals, we simulated how different levels (% of participants not having all PCs or mobile devices covered) of undercoverage would cause a bias for a set of univariate and multivariate estimates from their true values (see SM 5 for more detailed information about this approach). As an example, we simulated the bias that device undercoverage could introduce to the results obtained for the measure "average time spent on the Internet", which represents the average time spent on any URL or app across all tracked targets, for the 15 days prior to the survey being answered. We estimated that, when using the TRI-POL dataset, tracking undercoverage could underestimate the average time spent by participants on the Internet by 3.9% (Spain, in a scenario with 25% of the sample without information about PC behaviours) to 25.8% (Spain, 75% of the sample without information about mobile behaviours). This would imply that, at the actual levels of participants in the TRI-POL samples with all their PCs or mobile devices not tracked, the observed average time spent on the Internet might be underestimated by around 14%, across the different countries. This estimate does not take into account the effect of individuals having some but not all of their PCs or mobile devices not tracked, hence, the real bias is most likely even higher. A more in-depth discussion of these results can be found in Bosch et al. (2023)

6.2.2. Technology limitations

To reduce errors produced by technological limitations, ideally, we would: (1) identify the digital traces and their characteristics and (2) design or select the technologies allowing for their collection with the highest level of accuracy possible. In the TRI-POL project, the control over (2) was limited since data were collected through an already existing metered panel. Nonetheless, we defined strategies to quantify the prevalence and potential impact of these limitations on the estimates.

Strategy to quantify errors: We asked Netquest to provide information on all their tracking solutions and their limitations (what devices and traces they could or could not track, their level of accuracy, potential limits and interactions between devices/OSs and technologies; see SM 1 and 6). We then combined this information with the paradata about the technologies used to track each participant to compute the proportion of participants who could be affected by specific technology limitations (e.g., not being able to track incognito tabs). This allowed us to compute of proportion of participants:

1. *Not trackable in incognito mode*: we found 13.5%, 6.2% and 8.1% of participants affected by this technological limitation in Italy, Portugal and Spain, respectively.
2. *Without subdomain information*: those represented 20.5% (Italy), 12.0% (Portugal) and 14.8% (Spain) of participants.
3. *Without in-app information*: given that the tracking solutions used by Netquest cannot capture in-app behaviour, 100% of TRI-POL participants were affected by this, across all countries.

The nature of these limitations, nonetheless, prevented us from assessing the extent to which they might influence the final estimates. For instance, since the information from in-app behaviours cannot be obtained with any current approach, it is not possible to assess how much and which type of information is missed.

6.2.3. Technology errors

Although researchers have little control over technology errors when using a metered panel, we designed the following strategies.

Strategy to minimise errors: First, we identified whether the tracking solutions used were susceptible to being shut down by energy-saving apps and/or built-in features of the devices. This was not the case. Second, we limited the sampling pool to participants with up-to-date meters, which are better equipped to deal with the latest OS versions. Third, due to some of the errors, meters can stop tracking entirely. As such, sampled participants might not produce any metered data during the tracking period. To avoid this, we excluded participants without any tracked behaviour during the last month from the sampling pool. This could exclude very low-frequency internet users, but their very presence should be rare in an opt-in online panel. Finally, manually configured proxies sometimes produce inaccurate results. An approach to avoid these problems could be excluding participants using iOS devices. We considered that the undercoverage errors introduced by this would outweigh its benefits. As such, TRI-POL data from iOS devices might potentially be affected by measurement errors, which should be quantified.

Strategy to quantify errors: The nature of proxy errors makes them hard to quantify since it is complex to understand when and how the classification algorithm might have failed deciding which traces to include or exclude. However, indirect strategies can be used to test whether iOS users present different measurement properties. As an example, we tested whether being tracked on an iOS was associated with the absolute difference between the self-reports and metered data for the variable "average time spent on the Internet" (Absolute difference = — *Self-reported time – Tracked time* —, see Araujo et al., 2017). Although we expect both the survey and metered measures to be affected by errors, a significant effect of being tracked on an iOS could indicate that participants tracked on an iOS present different measurement properties. SM 7 discusses in more depth the exact analyses conducted. We found that being tracked on an iOS significantly increases the absolute difference in Italy and Spain, but not in Portugal. Specifically, in the TRI-POL dataset, being tracked on an iOS device is associated with having and absolute difference 56.8 and 57.6 minutes larger than for those not tracked on an iOS. Therefore, controlling for different potential confounders, we see that the mismatch between the survey and the metered data measures for iOS respondents is substantially higher than for those not tracked on an iOS. This could indicate that measures coming from iOS devices present different measurement properties, potentially of lower quality.

6.2.4. Hidden behaviours

Participants had two potential ways of hiding their behaviours: blacklisting domains and disconnecting their tracking technologies.

Strategy to minimise errors: We asked Netquest for their blacklisted domains in order to know whether some behaviours would be missing by design. None of TRI-POL's defined URLs were blacklisted.

Strategy to quantify errors: No information was available about whether participants disconnected their trackers or not, nor was there information about the types of content triggering this. Besides, considering that participants had the tracking technologies installed before sampling them, we could not apply quasi-experimental approaches like the ones proposed by Toth and Trifonova (2021). Consequently, we could not quantify (1) whether participants disconnected their meters and (2) the extent to which this could cause a bias in TRI-POL's estimates.

6.2.5. Misclassifying non-observations

Minimising and quantifying misclassifications required collecting as much information as possible about the errors which could provoke error-induced non-observations, to clearly differentiate which non-observations should be considered real and which not.

Strategy to minimise errors: Information about other error sources was used to decide which non-observations were coded as real (0), as error-induced (NA) or as unclear (specific code). For instance, for a few concepts, we directly asked participants whether they had visited specific domains with non-tracked targets (see SM 7 to check the questions asked). This information allowed us to discern between real and undercoverage-induced non-observations, potentially reducing the extent to which misclassification could affect the estimates.

Strategy to quantify errors: This same information also allowed us to compute the prevalence of undercoverage-induced true non-observations across different topics, individuals and countries. For instance, we computed the proportion of individuals with undercoverage-induced non-observations for Facebook, Twitter and the five most popular news media outlets in each country (see SM 8 for a more in-depth explanation). Across the different countries and domains, we found between 3.1% (Italy, GazzetaSud) and 25.9% (Spain, RTVE) of participants with undercoverage-induced non-observations. These proportions, nonetheless, varied highly depending on the domain of interest. In terms of social media, Facebook presents a low proportion of participants with undercoverage-induced non-observations (from 6.7% in Portugal to 8.4% in Spain), whereas the proportion of affected participants is substantially higher for Twitter (11.7% in Spain - 19.0% in Italy). The proportion of individuals with undercoverage-induced non-observations for news media outlets was on average between 10.1% (Italy) and 15.1% (Portugal). Thus, in the TRI-POL dataset, there is a non-negligible risk of increasing the size of the estimate's measurement errors if these participants are not excluded from the analyses.

In order to quantify the extent to which misclassifying these non-observations as true lacks of behaviours could introduce bias, one could apply a similar simulation strategy as for undercoverage (see 6.2.1): undercoverage scenarios can be combined with misclassification scenarios; simulated non-observations should be randomly misclassified to approximate the effect that this could have in the full dataset.

6.3. Processing errors in TRI-POL

6.3.1. Coding/categorisation errors

The TRI-POL project measured variables about the participants' consumption of specific news media content (e.g., political, national, opinion). Thus, we had to define what we considered, for instance, to be "political" or "opinion" content. Using these definitions, we created guidelines to code subdomains of the tracked news media outlets as "political" or not, "opinion" or not, etc. (see https://www.upf.edu/web/tri-pol/documentation-and-data-archive). Using the guidelines, human coders went through all the listed news media outlets in each country and categorised their subdomains as mostly containing national, international, regional, opinion or other articles (e.g., theguardian.com/politics mostly lists URLs covering political news and, as such, it is considered as political).

Strategy to minimise errors: Coders were required to have language and country-specific knowledge. To supervise that the coding approach was applied in a comparable way across countries and to spot potential coding errors, another researcher supervised all the coders' work.

Strategy to quantify errors: Given the time and budget constraints, only one coder per country was used for the full dataset, preventing us from computing inter-coder reliability. Therefore, it was not possible to quantify how codes varied across coders.

7. Discussion

Metered data are increasingly being used to understand people's online behaviours. In this article, we propose a Total Error framework for Metered data, the TEM, which documents and conceptualises the data generation and analysis process of metered data. It also distinguishes the various kinds of errors that might arise during each step in the process. By expanding the TSE for metered data, the TEM framework can be used by researchers from different backgrounds, to improve documenting, quantifying and, when possible, minimising errors. Given the framework's flexibility, the TEM can be applied to stand-alone metered data projects and projects combining metered data with surveys, to probability and non-probability-based sampling approaches and to cross-sectional and longitudinal research.

7.1. Limitations and future research

The TEM framework presents some limitations. First, as the TSE, the TEM only considers a definition of data quality. Other factors should be considered when deciding whether to use metered data (e.g., cost, timeliness, risks). In particular, privacy and ethical issues must be considered when planning to collect metered data. Truly informed consent might be more difficult to obtain than in surveys due to the limited understanding of some participants regarding the data being collected (Revilla, 2022). Moreover, metered data should not be shared nor made publicly available in its raw form if it can represent any risk to the participants' privacy, even if this could make some results not fully reproducible.

Second, the TEM considers the errors of the metered data independently. Nonetheless, metered data have mostly been collected in parallel with surveys. In such cases, one must consider the potential trade-offs between the active (surveys) and passive (metered) data collection parts of their projects. These trade-offs might vary across the many potential ways in which surveys and metered data can be combined (Revilla, 2022). In-the-moment surveys represent a good example (Ochoa and Revilla, 2022). By tracking what people do online with metered data, in-the-moment surveys can be triggered when a participant shows a specific behaviour, allowing (1) for the measurement of new survey concepts and (2) the enhancement of metered data (for instance, by recording the reasons behind behaviours). Nonetheless, sending surveys after a specific behaviour might make participation in the project more burdensome, as well as the monitoring of their behaviours more evident. This could potentially increase social desirability and the likelihood of dropping out even if the (limited) existing evidence does not seem to support these ideas (Ochoa and Revilla, 2021). Although our framework does not directly address the particularities of these projects, its similarity with the TSE allows researchers to use both frameworks simultaneously to consider potential interactions. Nonetheless, further research is needed to get a better understanding of how to theoretically and empirically do this.

Third, although this paper discusses and exemplifies the strategies followed in the TRI-POL project to quantify error sources, standardised approaches to quantify metered data errors still need to be developed and tested. The TEM framework, however, can serve as the foundation for future empirical research. Based on our experience, future research should explore at least three areas: (1) since the representation and measurement processes of metered data resemble those of surveys, approaches used to quantify surveys errors could, in principle, be adapted to metered data (e.g., GMTMM for administrative and survey data, Oberski et al.). (2) Combining metered data with survey data can help have a better understanding of the quality of both sources. (3) Metered data allow for the creation, for each concept of interest, of dozens or hundreds of variables by simply altering the specifications of the queries. This might allow for ML algorithms to be fed with hundreds or thousands of quality estimates, for a single study, to predict what characteristics might yield the least biased estimates.

Fourth, even if this paper presents some empirical evidence regarding the data quality of the TRI-POL datasets, the prime goal of these examples is to showcase how the TEM might help when designing instruments based on metered data, and how the quality of these can be empirically tested. More in-depth research must be conducted to get a good understanding of a) the best approaches to measure metered data quality, b) the general quality of metered data and c) the relevance of each error source.

Fifth, in the coming years, ML might play an important role in future uses of metered data. For instance, ML might increasingly be used to process metered data and/or the use of metered data in predictive statistics might become more prevalent (Hofman et al., 2021). If this happens, the TEM framework can be used to accommodate current debates on the challenges of ML for big data sources, and potential solutions (see Qiu et al., 2016 for a good summary of this). This could mean understanding how ML might amplify (Wang and Russakovsky, 2021) or reduce (Ramirez, Abrajano, and Alvarez, 2019) already existing biases or how and when automatisation errors could introduce new sources of error.

Finally, although key differences separate metered data from other digital trace data sources, the TEM framework can be applied to better understand the errors of other digital trace data sources or used as the foundation for new frameworks. On the one hand, metered data are similar to platform trace data or data donations; error sources like defining what qualifies as valid information or tracking undercoverage can be translated. On the other hand, parallels can be drawn with traces like GPS or call log data. In these cases, most measurement, processing, missing data and coverage errors might apply (with potential differences in the details), since these have to do with the technologies being installed and properly working.

7.2. Conclusions and practical recommendations

Well-designed metered data can be a very useful resource. By developing the TEM, our goal was twofold: to help researchers (1) understand the limitations of already published research using metered data, and (2) to design metered data collection strategies that minimise the error size and allow for the quantification of the remaining errors.

Most research done to date has not discussed the potential limitations of metered data in enough detail. For instance, it is common not to report the tracking solutions used, nor how measurements have been defined, nor the prevalence of errors such as tracking undercoverage. This is not an adequate practice since it does not provide enough information to judge data quality, and it is not in line with current good practices in survey research (e.g., comparable to not reporting response rates). This is even more pressing if we consider that, when applying the TEM to quantify the error sources affecting the TRI-POL datasets, we have found that some errors such as tracking undercoverage, invalidity, technology errors or the misclassification of non-observations are highly prevalent. Moreover, we showed that invalidity and tracking undercoverage have the potential to bias the results obtained (e.g., underestimating univariate estimates or reducing the predictive power of variables in multivariate models).

Considering this, and based on the TEM and our experience with TRI-POL, we propose some practical recommendations when using metered data:

1. Clearly define the list of traces (and how to transform them) to create valid measurement instruments. If it is not possible to make an informed decision and differences in the validity of different design choices are expected, create several measurements and test the robustness of the results and/or validity.

2. Consider the potential consequences that the different technologies can have on data quality before deciding which one(s) to use. If this is out of control, list all their limitations and report their prevalence and how these might affect the final estimates.

3. Clearly define what targets need to be tracked and try to maximise their coverage. If this is not possible or is out of your control, collect auxiliary information to assess its prevalence and potential to introduce bias.

4. Be mindful that tracked devices might be used by non-participants. Try to develop strategies to minimise or assess how this can affect the estimates.

5. Non-observations might be caused by errors. Auxiliary information should be collected to classify non-observations as real or induced by errors, in order to deal with them accordingly.

6. Develop strategies to minimise and correct for human or machine induced errors when extracting and transforming the raw data into observed variables. Metered data projects can quickly become complex, involving many steps that might be sensible to errors.

To conclude, collecting high quality metered data is complex and involves a high degree of uncertainty. Decisions often require balancing many pros and cons with limited information. This does not imply that metered data should not be used, or that previous research might be biased. It means, instead, that working with metered data requires a high degree of care and transparency and that further research is needed to help researchers to optimise their use of such data.

Acknowledgements: I would like to thank Patrick Sturgis, Jouni Kuha, and Mariano Torcal for their support and insightful comments

Funding: Fundación BBVA; H2020 European Research Council, Grant/Award Number: 849165; Ministerio de Ciencia e Innovación, Grant/Award Number: PID2019-106867RB-I00/AEI/10.13039/501100011033

Data: The data that support the findings of this study is openly available in OSF at https://osf.io/3t7jz/ (DOI: 10.17605/OSF.IO/3T7JZ).

Bibliography

Amaya, Ashley, Paul P Biemer, and David Kinyon. 2020. "Total Error in a Big Data World: Adapting the TSE Framework to Big Data." *Journal of Survey Statistics and Methodology* 8:89–119.

Bach, Ruben L., Christoph Kern, Ashley Amaya, Florian Keusch, Frauke Kreuter, Jan Hecht, and Jonathan Heinemann. 2019. "Predicting Voting Behavior Using Digital Trace Data." *Social Science Computer Review* p. 089443931988289.

Barberá, Pablo. 2015. "Birds of the Same Feather Tweet Together: Bayesian Ideal Point Estimation Using Twitter Data." *Political Analysis* 23:76–91.

Biemer, Paul P. 2010. "Total survey error: Design, implementation, and evaluation." *Public Opinion Quarterly* .

Bosch, Oriol J. 2023. "Validity and Reliability of Digital Trace Data in Media Exposure Measures: A Multiverse of Measurements Analysis."

Bosch, Oriol J. and Melanie Revilla. 2021. "The Quality of Survey Questions in Spain: A Cross-National Comparison." *Revista Española de Investigaciones Sociológicas* 175:3–26.

Bosch, Oriol J. and Melanie Revilla. 2022. "The challenges of using digital trace data to measure online behaviors: lessons from a study combining surveys and metered data to investigate affective polarization." *SAGE Research Methods Cases* .

Bosch, Oriol J., Melanie Revilla, and Ezequiel Paura. 2019. "Answering mobile surveys with images: an exploration using a computer vision API." *Social Science Computer Review* 37:669–683.

Bosch, Oriol J, Patrick Sturgis, Jouni Kuha, and Melanie Revilla. 2023. "Uncovering digital trace data biases: tracking undercoverage in web tracking data."

Breuer, Johannes, Libby Bishop, and Katharina Kinder-Kurlanda. 2020. "The practical and ethical challenges in acquiring and sharing digital trace data: Negotiating public-private partnerships." *New Media & Society* 22:2058–2080.

Cardenal, Ana S., Carlos Aguilar-Paredes, Carol Galais, and Mario Pérez-Montoro. 2019. "Digital Technologies and Selective Exposure: How Choice and Filter Bubbles Shape News Media Exposure." *The International Journal of Press/Politics* 24:465–486.

Christner, Clara, Aleksandra Urman, Silke Adam, and Michaela Maier. 2021. "Automated Tracking Approaches for Studying Online Media Use: A Critical Review and Recommendations." *Communication Methods and Measures* 0:1–17.

Cid, Enric. 2018. "3 steps to adopt online behavioral data."

Dagdoug, Mehdi, Camelia Goga, and David Haziza. 2021. "Model-Assisted Estimation Through Random Forests in Finite Population Sampling." *Journal of the American Statistical Association* 0:1–18.

Dvir-Gvirsman, Shira, Yariv Tsfati, and Ericka Menchen-Trevino. 2016. "The extent and nature of ideological selective exposure online: Combining survey responses with actual web log data from the 2013 Israeli Elections." *New Media & Society* 18:857–877.

Grimmer, Justin, Margaret E. Roberts, and Brandon M. Stewart. 2021. "Machine Learning for Social Science: An Agnostic Approach." *Annual Review of Political Science* 24:395–419.

Groves, Robert M., Floyd J. Fowler, Mick P. Couper, James M. Lepkowski, Eleanor Singer, and Roger Tourangeau. 2009. "Survey Methodology, 2nd Edition." *Wiley series in survey methodology* .

Groves, Robert M. and Lars Lyberg. 2010. "Total Survey Error: Past, Present, and Future." *Public Opinion Quarterly* 74:849–879.

Groves, Robert M., Eleanor Singer, James M. Lepkowski, Steven G. Heeringa, and Duane F. Alwin. 2010. "Survey methodology." In *A Telescope on Society: Survey Research and Social Science at the University of Michigan and Beyond*.

Guess, Andrew, Brendan Nyhan, and Jason Reifler. 2018. "Selective Exposure to Misinformation: Evidence from the consumption of fake news during the 2016 U. S. presidential campaign." *European Research Council* .

Guess, Andrew M., Dominique Lockett, Benjamin Lyons, Jacob M. Montgomery, Brendan Nyhan, and Jason Reifler. 2020. ""Fake news" may have limited effects on political participation beyond increasing beliefs in false claims." *Harvard Kennedy School Misinformation Review* 1.

Haim, Mario, Johannes Breuer, and Sebastian Stier. 2021. "Do News Actually "Find Me"? Using Digital Behavioral Data to Study the News-Finds-Me Phenomenon." *Social Media + Society* 7:20563051211033820.

Harari, Gabriella M., Nicholas D. Lane, Rui Wang, Benjamin S. Crosier, Andrew T. Campbell, and Samuel D. Gosling. 2016. "Using Smartphones to Collect Behavioral Data in Psychological Science: Opportunities, Practical Considerations, and Challenges." *Perspectives on Psychological Science* 11:838–854. PMID: 27899727.

Hofman, Jake M., Duncan J. Watts, Susan Athey, Filiz Garip, Thomas L. Griffiths, Jon

Kleinberg, Helen Margetts, Sendhil Mullainathan, Matthew J. Salganik, Simine Vazire, Alessandro Vespignani, and Tal Yarkoni. 2021. "Integrating explanation and prediction in computational social science." *Nature* 595:181–188.

Hsieh, Yuli Patrick and Joe Murphy. 2017. "Total Twitter Error." In *Total Survey Error in Practice*.

Jürgens, Pascal, Birgit Stark, and Melanie Magin. 2019. "Two Half-Truths Make a Whole? On Bias in Self-Reports and Tracking Data." *Social Science Computer Review* p. 089443931983164.

Lavrakas, Paul. 2008. "Total Survey Error (TSE." In *Encyclopedia of Survey Research Methods*.

Lynn, Peter and Peter Lugtig. 2017. "Total survey error for longitudinal surveys." In *Total survey error in practice*, pp. 279–298. Wiley Sons.

Mercer, Andrew W., Frauke Kreuter, Scott Keeter, and Elizabeth A. Stuart. 2017. "Theory and Practice in Nonprobability Surveys: Parallels between Causal Inference and Survey Inference." *Public Opinion Quarterly* 81:250–271.

Northcutt, Curtis G., Anish Athalye, and Jonas Mueller. 2021. "Pervasive Label Errors in Test Sets Destabilize Machine Learning Benchmarks." .

Oberski, D. L., A. Kirchner, S. Eckman, and F. Kreuter. 2017. "Evaluating the Quality of Survey and Administrative Data with Generalized Multitrait-Multimethod Models." *Journal of the American Statistical Association* .

Ochoa, Carlos and Ezequiel Paura. 2018. "The Value of Machine Learning in Privacy: Results-oriented machine learning solution in securing PII data anonymisation." In *ESOMAR FUSION 2018: BIG DATA WORLD*.

Ochoa, Carlos and Melanie Revilla. 2021. "Willingness to participate in in-the-moment surveys triggered by online behaviours." In *RECSM Webinar*.

Ochoa, Carlos and Melanie Revilla. 2022. "Acceptance and coverage of fast invitation methods to in-the-moment surveys." *International Journal of Market Research* 0:14707853221085204.

Peterson, Erik and Emily Damm. 2019. "A Window To the Worlds: Americans' Exposure to Political News from Foreign Media Outlets."

Peterson, Erik, Sharad Goel, and Shanto Iyengar. 2018. "Echo chambers and partisan polarization: Evidence from the 2016 presidential campaign." *Unpublished manuscript. https://5harad. com/papers/selecfive-exposure. pdf* .

Pew Research Center. 2016. "Evaluating Online Nonprobability Surveys." Technical report.

Pew Research Center. 2020. "Measuring News Consumption in a Digital Era." Technical report.

Qiu, Junfei, Qihui Wu, Guoru Ding, Yuhua Xu, and Shuo Feng. 2016. "A survey of machine learning for big data processing." *EURASIP Journal on Advances in Signal Processing* 2016:67.

Ramirez, Christina M., Marisa A. Abrajano, and R. Michael Alvarez. 2019. "Using Machine

Learning to Uncover Hidden Heterogeneities in Survey Data." *Scientific Reports* 9:16061.

Revilla, Melanie. 2022. "How to enhance web survey data using metered, geolocation, visual and voice data?" *Survey Research Methods* 16:1–12.

Revilla, Melanie, Mick P. Couper, Ezequiel Paura, and Carlos Ochoa. 2021. "Willingness to Participate in a Metered Online Panel." *Field Methods* 33:202–216.

Revilla, Melanie, Carlos Ochoa, and Germán Loewe. 2017. "Using Passive Data From a Meter to Complement Survey Data in Order to Study Online Behavior." *Social Science Computer Review* 35:521–536.

Ricciato, Fabio, Albrecht Wirthmann, and Martina Hahn. 2020. "Trusted Smart Statistics: How new data will change official statistics." *Data amp; Policy* 2:e7.

Saris, Willem E. and Irmtraud N. Gallhofer. 2014. *Design, Evaluation, and Analysis of Questionnaires for Survey Research*. Hoboken, NJ, US: John Wiley Sons, Inc., second edi edition.

Saris, Willem E., Daniel Oberski, Melanie Revilla, Diana Zavalla, Laur Lilleoja, Irmtraud Gallhofer, and Tom Gruner. 2011. "The development of the program SQP 2.0 for the prediction of the quality of survey questions."

Schwemmer, Carsten, Carly Knight, Emily D. Bello-Pardo, Stan Oklobdzija, Martijn Schoonvelde, and Jeffrey W. Lockhart. 2020. "Diagnosing Gender Bias in Image Recognition Systems." *Socius* 6:2378023120967171.

Sen, Indira, Fabian Flöck, Katrin Weller, Bernd Weiß, and Claudia Wagner. 2021. "A Total Error Framework for Digital Traces of Human Behavior on Online Platforms." *Public Opinion Quarterly* 85:399–422.

Smith, Brianna, Scott Clifford, and Jennifer Jerit. 2020. "TRENDS: How Internet Search Undermines the Validity of Political Knowledge Measures." *Political Research Quarterly* 73:141–155.

Sturgis, Patrick and Rebekah Luff. 2021. "The demise of the survey? A research note on trends in the use of survey data in the social sciences, 1939 to 2015." *International Journal of Social Research Methodology* 24:691–696.

Torcal, Mariano, Emily Carty, Josep Maria Comellas, Oriol J. Bosch, Zoe Thomson, and Danilo Serani. 2023. "The dynamics of political and affective polarisation: Datasets for Spain, Portugal, Italy, Argentina, and Chile (2019-2022)." *Data in Brief* 48:109219.

Toth, Roland and Tatiana Trifonova. 2021. "Somebody's Watching Me: Smartphone Use Tracking and Reactivity." *Computers in Human Behavior Reports* 4:100142.

Unangst, Jennifer, Ashley E Amaya, Herschel L Sanders, Jennifer Howard, Abigail Ferrell, Sarita Karon, and Jill A Dever. 2019. "A Process for Decomposing Total Survey Error in Probability and Nonprobability Surveys: A Case Study Comparing Health Statistics in US Internet Panels." *Journal of Survey Statistics and Methodology* 8:62–88.

Wang, Angelina and Olga Russakovsky. 2021. "Directional Bias Amplification." In *Proceedings of the 38th International Conference on Machine Learning*, edited by Marina Meila and Tong Zhang, volume 139 of *Proceedings of Machine Learning Research*, pp. 10882–10893. PMLR.

Chapter 4

Uncovering digital trace data biases: tracking undercoverage in web tracking data

Oriol J. Bosch, Patrick Sturgis, Jouni Kuha, and Melanie Revilla

Abstract

In the digital age, understanding people's online behaviours is vital. Digital trace data has emerged as a popular alternative to surveys, many times hailed as the gold standard. This study critically assesses the use of web tracking data to study online media exposure. Specifically, we focus on a critical error source of this type of data, tracking undercoverage: researchers' failure to capture data from all the devices and browsers that individuals utilize to go online. Using data from Spain, Portugal, and Italy, we explore undercoverage in commercial online panels and simulate biases in online media exposure estimates. The paper shows that tracking undercoverage is highly prevalent when using commercial panels, with more than 70% of participants affected. In addition, the primary determinant of undercoverage is the type and number of devices employed for internet access, rather than individual characteristics and attitudes. Additionally, through a simulation study, it demonstrates that web tracking estimates, both univariate and multivariate, are often substantially biased due to tracking undercoverage. This represent the first empirical evidence demonstrating that web tracking data is, effectively, biased. Methodologically, the paper showcases how survey questions can be used as auxiliary information to identify and simulate web tracking errors.

Keywords:
Digital trace data · Web tracking data · Undercoverage · Bias · Media exposure · Monte Carlo simulation

1. Introduction

In the age of the internet, measuring how people behave and what they consume online is crucial for academic, policy, and commercial researchers. Whether they are studying the use and potential effects of fertility apps (Rampazzo et al., 2022); gender inequalities in access and participation in online platforms (Kashyap et al., 2020); or how filter bubbles shape news media exposure (Cardenal et al., 2019), high-quality measures of online behaviours are essential. Although respondent self-reports have traditionally been used as the main instrument to measure online behaviours (González-Bailón and Xenos, 2022), there have long been good reasons to doubt their accuracy (Parry et al., 2021). These relate, most notably, to social desirability bias and recall error, as well as the cognitive burden they place on respondents, which can have negative impacts on response rates and sample composition. Consequently, researchers have long been searching for alternative ways of measuring online behaviours.

In this context, the collection of *digital trace data* has become prominent in recent years. This type of data records the interactions of users with specific digital systems (Howison et al., 2011), such as online transaction systems, telecommunication networks, websites, social media platforms, smartphone apps, sensors built in wearable devices, and digital devices (Stier et al., 2019). Given the 'objective' and granular nature of these digital traces, some have advocated for the possibility of using them to enhance or substitute self-reports (Revilla, 2022). Indeed, recent studies have already begun to treat digital trace data as the de facto gold standard when measuring online behaviours (Araujo et al., 2017; Scharkow, 2016), with some authors recommending substituting survey self-reports for digital traces when measuring what people do and consume online (Konitzer et al., 2021).

Nonetheless, digital trace data is itself subject to a wide range of different errors, which may introduce bias to estimates and substantive conclusions (Bosch and Revilla, 2022b; Sen et al., 2021; Amaya et al., 2020). Probably the most concerning of these errors is *tracking undercoverage*. It occurs when researchers fail to capture data from certain digital systems that individuals utilize to engage in specific online behaviours. For instance, when investigating an individual's interaction with harmful content on social media platforms, researchers must gather data from all the various social media platforms and accounts that the person employs. Failure to track data from all relevant digital systems leads to incomplete measurements. This, in turn, can introduce bias to the estimates produced with digital trace data, if the observed behaviours differ from participant's true ones. This issue arises because digital trace data is generated at the level of the digital systems individuals employ, rather than at the individual level. Therefore, to obtain individual-level measurements, it is crucial to collect information from all pertinent digital systems and devices utilized by each respondent.

In this paper we focus on understanding the impact of tracking undercoverage for one of the most widely used approaches of collecting individual-level digital trace data: **web trackers**. These technologies, also known as meters (Revilla et al., 2021), can be installed on participants' browsing devices with their consent. Meters enable researchers to track the traces left by participants while interacting with their devices, such as visited URLs, apps, timestamps, and sometimes even HTML content. We consider the consequences of tracking undercoverage when studying **online media exposure**, a behaviour that is well-suited to measurement with web trackers.

The remainder of this article proceeds as follows: in the next section, we review the literature on measurement of media exposure and web trackers. In the third section, we set out how tracking undercoverage can introduce bias to survey estimates. Next, in the fourth section, we describe the data and variables used, and in the fifth section, we describe our analytical approach to estimating bias due to tracking undercoverage. In the final section we present the results and, in the seventh, we conclude.

2. Measuring online media exposure with web trackers

Survey self-reports, while widely used for studying media use, face significant and well-known challenges (Price and Zaller, 1993; Chang and Krosnick, 2003; Prior, 2009). Research has shown that participants tend to overstate their media exposure, due to the complexities of the recall task. This has negative consequences on the accuracy of media exposure measures, evidenced by the low levels of agreement between self-reported and more 'objective' measures of media exposure (Prior, 2009; Araujo et al., 2017; Parry et al., 2021; Scharkow, 2016).

To overcome these challenges, social scientists have made significant efforts to develop alternative approaches for measuring online media exposure that do not rely on participants' memory and can also collect more granular data in greater vol-

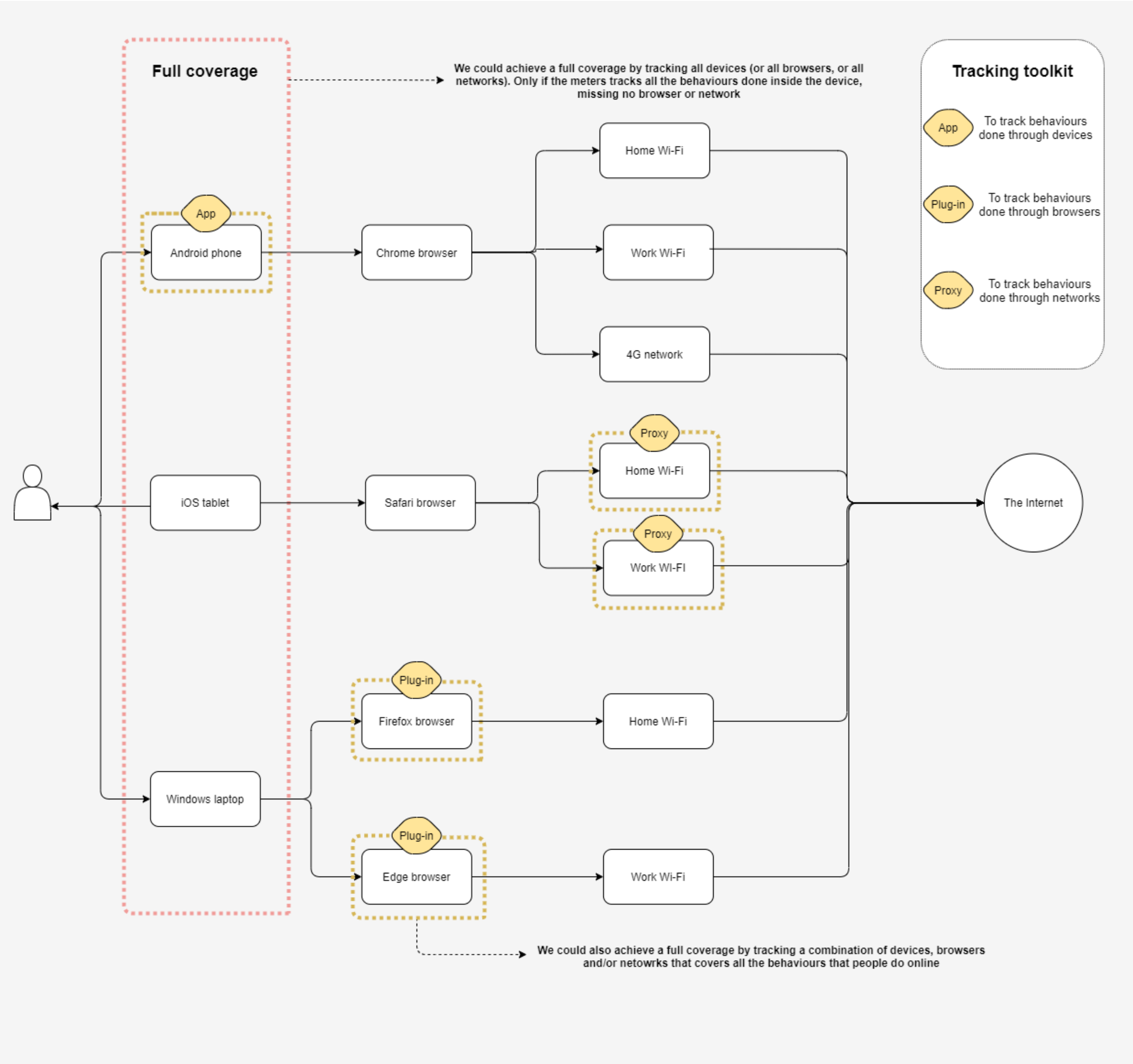
Full coverage
We could achieve a full coverage by tracking all devices (or all browsers, or all networks). Only if the meters tracks all the behaviours done inside the device, missing no browser or network
Tracking toolkit
App
To track behaviours done through devices
Plug-in
To track behaviours done through browsers
Proxy
To track behaviours done through networks
App
Android phone
Chrome browser
Home Wi-Fi
Work Wi-Fi
4G network
iOS tablet
Safari browser
Proxy
Home Wi-Fi
Proxy
Work WI-FI
The Internet
Plug-in
Firefox browser
Home Wi-Fi
Windows laptop
Plug-in
Edge browser
Work Wi-Fi
We could also achieve a full coverage by tracking a combination of devices, browsers and/or netowrks that covers all the behaviours that people do online

Figure 4.1: Diagram to illustrate the impact of device undercoverage

umes. Recent examples include public application programming interfaces (He and Tsvetkova, 2023) or data donations (Bosch et al., 2023). However, the most common approach is through use of web tracking technologies. Substantive researchers have used meters to quantify the prevalence of dubious media exposure during elections (Guess et al., 2020), the overlap in political media diets between partisans (Guess, 2021), and the extent to which online news environments are segmented by age groups (Mangold et al., 2021). Another salient area of inquiry is the degree to which social media and other sites serve as intermediaries for online media exposure (Cardenal et al., 2019; Jürgens and Stark, 2022; Scharkow et al., 2020; Stier et al., 2021).

However, web trackers are not immune to errors themselves. For researchers to obtain complete data on what an individual does online, they must track respondent behaviours across all relevant digital systems they use to connect to the Internet. How this is achieved, however, can vary a lot depending on the meters available as well as the number and type of devices used by respondents. Figure 1 provides an illustration of device undercoverage. Here, the target individual browses the Internet using several browsers (sometimes more than one within a device) and connecting though different networks. Collecting data from all devices can be achieved in different ways. Participants may be asked to install tracking apps in all their devices. Alternatively, if it is not possible for one or more devices, they can ask participants to install tracking plug-ins (i.e., VPNs) in each of the browsers they use on those untracked devices. Or a proxy can be manually configured to collect data at the network level. In practice, most tracking projects require a combination of different meters and tracking approaches, as exemplified in Figure 1. This is because there is rarely a one-size-fits-all tracking technology available: PCs are better tracked with browser plug-ins, Android devices with tracking apps, and iOS devices can (generally) only be tracked with proxies. However, achieving the goal of full coverage can be very challenging (Bosch and Revilla, 2022a,b). For instance, some participants are not willing to install all the required technologies on all their devices, or the meters used by researchers might not be installable in some types of devices/browsers.

Research has shown that between 53% (Spain, Revilla et al. (2021)) and 68% (United States, Pew Research Center (2020)) of participants in web tracking surveys do not have the meters installed in all the devices they use to go online. Hence, web tracking data likely only captures some but not all of the online behaviours. Nonetheless, so far, most research comparing self-reported and metered data measures of the same concepts has treated metered data as the gold standard, considering differences between these measures as attributable to errors in the self-reports. For example, research has shown that metered measures of Internet use and media exposure tend to be substantially lower compared to self-reports (Araujo et al., 2017; Scharkow, 2016), and correlate only modestly with digital trace data measures (Parry et al., 2021). Researchers have concluded that these differences derive from errors in the side of self-reports. For instance, (Parry et al., 2021, pp. 1541) concluded "self-report measures of media use may not be a valid stand-in for more objective measures." while, (Ernala et al., 2020, pp. 10) went as far as recommending "using logging applications rather than self-reports where feasible and appropriate, treating self-reports as noisy estimates rather than precise values." On the other hand, a Pew Research Center (2020) study found that participants with uncovered devices had a higher probability of a mismatch between self-reported and metered data measures (self-reported measures being higher than metered data) than those fully covered. This suggests that device undercoverage results in underestimation of online behaviours using metered data, an intuitively plausible expectation.

Nonetheless, much is still uncertain. Nothing is known about the characteristics of tracking undercoverage nor the mechanisms behind this phenomenon. Additionally, the size and direction of tracking undercoverage bias is still unknown. Considering this, the paper explores three research questions:

1. What is the prevalence of tracking undercoverage? (**RQ. 1**)
2. What characteristics are associated with tracking undercoverage? (**RQ. 2**)
3. To what extent does device undercoverage bias estimates of media exposure? (**RQ. 3**)

3. Defining tracking undercoverage bias

Here we provide a formal definition of tracking undercoverage and undercoverage bias. Consider respondents $i = 1, \ldots, n$ in a sample from a target population and three kinds of variable.

Variables that will be measured using web tracking data are denoted Y_{ijk} for individual i, variable $j = 1, \ldots, J$, and device $k = 1, \ldots, K$. We can, without loss of generality, assume that $Y_{ijk} \geq 0$, and take them to represent time/visits spent in a given period on specific internet activities on different devices. Suppose that in each case the variable of interest is the sum $Y_{ij} = \sum_{k=1}^{K} Y_{ijk}$ of the values across all the devices. An example of this kind of variable is the time an individual spends on a specific website, on each device (Y_{ijk}) and in total (Y_{ij}). Let Y_i=(Y_{i1},... ,Y_{iJ}) denote the vector of these variables for individual i. A vector of variables Z_i that will be measured in the survey part of the data collection. Examples include the respondent's level of education and self-reported level of political interest. And a vector of variables V_i which characterise the internet sources and activities under consideration.

Z_i and V_i can be combined with Y_{ij} to derive further variables related to internet behaviour. They could be obtained from external sources (in which case they do not depend on respondent i) or from the survey. An example of the former is the identification of a website as a news site vs. not, and an example of the latter the respondent's self-report of which news site they have most trust in.

Suppose that the parameter of interest is some function of (Y_i,Z_i,V_i) across the individuals in the population. For example, this could be the average daily time a person spends on the internet, the proportion of their internet time they spend on news sites, the average variance of political ideology scores of the news sites that they visit, or the correlation between the time a person spends on news sites and their self-reported voting behaviour in elections.

Define $d_{ik} = 1$ if individual i uses device k at all, and $d_{ik} = 0$ if they do not. Similarly, let $d^*_{ik} = 1$ if device k is recorded in web tracking data for individual i and $d^*_{ik} = 0$ if it is not, and let $e_{ik} = 0$ if $d^*_{ik} = d_{ik}$ and $e_{ik} = 1$ if $d^*_{ik} \neq d_{ik}$. There is undercoverage of device k for individual i if $e_{ik} = 1$, i.e. $d_{ik} = 1$ but $d^*_{ik} = 0$ (we assume that the false recording case $d_{ik} = 0$, $d^*_{ik} = 1$ does not occur).

The true value of Y_{ij} can be written as $Y_{ij} = \sum_{k=1}^{K} d_{ik}Y_{ijk}$ and its measured values from web tracking data as $Y^*_{ij} = \sum_{k=1}^{K} d^*_{ik}Y_{ijk}$. Undercoverage leads to measurement error in the measured values if $Y^*_{ij} - Y_{ij} = \sum_{k=1}^{K} e_{ik}Y_{ijk} \neq 0$. The expected size of this error depends on the probabilities of undercoverage $P(e_{ik} = 1)$, but also on the distribution of Y_{ijk} across the devices and on the correlations between e_{ik} and Y_{ijk}.

The measurement error can lead to *undercoverage bias* in estimates of the parameters of interest that use Y^*_{ij} instead of the true values Y_{ij}. The size of this bias will depend not only on the magnitude of the measurement errors but also on the definition of the parameter and the joint distribution of all the variables involved in it. Here it is worth noting, in particular, that the bias does not need to be toward zero, i.e., undercoverage does not need to lead to underestimation. Downward bias is inevitable only if the parameter is a simple total or mean, such as the average time spent on the internet, since then every missed device can only reduce the estimate. For other types of parameters, however, the undercoverage bias can be in any direction. For example, the average proportion of time that individuals spend on news sites will be underestimated, overestimated, or estimated with little bias, depending on whether and how the individual uses the devices that are included in tracking tend to differ from how they use the devices that are not included.

4. Data and variables

4.1. The TRI-POL dataset

We use data from the first wave of the TRI-POL project (Torcal et al., 2023), the goal of which is to understand whether and how online behaviours are related to affective polarisation across Southern European and Latin American countries (https://www.upf.edu/web/tri-pol)[1]. TRI-POL conducted a three-wave survey between September 2021 and March 2022. Questionnaire responses were matched at the individual level with metered data. Data were collected through the Netquest opt-in metered panels (https://www.netquest.com), which consist of individuals who have meter(s) already installed in their devices and who can also be contacted to conduct surveys. We can thus link these respondents' online behaviour with their questionnaire data. When the panellists join the metered panels, they must agree to install the meter on at least one device (PC, tablet, or smartphone), and they receive more incentives if they install it on more devices (up to a maximum of three). Here we use the data collected in Italy, Portugal, and Spain.

Cross quotas for age and gender, and quotas for educational level, and region were used in each country to ensure a sample matching on these variables to the general online populations. Survey questions were used to measure attitudinal and demographic variables, while metered data were used to measure variables related to the general Internet use as well as consumption of specific news media outlets, political news, and social media (see the TRI-POL data protocols in footnote 2 to check the specific URLs defined to measure these concepts). Metered data was collected

[1] More information about the data collection strategy of both survey and digital trace data can be found in the TRI-POL data protocols: https://osf.io/3t7jz/

for the 15 days prior to and following participants completing the questionnaire. The meter logged each URL accessed by the panellists, along with timestamps indicating the initial visit to the URL, and the duration in seconds during which the URL remained the active content within the browser, or in the case of mobile devices, on the smartphone screen. It is important to note that a URL or app was classified as 'active' when it was the foremost content displayed in the browser or on the device's screen. This definition excludes any other URLs or apps that might have been open in separate tabs or screens, as they were not considered active during this time frame. The duration of active engagement was computed as the elapsed time between the moment the URL or app first gained 'active' status within the browser or device and the point at which a different URL or app took over as the active content in the browser or device. A visit was defined as any opened URL/app lasting one second or more. Participants were tracked on iOS and Android mobile devices, and Windows and MAC computers, using the tracking solutions provided by Wakoopa (https://www.wakoopa.com/). Windows and MAC devices were tracked with desktop apps and/or web browser plug-ins, Android devices through apps and iOS devices through manually configured proxies. More information about the collectable data and the characteristics of each of the tracking technologies used can be found in Torcal et al. (2023).

Challenges were faced when filling some of the specific cross-quotas with participants from the metered panel. Hence, in some cases panellists were invited without a meter installed to fill some of the quotas. Thus, in total, for the first wave, 3,548 respondents completed the survey, but only 2,653 had the meter installed in at least one mobile (smartphone or tablet) or PC device: 993 in Spain, 818 in Portugal and 842 in Italy. For the analyses, we combine the samples from the three countries. No significant differences are observed between the full sample and the subsample of tracked participants, across a selection of demographic, political and technological variables (see Supplementary Material 1, i.e., SM 1). Nonetheless, to correct for any unobserved difference in the sample characteristics between the full sample of respondents, and those with the meter installed, estimates from the subsample of metered participants were weighted with inverse probability weights, computed using the random forest relative frequency method (Buskirk et al., 2015).

4.2. Identifying undercoverage and its characteristics

We first had to identify when a participant was undercovered, and through which devices. To do so, two pieces of information were needed: which sources were tracked, and which sources panellists used to go online. The first piece of information was obtained using paradata about the technology with which participants were being tracked, the type of device, the operating system, whether it was a tablet or smartphone and, for plug-ins, the browser in which they were installed[2]. The second piece of information was obtained by asking participants questions about which devices and browsers they used to access the internet during the 15 days before the start of the survey. SM 2 shows the wordings of these questions (English translations), as well as the paradata available. We were able to identify any mismatches between the self-reported devices used to go online, and the paradata of those that we tracked. With this, we were able to identify how many of devices were not covered, and which type of devices they were (e.g., Windows or MAC).

In terms of browser undercoverage, we did not have information about the number of browsers that participants used within each of their devices, but only the types of browsers (e.g., Chrome or Firefox) used on all their Windows PC, MAC, and Android devices (see SM 2 to see the exact question used). Hence, for fully undercovered panellists, it would not be possible to discern whether a browser is not covered because the panellist did not install it in a tracked device, or because it was installed on an untracked device. Conversely, for fully covered panellists, we know that all their devices have a tracking technology installed; if a browser is not tracked, it is because the technology installed in the device is not tracking that browser. Hence, for fully covered panellists we were able to identify the types of browsers that they were not covered, in general and within each type of device (e.g., Windows PC). No information was computed for undercovered panellists.

It should be noted, nonetheless, that this identification strategy is affected by errors, given that the self-reported measures of the number and types of devices and browsers used by participants can be affected by measurement errors.

[2] No information about the networks in which proxies were configured was available.

4.3. Predictors of tracking undercoverage

To assess which individual characteristics are associated with undercoverage, we conducted three logistic regressions predicting whether a participant was not fully tracked in terms of device (1= at least one device not tracked, 0 = fully covered). In the first model, we focus on the profile of the people more likely of being undercovered. Hence, as independent variables, we introduced sociodemographic information about participants' sex (male = 0, female = 1), education (0 = not completed high education, i.e., post-secondary education such as university or superior technical training, 1 = completed high education), age group (18-24 = 1, 25-34 = 2, 35-44 = 4, 45-54=5, +55 = 6), and country of residence. Also included were political variables measuring participants' political ideology (0 = left, 10 = right) and political interest (1= not at all, 2= a little, 3= a fair amount, 4= a lot), as well as a measure of panel loyalty (the number of years a participant had been part of the Netquest panels) a self-reported measure of participants' Internet use (as hours spent on the Internet on a typical day). The second model introduced variables relating to the devices that people use. Therefore, as predictors, we introduced variables for the self-reported number of Windows PCs, MACs, Android, and iOS devices that participants used to go online during the 15 days prior to the survey. We also included dummy variables for whether the participants used both PCs and mobile devices to go online, or only mobile or PCs. In Model 3, we combined all predictors.

5. Simulated estimates of undercoverage bias

The bias introduced by device undercoverage cannot be directly quantified, since data that would be collected from the noncovered devices is, by definition, not observed. Hence, to assess the likely extent of this bias, we developed a Monte Carlo simulation to examine the magnitude of bias introduced by device undercoverage. To do so, we used data from the subset of 688 participants identified as fully covered in terms of device (i.e., match between self-reports and paradata on the number and type of devices used), considering that they are the only ones for which we can observe their complete online behaviours in terms of device. We take this to be the true complete sample of n respondents in the simulation. No significant difference is found between this sample and the full sample of participants in a selection of sociodemographic, political, and technological variables, except for the average number of devices (see

SM 3 for the results).

Table 4.1: Scenarios for the simulations

Scenario	**P(PC undercoverage)**	**P(Mobile undercoverage)**
1	0.25	0.0
2	0.50	0.0
3	0.75	0.0
4	0.0	0.25
5	0.0	0.50
6	0.0	0.75
7	0.25	0.25
8	0.25	0.50
9	0.25	0.75
10	0.50	0.25
11	0.50	0.50
12	0.75	0.25
13*	0.33	0.33

Note: *Scenario 13 represents the actual undercoverage in the sample

For this group of fully covered participants, we generated simulated observed samples by setting one or more of their devices to be non-covered and the metered data for these devices to be missing (i.e., setting d^*_{ij} to 0, in the notation of Section 3). Ideally, we would do this at the level of individual devices. However, we did not have meter data for respondents at the level of devices but only aggregated at the level of PC and mobile devices. Hence, in our simulation we consider all PCs together as one device ($k = 1$) and all mobile devices as one device ($k = 2$). We defined 13 simulation scenarios which vary in the probability of participants having no PC or mobile device covered ($P(d^*_{ij} = 0)$). These probabilities were independent for different participants. For scenarios simulating both PC and mobile undercoverage, this independence was relaxed to constrain the simulations to not allow both PC and mobile undercoverage for the same participant. Hence, the simulation took two steps: in the first step, all participants had a random chance to be selected to have their PCs untracked. In the second step, a subsample of those fully tracked on PC were randomly selected to have their mobile devices untracked. Additionally, fully covered participants using only PCs or mobile devices were kept in the sample, but were not affected by undercoverage, given that this would have resulted in them having no device covered whatsoever, an impossible scenario. Table 1 presents the details for each scenario.

For each scenario, we created 1,000 random allocations. In each of them, devices were set to be uncovered with the probabilities shown in Table 1. Estimates of several parameters of interest, as described below, were calculated for each such dataset[3]. Their average over the 1,000 simulations is the expected value of the undercovered estimate in each setting, and the difference between it and the fully covered estimate (which uses all the data) estimates the undercoverage bias. Monte Carlo standard errors were computed using the R “rsimsum” package (White, 2010; Gasparini, 2018).

Simulations were conducted for a range of univariate and multivariate statistics selected based on the conceptualisation of undercoverage bias presented in section 3. The univariate statistics were the following:

- **Average time spent on the Internet.** This measure captures the time spent by each participant on any URL and online app (Araujo et al., 2017).
- **Average time spent on Social Network Sites (SNSs).** This corresponds to the time spent on any URL or app identified as being a SNS (i.e., Facebook, Instagram, Snapchat, TikTok, Twitter, WhatsApp, Messenger, YouTube) (Scharkow et al., 2020).
- **Proportion of non-users of online news.** We replicate Reiss (2022) approach to measure news avoiders. Non-users were defined as those who during the period of one month never visited an URL or app defined as “news” (Palmer et al., 2020). In our case, we only consider the consumption of written “news.”
- **Total number of media consumed.** We replicated Padró-Solanet and Balcells (2022) and computed a statistic showing the total number of media consumed by participants (during the first wave), to measure the variety of their media diet.
- **Proportion of Internet time spent on SNSs.** We computed a measure of importance of SNSs consumption over all the time participants spend on the Internet: ((Time on SNSs)/(Time on the Internet)×100).

[3] Weights applied to compute these estimates. The weights bring closer the already very close sample of fully covered participants and the full sample, on a selection of sociodemographic, political, and technological variables.

- **Proportion of Internet time spent on news media outlets.** We computed a measure of importance of news media consumption over all the time participants spend on the Internet: ((Time on news media outlets)/(Time on the Internet)×100).

For multivariate statistics, we focused on the following five statistics:

- **Correlation between the average time spent on SNSs and trust in SNSs.** We computed the Spearman[4] correlation between trust in SNSs (0 to 10 scale, 0 being "I don't trust it at all", and 10 "Completely trust"), and the average time spent on SNSs during the month of tracking.
- **Association between the trust in news and news avoidance.** We ran a logit regression with news avoidance as the dependant variable, trust in news as the main independent variable (0 to 10 scale, 0 being "I don't trust it at all", and 10 "Completely trust"), and several common control variables (age, gender, higher education, left-right self-placement, and country).
- **Association between the total number of media consumed and ideological extremism.** Similarly as Padró-Solanet and Balcells (2022), we ran an OLS regression with a measure of ideological extremism as the dependent variable , the total number of media as the main independent variable, and several common control variables (age, gender, higher education, political interest, left-right self-placement, and country).
- **Correlation between age and Instagram use.** We computed the Spearman correlation between the age of the participant (continuous), and the average time spent on Instagram during the month of tracking.
- **Correlation between the average time spend on SNSs and on news sites.** We computed the Spearman correlation between the average time spent on SNSs during the month of tracking, and the average time spent on news media sites.

[4] Given the zero-inflated nature of the web tracking measures used, we use Spearman instead of Pearson. Spearman correlation is non-parametric and does not rely on the distributional assumptions of the data that Pearson does.

Table 4.2: Proportion of participants fully covered, and median number of untracked devices, in general and per specific sub-groups of participants

Coverage	**n**	**% Fully Covered**	**Median Number of Untracked Devices**
General Coverage			
All Participants	2653	26	2
Participants who reported using…			
1 device	207	100	0
2 devices	1103	34	1
3 devices	611	13	2
4 devices	305	1	3
+5 devices	416	0	6
Only PC	66	77	1
Only Mobile	264	66	1
Both PC and Mobile	2312	20	2
Device Specific Coverage†			
Participants who reported using…			
PC			
Any type	2379	47	1
Windows	2305	49	1
MAC	302	27	1
Mobile			
Any type	2577	41	1
Android	2340	52	1
iOS	782	10	1

Note: † Values for the device specific undercoverage are computed over all members of those subsamples. Therefore, the 49% fully covered of “Windows” means that, out of all Windows users, 49% have all their self-reported Windows devices covered.

6. Results

6.1. The prevalence and characteristics of tracking undercoverage

Table 2 presents the proportion of participants that had all their devices covered, and the median number of untracked devices for those who were not fully tracked. Results are additionally presented for the number and types of devices used. Only 26% had all reported devices tracked, which is very similar to the 28% of fully tracked participants that the Pew Research Center (2020) found in a probability-based sample in the United States. Of those who were not fully tracked, the median number of untracked devices was 2.

Table 2 also shows that the prevalence of device undercoverage differs depending on the number and types of devices of the participant. The more reported devices, the higher the proportion of participants not fully tracked and the median number of untracked devices. Specifically, while 34% of individuals using two devices were fully tracked, this number drops to 1% and 0% for those using four and five or more devices, respectively. There are also clear differences between those who use only PCs and mobiles devices, and those who use both, with the latter group yielding a three-times lower proportion of fully covered participants.

These estimates also show notable differences in the prevalence of undercoverage depending on the types of devices that participants use. For PCs, while 49% of those using a Windows PC had all their Windows PCs tracked, only 27% of MAC users had all their devices tracked. Similarly for mobile, while 52% of Android users had all their devices tracked, only 10% of iOS users had all their devices tracked. Therefore, tracking undercoverage is more prevalent for Apple devices, especially iPads and iPhones. Indeed, 85% of participants with at least one iOS mobile device had none of them tracked. This means that, for 29% of the participants, almost everything that they did through their iOS devices was missed. This is likely to be because Netquest, the panel provider, tracks these devices using proxies, which require participants to manually configure them according to a complex and burdensome process.

Even if all devices are covered, we might still miss people's behaviours if we do not track the browsers that they use within those devices. Table 3 presents the proportion of participants that, having all their devices tracked, also have full browser coverage. Results are additionally presented for different subgroups, depending on the types of devices and browsers that they use to go online.

Table 4.3: Browser coverage conditional on device full coverage

Coverage	**n**	**% Fully Covered Browser**
Device Specific Browser Coverage		
Fully covered in terms of...*		
All devices	688	38
Windows PC	1109	51
MAC	83	48
Android	1224	91
iOS	79	100
Browser Specific Coverage		
Fully covered using...†		
Internet Explorer	130	0
Chrome	401	97
Firefox	124	26
Safari	7	0
Other	132	15

Note: *Values for the device specific browser undercoverage are computed over those participants that are fully covered in terms of the specific devices listed. Therefore, the 91% fully covered of "Android" means that, out of all those participants that have all their Android devices covered, 91% have all their browsers covered.† Values for the specific types of browsers are computed over all those fully covered participants that self-reported using the listed browsers. Therefore, the 97% fully covered of "Chrome" means that, out of all fully covered (in terms of device) participants using Chrome, 97% have all their self-reported Chrome covered.

Only 38% of those who were fully covered in terms of device, also had all browsers tracked. This shows that even if we track people on all the devices that they use to go online, there is still a high rate of undercoverage. Focusing on the different subgroups by device used, we see that browser undercoverage is mainly a phenomenon affecting PCs. Of those with all Android and/or iOS devices covered, respectively 91% and 100% of them had all browsers within those devices covered. Conversely, these numbers go down to 51% and 48% for fully covered Windows and MAC users. This is to be expected; while mobile tracking technologies tend to track all browsers used within a device[5], most technologies used to track PCs have to be installed as plug-ins in each browser that participants use. Additionally, it is more common to have multiple browsers on a PC than a mobile device. This might result from participants not installing the plug-ins in all the browsers they use to go online. Another potential hypothesis might be that participants satisfice by installing trackers on browsers they do not use, to get the incentives without having their behaviours tracked.

Furthermore, Table 3 shows that the prevalence of browser undercoverage is highly dependent on the browsers that people use. While we observe very high coverage rates for Chrome (97%), all other browsers are substantially lower (0 to 26%). In particular, for Internet Explorer or Safari users, none of those participants had all those specific browsers covered. This points to technological limitations on the side of the panel provider: if most panellists are tracked with web browser plug-ins, and these are only available for Chrome and Firefox, almost all behaviours done through other types of browsers will be missed.

Table 4 presents the results of three logistic regression predicting whether an individual was not fully tracked in terms of device (1= at least one device not tracked, 0 = fully covered). Results show that the more longstanding the panellists and people who completed tertiary education have a lower probability of being undercovered, Model 1 also shows significant differences in terms of the country of residence, with people residing in Italy and Portugal being more likely to not be fully covered than those from Spain. In Model 2 we observe that for each additional device, the odds of being undercovered increase by a factor of 4.7 (Android), 6.9 (iOS), 7.1 (Windows PC), and 10.1 (MAC). The odds of a participant who uses only PCs being undercovered are 83% lower than for those using both PCs and mobile devices.

Model 3 shows that education and country of residence of individuals is no longer significant (and the odds ratios are smaller), suggesting that the main driver behind those effects is that the number and types of devices used varies across educational levels and country groups. Additionally, we find that women are slightly more likely to be undercovered than men. All in all, these models seem to suggest that, although there might be some slight differences in the demographics of those being

[5] This is, however, not always the case. Depending on the OS version, and the tracking technology used, uncommon browsers might not be trackable even with a tracking app. This has been accounted in our approach to identify browser undercoverage. Hence, why there is some proportion of undercoverage for Android devices.

fully tacked and those not, the biggest driver behind someone being undercovered is the type and number of devices that they use to go online.

Table 4.4: Characteristics associated with tracking undercoverage

Variables	Model 1		Model 2		Model 3	
	Coefficient	SE	Coefficient	SE	Coefficient	SE
Type of user						
Mobile only			.64	.17	0.62	0.18
PC only			.17***	.07	.19***	.09
Self-reported number of...						
Windows PC			7.07***	1.45	7.37***	1.56
MAC			10.13***	2.92	11.93***	3.69
Android			4.72***	.63	4.31***	.60
iOS			6.88***	1.03	7.13***	1.16
Internet consumption †	1.02	.02			0.97	.02
Years in panel	.96**	.01			.93***	.00
Ideology	1.01	.02			1.02	.02
Political interest	1.06	.05			1.01	.06
Age						
25-34	.95	.19			1.08	.25
35-44	.85	.16			0.95	.22
45-54	.81	.15			0.87	.20
55+	.81	.15			1.04	.24
Female	.86	.07			1.25*	.13
Tertiary education	1.32**	.12			0.98	.11
Country						
Portugal	1.27*	.15			1.12	.16
Italy	1.33**	.14			1.26	.16
Constant	2.99**	.76	.05***	.01	.07***	0.1
AIC	3307.1		2542.3		2263.7	
n	2374		2641		2366	

Note: Coefficients reported in odds ratios. $*p < .05$, $**p < .01$, $***p < .001$. † In hours

6.2. The bias introduced by tracking undercoverage

Table 5 presents the results of the Monte Carlo simulations for estimating biases due to device undercoverage. The top row (in bold) shows the 'true value' which is taken as the estimate from the subset of fully covered participants, and the following rows show the estimates under the different undercoverage scenarios.

Table 4.5: Average estimates of univariate statistics under different scenarios of simulated undercoverage.

Undercoverage Scenarios		**Univariate Estimates**					
P(PC	**P(Mobile**	**Time**	**Time**	**% News**	**Num. Media**	**% Internet Time**	**% Internet Time**
Undercoverage)	**Undercoverage)**	**Internet**	**SNSs**	**Avoiders**	**Exposed**	**SNS**	**News**
0.00	0.00	221'	64'	17%	4.3	29%	1.6%
(Full Coverage)		(.18)	(.08)	(.01)	(.29)	(1.22)	(.23)
0.25	0.00	206	61	22	3.4	30	1.4
		(.06)	(.02)	(.02)	(.01)	(.01)	(.00)
0.50	0.00	191	58	27	3.0	31	1.3
		(.06)	(.02)	(.03)	(.01)	(.01)	(.00)
0.75	0.00	176	55	32	2.7	32	1.1
		(.06)	(.02)	(.02)	(.01)	(.01)	(.00)
0.00	0.25	188	53	23	3.5	25	1.8
		(.13)	(.05)	(.03)	(.01)	(.01)	(.00)
0.00	0.50	153	42	28	3.2	23	2.0
		(.16)	(.06)	(.03)	(.01)	(.01)	(.00)
0.00	0.75	119	31	33	2.8	20	2.2
		(.13)	(.05)	(.03)	(.01)	(.01)	(.00)
0.25	0.25	172	50	27	2.4	27	1.6
		(.12)	(.04)	(.04)	(.01)	(.02)	(.00)
0.25	0.50	138	39	33	.9	24	1.9
		(.13)	(.05)	(.04)	(.01)	(.02)	(.00)
0.25	0.75	104	28	38	.2	21	2.1
		(.13)	(.05)	(.04)	(.01)	(.02)	(.00)
0.50	0.25	157	47	32	1.5	28	1.5
		(.13)	(.05)	(.04)	(.01)	(.02)	(.00)
0.50	0.50	123	36	37	1.0	25	1.7
		(.12)	(.04)	(.04)	(.01)	(.02)	(.00)
0.75	0.25	142	44	37	.5	29	1.3
		(.16)	(.05)	(.04)	(.01)	(.02)	(.00)
0.33*	0.33*	157	45	31	2.0	26	1.7
(Sample Undercoverage)		(.14)	(.02)	(.04)	(.01)	(.02)	(.00)

Note: Averages computed over the 1,000 simulated scenarios, with the exception of the full coverage scenario, which represents the sample estimate. Empirical Monte Carlo standard errors in brackets.

We can see that tracking undercoverage results in biases for most of the univariate statistics considered. In many instances, these biases are of a substantial magnitude, both in absolute and relative terms (see Figure 2). In absolute terms, the direction of the effects is as expected: while device undercoverage reduces the average time participants spend on the Internet (by between 15 to 117 minutes) and SNSs (by 3 to 36 minutes), it increases the estimated proportion identified as news avoiders (by 5 to 21 percentage points). Furthermore, undercoverage reduces the average number of media outlets exposed to (by 0.9 to 4 fewer media), which implies that people do not have consistent media diets across devices.

The effect of tracking undercoverage is less pronounced for the percentage of time spent on SNSs and on news. In these cases, the bias would be introduced if tracking undercoverage had a different effect on the numerator (time on SNSs / news) than the denominator (time on the Internet). What we observe is that, while undercovered estimates deviate from the full coverage, the deviations are smaller and more varied. For instance, for percentage of time spent on SNSs, PC undercoverage alone inflates the estimates by around 1 to 3 percentage points, while mobile undercoverage reduces it by 4 to 9 percentage points.

If we focus on the relative size of these biases (see Figure 2), undercoverage is responsible for a relative overestimation of 29 to 123% of the participants identified as news avoiders, and an underestimation of 21 to 93% of the number of media that people consume on average. Figure 2 also shows that undercoverage leads to an underestimation of 5 to 53% of the average time spent on the Internet and SNSs. Although the relative bias for the percentage of time spent on SNSs and on news is smaller than for the other statistics, in some scenarios these are still substantial, with the estimated percentage of Internet time spent on SNSs underestimated by up to 31%, while the percentage of Internet time spent on news can be both under- and overestimated by up to 31%.

Additionally, as we would expect, there is an association between the extent of undercoverage and the size of the bias. In the scenarios with only PC or mobile undercoverage, the scenarios with 50% and the 75% undercoverage reveal biases two and three times the size of the scenario with 25% of the sample, respectively. In general, mobile undercoverage introduces more bias than PC undercoverage. This can be observed both in the scenarios with mobile and PC undercoverage alone, and when combined.

Table 5 also shows the estimated bias that undercoverage would introduce at the observed undercoverage level of TRI-POL dataset (around 33% for both PCs and mobile devices), which is expected to be very similar to what other studies have experienced. This shows that the observed time spent on the Internet and on SNSs is underestimated by 29 and 30%, or between 64' and 19' lower than what we would observe without undercoverage, respectively (221' vs 157' / 64' vs 45'). In addition, undercoverage is estimated to lead the observed proportion of news avoiders to be overestimated by 82%, or 14 percentage points higher than what it would be without undercoverage (17% vs 31%). The number of media outlets that we observe people being exposed to is less than half of what we would observe with full coverage (4.3 vs 2.1). On the other hand, at this level of undercoverage both the estimated percentage of time spent on SNSs and on news would be very similar to full coverage (29% vs 26% / 1.6% vs 1.7%).

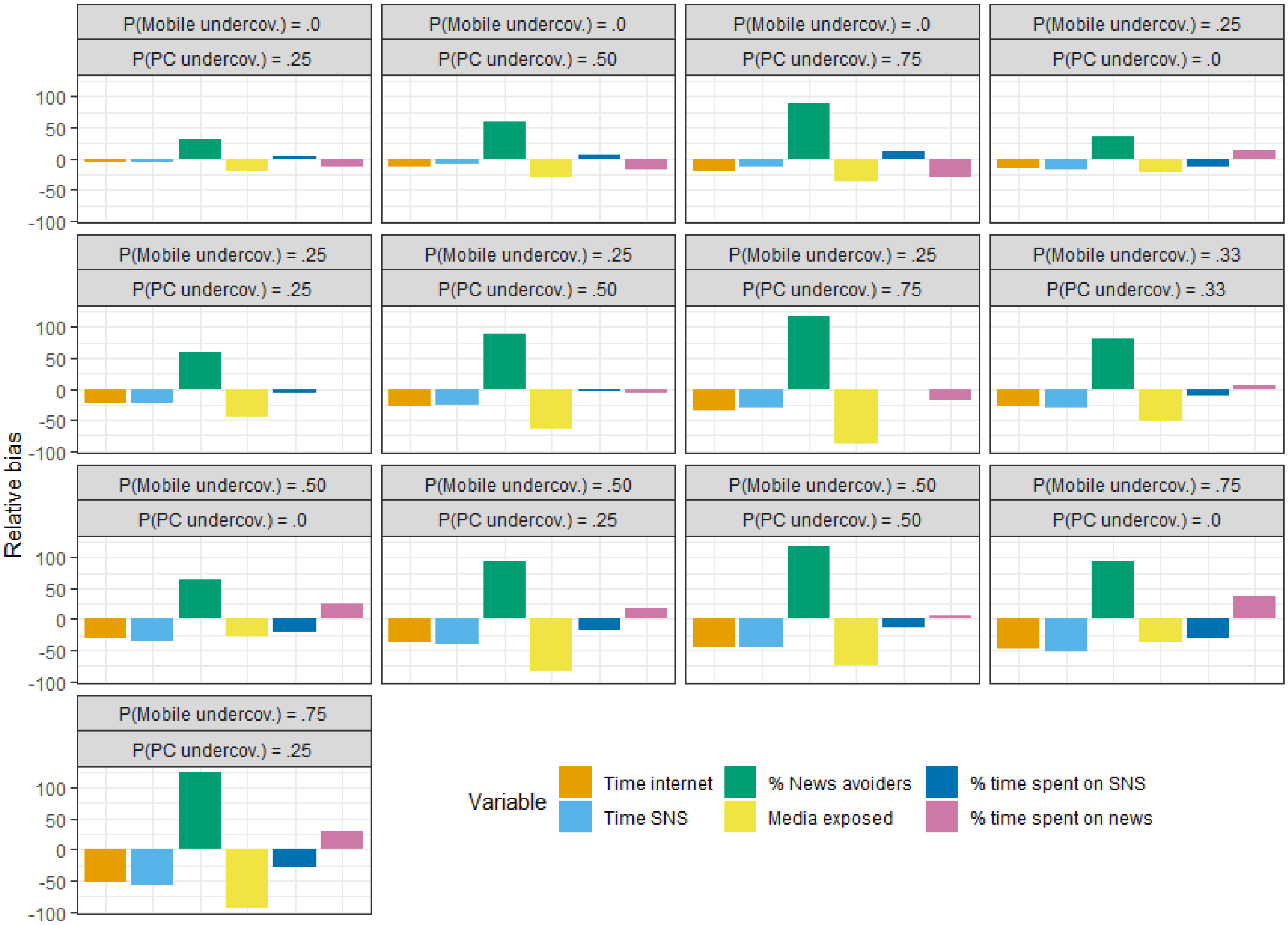

Figure 4.2: Estimates of relative bias for the univariate estimates

Table 6 presents the Monte Carlo simiulation estimates of bias for the set of multivariate statistics, again the ‘true scores’ are in bold in the first row of the table. This shows that, overall, the biases of tracking undercoverage are considerably smaller than for the univariate quantities but that there is wide variability between them. For three of the statistics the biases are close to zero. Two of them are severely affected by tracking undercoverage: the correlations between time spent on the Internet and both age and time spent on SNSs and news media outlets.

Table 4.6: Average estimates of univariate statistics under different scenarios of simulated undercoverage.

Undercoverage Scenarios		**Univariate Estimates**				
P(PC Undercoverage)	**P(Mobile Undercoverage)**	**Time SNSs ˜ Trust SNSs**	**News Avoidance ˜ Trust news**	**% Polarization ˜ No media Consumed**	**Time Instagram ˜ Age**	**% Time SNSs ˜ Time News**
0.00	0.00	.03	.89	-.01	.41	.16
(Full Coverage)		(.02)	(.02)	(.02)	(.02)	(.02)
0.25	0.00	.03	.92	-.01	-.41	.19
		(.00)	(.00)	(.00)	(.00)	(.00)
0.50	0.00	.03	.93	-.01	-.40	.22
		(.00)	(.00)	(.00)	(.00)	(.00)
0.75	0.00	.02	.95	-.01	-.39	.24
		(.00)	(.00)	(.00)	(.00)	(.00)
0.00	0.25	.02	.90	-.01	-.32	.27
		(.00)	(.00)	(.00)	(.00)	(.00)
0.00	0.50	.00	.91	-.01	-.24	.35
		(.00	(.00)	(.00)	(.00)	(.00)
0.00	0.75	-.02	.91	-.01	-.17	.40
		(.00)	(.00)	(.00)	(.00)	(.00)
0.25	0.25	.01	.92	-.01	-.31	.28
		(.00)	(.00)	(.00)	(.00)	(.00)
0.25	0.50	-.01	.92	.02	-.23	.33
		(.00)	(.00)	(.00)	(.00)	(.00)
0.25	0.75	-.03	.92	.03	-.17	.35
		(.00)	(.00)	(.00)	(.00)	(.00)
0.50	0.25	.01	.93	.00	-.31	.27
		(.00)	(.00)	(.00)	(.00)	(.00)
0.50	0.50	-.01	.93	.00	-.23	.29
		(.00)	(.00)	(.00)	(.00)	(.00)
0.75	0.25	.00	.94	.01	-.30	.26
		(.00)	(.00)	(.00)	(.00)	(.00)
0.33*	0.33*	.01	.93	-.01	-.29	.29
(Sample Undercoverage)		(.00)	(.00)	(.00)	(.00)	(.00)

Note: Averages computed over the 1,000 simulated scenarios, with the exception of the full coverage scenario, which represent the sample estimate. Empirical Monte Carlo standard errors in brackets. "Time SNSs ˜ trust SNSs", "Time Instagram ˜ Age" and "Time SNSs ˜ Time News" expressed as correlation coefficient, "News avoidance ˜ trust news" as odds ratios, and "Polarization ˜ No media consumed" as non-standardised regression coefficients.

Focusing on the association between the time spent on Instagram and age, device undercoverage reduces the estimated correlation by 24% (-.41 vs -.29), a difference with potentially important substantive implications. For the correlation between the time spent on SNSs and on news media outlets, we observe the opposite effect: device undercoverage now inflates the correlation, by between 3 and 24%;

the estimated correlation increases from .16 to .29. This might be due to the difference in the strength of the associations in the full coverage scenario: while the first three statistics show small and weak associations under full coverage, the associations showed for the other two statistics are substantially higher.

7. Discussion

There has been much excitement about the possibility of obtaining unbiased measures of online behaviours using web tracking data, and digital trace data in general. This has led to an increasing use of this type of data in the social sciences, under the assumption that they fix the well-established problems of self-reports. In this paper, however, we have shown that web tracking data suffers from severe limitations of its own. Specifically, tracking undercoverage is very prevalent in web tracking research and this can produce substantial biases across a broad range of univariate and multivariate statistics that are of substantive interest to scholars of media, communication, and public opinion.

Overall, we found that tracking undercoverage is highly prevalent in a commercial panel in the three studied countries: 74% of participants had at least one of the devices they use to go online not tracked. These results are in line with the 68% of device undercoverage found by the Pew Research Center (2020) in a US probability-panel. Additionally, of those fully tracked, 62% had at least one web browser uncovered, showing that undercoverage is multi-layered: even when we track all participant devices, we can still fail to observe online behaviours within those devices. For the vast majority of participants in this study we could not track at least some of what they did online during the period of observation.

Besides describing the prevalence of tracking undercoverage, our results also identify the main issues faced when trying to track online behaviours, and potential approaches to reduce them. Our data shows that, at least in the context of our panel company and period studied, there are specific difficulties when tracking devices, and browsers other than Chrome and Firefox. Indeed, between 90% of those participants that reported browsing online with iPhones and/or iPads had at least one of those not tracked (a large majority had all of them untracked), and none of the self-reported Internet Explorer and Safari browsers were tracked. Considering that our tracking approach is based on current standard research practice (provided by Wakoopa), these results highlight the technological limitations that the field still faces when tracking anything apart from Android and Windows devices, and mainstream browsers. Our analysis also shows that the main determinant of tracking undercoverage is the number, type, and combination of devices that participants use, pointing to the importance of creating tailored recruitment approaches based on participants' self-reported information on what devices and browsers they use to go online. This has several practical implications. First, if the behaviours that people do online vary across devices and browsers, web tracking data will systematically miss some behaviours more than others. Hence, the size of the biases will differ across statistics of interest. Second, if the probability of using of these devices and browsers varies across key demographics, tracking undercoverage will introduce differential errors.

The results of our Monte Carlo simulations show that most statistics derived with web tracking data are highly likely to be biased due to tracking undercoverage. Specifically, when simulating the bias at the level found in the TRI-POL dataset, tracking undercoverage led to very large biases across a broad range of quantities of substantive interest. For example, the estimated proportion of participants identified as not consuming news almost doubled from 17% to 31% and the number of media outlets exposed dropped by 50% under this scenario. Our results confirm that the higher the level of undercoverage, the larger the bias introduced. The type of devices missed is also important; mobile undercoverage leads to higher biases than PC undercoverage. There are two potential and complementary explanations for this. On the one hand, internet consumption is more common through mobile devices (StatCounter, 2017). On the other hand, most of the online behaviours that we have tested are done more often through mobile devices (Festic et al., 2021). The bias due to device undercoverage also varies depending on the statistic of interest. For univariate statistics, undercoverage underestimates count variables (e.g., counting number of visits, time, or media), but will under- or overestimate proportions engaging in specific behaviours when the underlying variables used are simple counts (e.g., people avoiding news). In terms of multivariate statistics, we observe big variations depending on the associations tested. Although just a speculation, results suggests that when the true association is small, undercoverage might be irrelevant. Nonetheless, for associations that are expected to be more substantial, our results propose that tracking undercoverage might heavily deviate the estimated correlation coefficients from their true value.

There are some limitations in our research design that should be acknowledged. First, participants were recruited using an opt-in online panel of already tracked participants. Although this is the most common approach in the literature, it is unclear whether these results would replicate when conducted on another opt-in panel using a different recruiting and tracking approach, as well as on a probability-based sample of less experienced respondents. Additionally, given that participants were already tracked by the fieldwork company, we had no control over the tracking technologies used or the recruitment techniques, preventing us from testing approaches to reduce the prevalence of undercoverage. Second, tracking undercoverage has been identified combining paradata and self-reports, the latter being sensitive to measurement errors. Since participants might have trouble properly recalling the devices/browsers used to go online, the estimates of undercoverage cannot be themselves expected to be free of errors. Furthermore, the results from the simulations must be understood together with several caveats: simulations are based on a small subsample of fully covered participants. Although no relevant differences are observed between the subsample used and the full sample of participants, and the weighting approach should reduce some of the potential problems introduced by this, it is to expect that these results to be biased on the side of representation, as well as noisier than desired.

Although we would ideally have simulated undercoverage at the level of individual devices and browsers, we were limited by the granularity of our dataset, forcing us to focus only on the effect of full mobile/PC undercoverage (i.e., having all mobile or PC devices undercoverage). Considering that partial undercoverage (i.e., some but not all mobile/PC devices undercovered) is more common, the simulated bias is expected to be lower than realistically expected. Furthermore, the mechanism leading to undercoverage in our simulation is at random (everyone has the same probability). This might not be realistic in real life given that some people being more prone to be undercovered because of some of their demographics. Additionally, undercoverage could also be associated with how individuals use specific devices or browsers, leading to missingness not at random. This could exacerbate the biases that might be affecting real web tracking studies. Future studies should consider simulating undercoverage under not-random scenarios, where undercoverage is dependant on the characteristics of the participants and their devices.

Our results have implications for improving best practice in this area. First, fieldwork companies offering panels of already tracked individuals should increase efforts to minimise device undercoverage. This can be achieved by increasing the resources allocated to making sure that participants install tracking technologies in all their devices/browsers, and they keep them installed and updated. Device coverage can also be increased by improving the capabilities of tracking beyond Windows, Android, Chrome, and Firefox. In parallel, they should be more transparent to their clients, disclosing the approaches used to assure full coverage, and up to date information of the level of undercoverage of each of their panellists.

Tech companies can also improve their practices. Currently, some of the challenges faced when tracking specific devices (e.g., iOS) can be linked to tech companies' terms of services, which limit the information that apps can track from users' devices. Although this can be beneficial in many instances, these companies need to acknowledge that if their products and services might lead to negative effects to their users, it should be possible for individuals to willingly access and share this information with academics for research purposes. Hence, research-based tracking technologies should not be treated in the same way than those that exploit personal data for commercial purposes.

Finally, researchers using metered data (and other digital trace data by extension) should not assume that this data is without error. In recent years there has been a great deal of excitement about the 'zero measurement error' of meters and how they can solve longstanding problems of self-reports in areas of research such as media consumption. Our research suggests treating meters as a gold standard is itself highly problematic. Therefore, researchers using web tracking data reflect these limitations in their analysis plans and reporting practices. Best practice when using metered data must involve:

1. Identify what participants are affected by undercoverage, and to what extent. Our approach can be used as a template to replicate or build on.

2. Report the proportion of people affected by tracking undercoverage, and some information about the characteristics. This should be similar as to what is done with nonresponse or drop-out rates in survey research, allowing readers and secondary data users to understand the quality of the data. The TRI-POL data protocols can be used as a good example of transparency (Torcal et al., 2023).

3. When possible, researchers should try to simulate the extent and ways in which tracking undercoverage might bias their results, in a similar way as robustness

checks are conducted when using survey data.

4. Potentially, researchers could try to develop imputation strategies to reduce the size of the biases.

Even though our findings raise serious concerns about the quality of meter data, there are also reasons to be optimistic. While device and browser undercoverage are highly prevalent and result in large biases, there is clear room for improvement in the future. Our findings point toward some of the areas where low-hanging fruit can lead to big improvements. Specifically, much of this undercoverage seems to be linked to the current limitations faced by the tracking technologies that we use. With extra investment, most of these limitations could be addressed. Our results shows that it is possible to develop approaches to identify, estimate, and report these errors. Doing so in a similar way as we did should be easy and mostly inexpensive to any other researcher dealing with digital trace data and, especially, web tracking data.

Acknowledgements: I would like to thank Mariano Torcal for his support

Funding: Fundación BBVA; H2020 European Research Council, Grant/Award Number: 849165; Ministerio de Ciencia e Innovación, Grant/Award Number: PID2019-106867RB-I00/AEI/10.13039/501100011033

Data: The data that support the findings of this study is openly available in OSF at https://osf.io/3t7jz/ (DOI: 10.17605/OSF.IO/3T7JZ). More information can be found in the following paper

Bibliography

Amaya, Ashley, Paul P Biemer, and David Kinyon. 2020. "Total Error in a Big Data World: Adapting the TSE Framework to Big Data." *Journal of Survey Statistics and Methodology* 8:89–119.

Araujo, Theo, Anke Wonneberger, Peter Neijens, and Claes de Vreese. 2017. "How Much Time Do You Spend Online? Understanding and Improving the Accuracy of Self-Reported Measures of Internet Use." *Communication Methods and Measures* 11:173–190.

Bosch, Oriol, Marc Asensio, and Caroline Roberts. 2023. *Data donations, are they worth the effort? The accuracy and validity of smartphone usage measures computed with self-reports and data donations*.

Bosch, Oriol J. and Melanie Revilla. 2022a. "The challenges of using digital trace data to measure online behaviors: lessons from a study combining surveys and metered data to investigate affective polarization." *SAGE Research Methods Cases* .

Bosch, Oriol J. and Melanie Revilla. 2022b. "When survey science met web tracking: Presenting an error framework for metered data." *Journal of the Royal Statistical Society Series A: Statistics in Society* 185.

Buskirk, T. D., T Saunders, and J Michaud. 2015. "Are sliders too slick for surveys? An experiment comparing slider and radio button scales for smartphone, tablet and computer based surveys." *Methods, data, analyses: a journal for quantitative methods and survey methodology* 9:229–260.

Cardenal, Ana S., Carlos Aguilar-Paredes, Carol Galais, and Mario Pérez-Montoro. 2019. "Digital Technologies and Selective Exposure: How Choice and Filter Bubbles Shape News Media Exposure." *The International Journal of Press/Politics* 24:465–486.

Chang, Lin Chiat and Jon A. Krosnick. 2003. "Measuring the frequency of regular behaviors: Comparing the "Typical week" to the "past week"." *Sociological Methodology* 33:55–80.

Ernala, Sindhu Kiranmai, Moira Burke, Alex Leavitt, and Nicole B. Ellison. 2020. "How well do people report time spent on Facebook?" *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* .

Festic, Noemi, Moritz Büchi, and Michael Latzer. 2021. "How long and what for? tracking a nationally representative sample to quantify internet use." *Journal of Quantitative Description: Digital Media* 1.

Gasparini, Alessandro. 2018. "RSIMSUM: Summarise results from Monte Carlo simulation studies." *Journal of Open Source Software* 3:739.

González-Bailón, Sandra and Michael A. Xenos. 2022. "The blind spots of measuring online news exposure: A comparison of self-reported and observational data in nine countries." *Information, Communication amp;amp; Society* 26:2088–2106.

Guess, Andrew M. 2021. "(Almost) Everything in Moderation: New Evidence on Americans' Online Media Diets." *American Journal of Political Science* .

Guess, Andrew M., Brendan Nyhan, and Jason Reifler. 2020. "Exposure to untrustworthy websites in the 2016 US election." *Nature Human Behaviour* 4:472–480.

He, Yuanmo and Milena Tsvetkova. 2023. "A method for estimating individual socioeconomic status of Twitter users." *Sociological Methods amp; Research* p. 004912412311686.

Howison, James, Andrea Wiggins, and Kevin Crowston. 2011. "Validity issues in the use of social network analysis with Digital Trace Data." *Journal of the Association for Information Systems* 12:767–797.

Jürgens, Pascal and Birgit Stark. 2022. "Mapping exposure diversity: The divergent effects of algorithmic curation on news consumption." *Journal of Communication* 72:322–344.

Kashyap, Ridhi, Ingmar Weber, Reham Al Tamime, and Masoomali Fatehkia. 2020. "Monitoring global digital gender inequality using the online populations of Facebook and Google." *Demographic Research* 43:779–816.

Konitzer, Tobias, Jennifer Allen, Stephanie Eckman, Baird Howland, Markus Mobius, David Rothschild, and Duncan J Watts. 2021. "Comparing estimates of news consumption from survey and passively collected behavioral data." *Public Opinion Quarterly* 85:347–370.

Mangold, Frank, Sebastian Stier, Johannes Breuer, and Michael Scharkow. 2021. "The overstated generational gap in online news use? A consolidated infrastructural perspective." *New Media amp;amp; Society* 24:2207–2226.

Padró-Solanet, Albert and Joan Balcells. 2022. "Media diet and polarisation: Evidence from Spain." *South European Society and Politics* 27:75–95.

Palmer, Ruth, Benjamin Toff, and Rasmus Kleis Nielsen. 2020. "“the media covers up a lot of things”: Watchdog ideals meet folk theories of journalism." *Journalism Studies* 21:1973–1989.

Parry, Douglas A., Brittany I. Davidson, Craig J. Sewall, Jacob T. Fisher, Hannah Mieczkowski, and Daniel S. Quintana. 2021. "A systematic review and meta-analysis of discrepancies between logged and self-reported digital media use." *Nature Human Behaviour* 5:1535–1547.

Pew Research Center. 2020. "Measuring News Consumption in a Digital Era." Technical report.

Price, Vincent and John Zaller. 1993. "Who gets the news? alternative measures of news reception and their implications for research." *Public Opinion Quarterly* 57:133.

Prior, Markus. 2009. "Improving Media Effects Research through Better Measurement of News Exposure." *The Journal of Politics* 71:893–908.

Rampazzo, Francesco, Alyce Raybould, Pietro Rampazzo, Ross Barker, and Douglas R. Leasure. 2022. *"update: I'm pregnant!": Inferring global use of fertility tracking apps* .

Reiss, Michael V. 2022. "Dissecting non-use of online news – systematic evidence from combining tracking and automated text classification." *Digital Journalism* 11:363–383.

Revilla, Melanie. 2022. "How to enhance web survey data using metered, geolocation, visual and voice data?" *Survey Research Methods* 16:1–12.

Revilla, Melanie, Mick P. Couper, Ezequiel Paura, and Carlos Ochoa. 2021. "Willingness to Participate in a Metered Online Panel." *Field Methods* 33:202–216.

Scharkow, Michael. 2016. "The Accuracy of Self-Reported Internet Use—A Validation Study Using Client Log Data." *Communication Methods and Measures* 10:13–27.

Scharkow, Michael, Frank Mangold, Sebastian Stier, and Johannes Breuer. 2020. "How social network sites and other online intermediaries increase exposure to news." *Proceedings of the National Academy of Sciences* 117:2761–2763.

Sen, Indira, Fabian Flöck, Katrin Weller, Bernd Weiß, and Claudia Wagner. 2021. "A Total Error Framework for Digital Traces of Human Behavior on Online Platforms." *Public Opinion Quarterly* 85:399–422.

StatCounter. 2017. "Desktop vs mobile vs tablet market share united kingdom."

Stier, Sebastian, Johannes Breuer, Pascal Siegers, and Kjerstin Thorson. 2019. "Integrating Survey Data and Digital Trace Data: Key Issues in Developing an Emerging Field." *Social Science Computer Review* p. 089443931984366.

Stier, Sebastian, Frank Mangold, Michael Scharkow, and Johannes Breuer. 2021. "Post post-broadcast democracy? news exposure in the age of online intermediaries." *American Political Science Review* 116:768–774.

Torcal, Mariano, Emily Carty, Josep Maria Comellas, Oriol J. Bosch, Zoe Thomson, and Danilo Serani. 2023. "The dynamics of political and affective polarisation: Datasets for Spain, Portugal, Italy, Argentina, and Chile (2019-2022)." *Data in Brief* 48:109219.

White, Ian R. 2010. "Simsum: Analyses of simulation studies including Monte Carlo Error." *The Stata Journal: Promoting communications on statistics and Stata* 10:369–385.

Chapter 5

Validity and Reliability of Digital Trace Data in Media Exposure Measures: A Multiverse of Measurements Analysis

Oriol J. Bosch

Abstract

Understanding online media exposure is critical, especially in contemporary politics. Given the doubts about survey self-reports, research on media exposure has turned to web tracking data, sometimes considered the gold standard. However, studies revealed that web tracking data is also biased. To improve the understanding of the quality of web tracking measures of media exposure, this paper estimates their predictive validity and true-score reliability. It additionally identifies design choices that optimize their validity and reliability. Using data from a three-wave survey in Spain, Portugal, and Italy, combined with web tracking, this paper conducts a multiverse analysis to assess the validity and reliability of +2,500 web tracking measures of media exposure. Results show an overall high, but imperfect, reliability (0.86). However, in terms of predictive validity, the association between media exposure measures and political knowledge appears weak. This raises questions not only about the predictive validity of web tracking measures but also about the overemphasis on similar critiques regarding survey-based measures. Additionally, results suggest that the design decisions made by researchers can have a substantial impact on the quality of the web tracking data. Methodologically, the paper presents the multiverse of measurements approach, allowing researchers to embrace uncertainty, and improve the transparency of web tracking research.

Keywords:
Digital trace data · Web tracking data · Media exposure · Reliability · Validity · Multiverse analysis

1. Introduction

The measurement of online media exposure, which refers to the extent to which individuals encounter specific media messages or content online (Slater, 2004), is of paramount importance for studying the uses and effects of online media. When researching pressing political phenomena, it becomes essential to have reliable and valid measures of people's online media exposure. However, obtaining suitable measures of media exposure has proven challenging for years.

Traditionally, survey self-reports have served as the primary instrument for measuring media exposure (González-Bailón and Xenos, 2022). Despite testing various approaches (see Golder and Macy 2014, for a summary), doubts persist regarding the validity and reliability of self-reports in measuring online behaviours in general (Parry et al., 2021) and media exposure in particular (Guess, 2015). Research indicates that participants tend to overstate their self-reported media exposure due to complexities in recall, contributing to poor reliability and validity of such measures. Evidence supporting the low quality of self-reports can be observed in the lack of agreement between self-reported and "objective" measures of media exposure (Prior, 2009; Parry et al., 2021).

In response to these challenges, researchers have made significant efforts to develop alternative approaches for measuring media exposure that do not rely on participants' memory. One increasingly popular method involves collecting *digital trace data*, which records users' interactions with specific digital systems (Howison et al., 2011), including telecommunication networks, websites, social media platforms, mobile apps, and digital devices (Stier et al., 2019). One of the most used approaches for collecting individual-level digital trace data is the use of web trackers, or *meters*, a group of technologies which can be installed on participants' browsing devices with their consent. These meters enable researchers to track various traces left by participants while they interact with their devices, such as visited URLs, apps, timestamps, and HTML content. Using the data collected through meters, researchers have investigated pressing questions related to media exposure. For instance, studies have explored the relationship between media consumption and affective polarization (Torcal et al., 2023), the influence of social networks on digital media exposure (Scharkow et al., 2020), and engagement with untrustworthy media (Guess et al., 2020).

However, contrary to the common assumption that web tracking data is the gold standard to measure media exposure (Araujo et al., 2017; Scharkow, 2016; Guess, 2015), recent research has highlighted errors that can affect this kind of data (Revilla et al., 2017; Bosch and Revilla, 2022a,b; Pew Research Center, 2020). In particular, Chapter 4 demonstrate that web tracking estimates of media exposure are often substantially biased due to panel companies' inability to track participants across all devices they use to go online. Such device undercoverage errors negatively affect the validity of web tracking measures and introduce bias into the research findings and conclusions, similar to what happens with surveys.

Given these considerations it is key to assess the quality of media exposure measures derived from web tracking data to understand the potential biases affecting the conclusions drawn from this new source of data. Consequently, this paper has two primary goals, to: 1) estimate the validity and reliability of media exposure measures derived from digital traces collected through meters, and 2) identify design decisions that maximize the validity and reliability of these measures. To explore this, this study uses data from a three-wave survey in Spain, Portugal, and Italy, combined at the individual level with web tracking data. The empirical part of the paper builds upon the multiverse analysis framework initially developed by Steegen et al. (2016), to create a ”multiverse of measurements.” Hence, instead of focusing on one or a few arbitrarily created measures of media exposure, I test the validity and reliability of more than 2,500 different measurements. Specifically, in this Chapter I explore the following research questions:

- What is the overall true-score reliability of news media exposure measures created using digital traces?(**RQ. 1**)
- What is the overall predictive validity of news media exposure measures created using digital traces? (**RQ. 2**)
- Does the predictive validity and reliability of media exposure measures fluctuate across different measurements? (**RQ. 3**)
- What design choices maximise the predictive validity and reliability of web tracking measures? (**RQ. 4**)

The remainder of this article is divided as follows: in the second section, I review the literature on the validity and reliability of media exposure measures. In the third section, the data used is presented. In the fourth section I conceptualise the "multiverse of measurements" and detail how it was built. Next, in the fifth section, the paper's analytical approach is detailed. In the sixth section I present the results and, in the seventh, I conclude.

2. Background

2.1. Conceptualising reliability and validity

When drawing inferences about individuals' media exposure, surveys and digital trace data are employed in a comparable manner. In both contexts, their objective is to quantify an underlying latent concept of interest: media exposure. This construct can be conceptualized as the degree to which an individual comes across media content or messages (Slater, 2004). In quantifiable terms, this could be the amount of time that an individual is typically exposed to media.

To measure this concept, measurements are devised. These can be understood as methods of collecting information about the latent concepts. In surveys, measurements typically take the form of survey questions. When these measurements are employed, for instance, by recording the answer of a person to our question, or tracking and combining a list of traces, we observe a specific quantity of interest (y). Nonetheless, a same respondent could answer a question in different ways if they were asked the same question on different occasions, depending on their mood, attention, or context. Hence, an observed score is only a single draw of an infinite number of different potential observations that could be produced using the specific measurement designed. We can imagine this as a probability density function of possible values of Y (Alwin, 2007).

Given these random fluctuations, hence, a single observation might not be an accurate representation of the true score of a measurement (τ). This true score can be though as the expected value of that hypothetical distribution of observations of Y, for a fixed person p (see Lord et al. 1968). These diversions from the true score are what in psychometrics is considered random measurement errors (ε):

$$\epsilon_p = y_p - \tau_p \tag{5.1}$$

Hence, the true score represents the value that the measurement should produce if no random errors occurred when producing the data. Knowing this, for a given population we can consider that the observed score (Y) is a product of the true score of the measurement (T), plus random error (E):

$$Y = rT + E \tag{5.2}$$

The association between the true score and the observed score (r), when standardised, is what we consider the population parameter of *reliability*. Or in other terms, reliability refers to the relative proportion of random error versus true variance in the measurement of Y, allowing to understand to what extent what we observe is due to noise or signal. Although random errors might not inherently bias univariate estimates, an inflated variance could potentially distort statistical estimates based on them, such as mean difference tests (Cleary et al., 1970), and potentially underestimate standardized relationships derived from correlations or regressions (Alwin, 2007). This model, nonetheless, only accounts for random measurement errors. The true score can still be affected by systematic errors, but this is not the focus of this paper (a better explanation of this is provided in Alwin 2007: 41-42).

Beyond reliability, an important issue of concern is whether the measurement is valid. In the true-score framework, this means: does the true score specified correspond to the conceptual variable that we intend to measure? Is it a valid representation of the theoretical quantity of interest? Across the literature, the concept of validity has a large number of different meanings. Essentially, nonetheless, the validity of a measurement can only be assessed with respect to some criterion (Alwin, 2007). What makes a measure valid? If the concept of interest can be objectively assessed in any way possible, validity will refer to the extent to which the observed values of the measurement and the objective score are the same. This is what is known as *construct validity*.

Conversely, in many cases it is not possible to obtain an objective measure of the construct of interest, because it is inherently unobservable, or there is no unbiased method available. Within the true score tradition, then, the main interest lays in *criterion validity* (Alwin, 2007). Criterion validity is defined simply as the correlation of the observed score with some other variable (X). This variable is assumed to be a criterion linked to the objective of the measurement. In this case, hence, the validity coefficient of a measurement Y with respect to a second measurement X is defined as the absolute value of their correlation coefficient:

$$\mathrm{COR}(X,Y) = \frac{\mathrm{COV}(X,Y)}{\sqrt{\mathrm{VAR}(X)\cdot\mathrm{VAR}(Y)}} \tag{5.3}$$

In this context, hence, the closer the association between the observed score and the criterion score to the true association between the construct and the criterion variable, the higher the validity of the measurement. Nonetheless, given that normally the true association is in itself unknown, it is assumed that higher associations between the observed and criterion score mean higher validity. Although imperfect, when backed by a strong theory, criterion validity can be used as an indicator of whether measures capture the concept of interest. If a relationship should be found and no correlation exists in actuality, it could mean that the designed measure is affected by specification errors and, hence, measures the incorrect parameter (Biemer, 2010).

Section 5 describes how these definitions of reliability and validity are analytically operationalised, and the data used to do so.

2.2. The reliability of media exposure measures

Traditional survey-based self-reports of media exposure are susceptible to a multitude of errors that may compromise the reliability of the reported responses. Conventional exposure measurements have been criticized for imposing a considerable cognitive burden on survey participants (Price and Zaller, 1993). Generating a reliable estimate of the hours (or even days) spent watching political (or news) content within a typical week necessitates an extended multistage recall process (Schwarz, 2001). Given the inclination for rapid responses, respondents often resort to mental shortcuts to provide offhand estimates. All these factors can lead participants to provide answers that diverge from their true score, potentially affecting the reliability of the measures (Prior, 2009).

Despite these concerns, a limited number of studies examining the true-score reliability of self-reported media exposure measures have reported acceptable to good true-score reliability estimates. Specifically, Bartels (1993) reported a true-score reliability of .75 for a traditional television exposure measure, while Dilliplane et al. (2012) found that a list-based measure assessing the number of political programs viewed on TV yielded a true-score reliability of .83. These findings align with the average reliability of factual survey questions (approximately .75) reported by Alwin (2007). Nevertheless, to the best of the author's knowledge, the true-score reliability of online media exposure remains unexplored.

While the process of gathering and processing digital traces through web trackers to formulate a specific variable distinctly differs from traditional survey methods, it remains susceptible to measurement errors. Notably, Bosch and Revilla (2022b) outlined eight theoretical sources of measurement errors inherent in digital trace data collected via web trackers. These sources include errors related to the tracking technologies, participants' incomplete installation of tracking technologies across their devices, and device-sharing scenarios with non-participants.

Some of these error sources might exhibit high variability, fluctuating in their direction and magnitude. For example, if patterns in device sharing with non-observed individuals are not constant, they could introduce random errors to measures. Similarly, technology errors can significantly fluctuate when participants use multiple devices, each with different tracking technologies, and varying usage patterns. Nonetheless, based on the limited available past research, it is expected that most web tracking data sources of error introduce systematic rather than random errors (see Chapter 4).

Empirical research to date has not undertaken the task of estimating the reliability of media exposure variables computed from digital trace data. Although we might expect measures created with web tracking data to be reliable, the lack of evidence leaves researchers without a comprehensive understanding of the extent to which prior and ongoing research might potentially yield underestimated standardized relationships and draw misleading conclusions regarding the significance of estimates.

2.3. The validity of media exposure measures

In the realm of surveys, specification errors often manifest when questions and scales fail to accurately encapsulate the intended concept. In today's intricate media landscape, the conventional approach of inquiring about participants' exposure to specific media, such as political news, is fraught with challenges. These challenges encompass determining the precise definition of ”news” and what qualifies as ”political,” as well as ensuring a uniform understanding among all participants. Given these complexities and others, concerns about the validity of self-reported measures of media exposure are pervasive.

The primary criticism levied against such measures centres on their predictive validity, a subtype of criterion validity. Specifically, self-reports gauging media exposure aim to assess (news) media exposure, yet the observed scores exhibit a notably weak correlation with their hypothesized outcomes: political knowledge (Zaller, 2002) and news recall (Price and Zaller, 1993; Chang and Krosnick, 2003). This lack of correlation has prompted some to advocate for the abandonment of self-reported measures of media exposure (Price and Zaller, 1993).

However, weak validity is not exclusive to self-reports (Jungherr, 2019). In the realm of digital trace data, measurements are based on specific pieces of information derived from participants' tracked online behaviour. These pieces are subsequently combined and, at times, transformed to compute a particular variable. In the context of media exposure, this process may involve categorizing all URLs related to political articles and aggregating the time users spend on those URLs. If the traces utilized for constructing the variables do not align accurately with the concept of interest, web tracking measures may also be rendered invalid, even in the absence of random or systematic errors during the collection and processing of the traces (Bosch and Revilla, 2022b).

While often lacking empirical validation, two potential mechanisms could give rise to specification errors in web tracking research. Firstly, discrepancies may emerge when mismatches occur between the traces defined for constructing the measurement and the intended concept. For instance, when aiming to measure exposure to political media, URLs must be categorized as either political media or not. This categorization can be achieved through manual or automated methods (e.g., Peterson et al. 2018; Bach et al. 2022). Any disparities between the actual nature of the URL and the manual or automated classifications will introduce specification errors. Secondly, media exposure measures can also suffer from weak validity if they incorporate by-design missing data (Bosch and Revilla, 2022b). To illustrate, Reiss (2022) demonstrates that when measuring the proportion of individuals avoiding news online, omitting information regarding news exposure through mobile apps yields problematic outcomes. Specifically, the author shows that disregarding app-based exposure leads to an overestimation of the proportion of individuals identified as news avoiders by 9 percentage points. If researchers want to measure whether an individual is generally a news avoider online, using this measure would introduce specification errors.

Despite the presence of some evidence, research examining the validity of media exposure measurements derived from digital trace data remains limited.

2.4. The effect of researchers' design decisions

The reliability and validity of a variable depend not only on the data source, such as surveys or digital trace data but also on the design decisions made during the formulation of the measurement instrument. For surveys, over 60 design characteristics can significantly influence their validity and reliability (Saris and Gallhofer, 2007, 2014), especially concerning how survey questions and their scales are designed and administered (DeCastellarnau, 2017; Bosch et al., 2018; Bosch and Revilla, 2021; Michaud et al., 2023). Therefore, various approaches can be employed to measure media exposure through survey self-reports, and the validity and reliability of these measurements can exhibit notable variation based on the design choices made (see Goldman and Warren 2019, for a comprehensive review).

Within the context of measuring media exposure using digital trace data, several key design questions arise, such as: which URLs and apps can be considered as news media articles? How do we define a visit to a URL that qualifies as exposure to media content? For how long do we need to track someone to capture their typical media exposure? These and other design questions are often approached in non-uniform ways across the literature. The definition of "news," the methods used to categorize URLs as "news," and the duration of the tracking period all exhibit substantial variability, for example. Interestingly, most of these design decisions are made once the raw dataset of traces has already been collected, in contrast to surveys, which require designing measurements before data collection. This presents both advantages and challenges for web tracking data. While it means that researchers can compute multiple measures of the same concept at no extra cost, it also implies that for every concept, there is nearly an endless number of different design choices that researchers could test.

Although empirical evidence remains scarce, previous investigations have indicated that specific design choices yield disparate outcomes. For instance, Mangold et al. (2021) demonstrated that altering the time threshold used to define a visit to a news outlet—ranging from a 3-second threshold to a 120-second threshold—resulted in divergent estimates regarding the proportion of millennials categorized as news avoiders, as well as the breadth of their news consumption. Similarly, Reiss (2022) explored whether the proportion of individuals identified as news avoiders differed based on the approach used to classify URLs as ”news” or otherwise. Findings indicated that employing a supervised text classification approach to categorize the textual content of articles accessed by participants, as opposed to designating any URL within a news media outlet webpage as containing news-related content, led to a higher prevalence of online news non-users.

Yet, the extent to which researchers should be concerned about each potential design choice when using digital trace data remains an open question. If digital trace data is inherently valid and reliable, one might assume that any measurement instrument created using this data source would yield acceptable levels of validity and reliability. Nonetheless, if fluctuations exist, comprehending the individual effects of each design choice on the resulting validity and reliability becomes crucial in crafting optimal measurement instruments and predicting the quality of digital trace data variables both a priori and a posteriori (Saris et al., 2011).

3. The TRI-POL dataset

I use data from the TRI-POL project (Torcal et al., 2023), the goal of which is to understand whether and how online behaviours are related to affective polarisation across Southern European and Latin American countries (https://www.upf.edu/web/tri-pol)[1]. TRI-POL conducted a three-wave survey between September 2021 and March 2022. Survey responses were matched at the individual level with metered data. Data were collected through the Netquest opt-in metered panels (https://www.netquest.com), which consist of individuals who have meter(s) already installed in their devices and who can also be contacted to conduct surveys. Panellists receive more incentives if they install the meter on more devices (up to a maximum of three). Respondents' online behaviours, hence, can be linked with their survey answers. In this study I focus on the data collected in Italy, Portugal, and Spain.

Cross quotas for age and gender, and quotas for educational level, and region were used in each country to ensure a sample similar on these variables to the general online population of those countries. Survey questions were used to measure attitudinal and demographic variables, while metered data were used to measure

[1] More information about the data collection strategy of both survey and digital trace data can be found in the TRI-POL data protocols: https://osf.io/3t7jz/

variables related to the general Internet use as well as consumption of specific news media outlets, political news, and social media (see the TRI-POL data protocols in footnote 2 to check the specific URLs defined to measure these concepts). Metered data was collected for the 15 days prior to and following participants starting the questionnaire. The meter logged each URL accessed by the panellists, along with timestamps indicating the initial visit to the URL, and the duration in seconds during which the URL remained the active content within the browser, or in the case of mobile devices, on the smartphone screen. It is important to note that a URL or app was classified as 'active' when it was the foremost content displayed in the browser or on the device's screen. This definition excludes any other URLs or apps that might have been open in separate tabs or screens, as they were not considered active during this time frame. The duration of active engagement was computed as the elapsed time between the moment the URL or app first gained 'active' status within the browser or device and the point at which a different URL or app took over as the active content in the browser or device. A visit was defined as any opened URL/app lasting one second or more. Participants were tracked on iOS and Android mobile devices, and Windows and MAC computers, using the tracking solutions provided by Wakoopa (https://www.wakoopa.com/). Windows and MAC devices were tracked with desktop apps and/or web browser plug-ins, Android devices through apps and iOS devices through manually configured proxies. Torcal et al. (2023) provides more information about the collectable data and the characteristics of each of the tracking technologies used. Overall, the sample presents a tracking undercoverage rate of 74%, meaning that three quarters of the participants were not tracked in all the devices they use to go online.

Challenges were faced when filling some of the specific cross-quotas with participants from the metered panel. Hence, in some cases panellists without a meter installed had to be invited to fill some of the quotas. Thus, in total, for the first wave, 3,548 respondents completed the survey, but only 2,653 had the meter installed in at least one mobile (smartphone or tablet) or PC device: 993 in Spain, 818 in Portugal and 842 in Italy. No significant differences are observed between the full sample and the subsample of tracked participants, across a selection of demographic, political and technological variables (see Supplementary Material 1, i.e., SM 1).

4. A multiverse of measurements

In this study, I build upon the multiverse analysis framework initially developed by Steegen et al. (2016), introducing the idea of a ”multiverse of measurements.” Instead of being limited to a single measurement, a concept can be captured through various alternative measurements. Each measurement within this extensive multiverse represents a distinct combination of design choices. As a result, a multiverse of measurements also leads to a variety of quality assessments, as each measurement produces its own set of validity and reliability parameters.

Traditionally, researchers often focus on a single set of choices and present it as the only approach taken in their analyses. This practice of selectively emphasizing one set of choices assumes that the chosen measurement either perfectly represent the entire universe of measurements, or it can be considered as the best multiversal alternative. However, the decisions made during the design phase are often arbitrary and lack clear justification. When researchers choose one measurement from the many possibilities, they overlook the range of quality estimates that could arise. The inherent uncertainty in the data and the sensitivity of the results are not fully considered, making it difficult to interpret a single result accurately. This is justified with surveys, given the complexity of asking participants repeated questions about the same concept. For web tracking data, nonetheless, it is not. Once the raw data from the web trackers is collected, it is mostly inexpensive and effortless to go from one measurement to thousands of them.

To address the challenges posed by selective reporting, I adopt a multiverse analysis which enhances transparency by revealing the sensitivity of the results to different design choices. Additionally, it allows identifying the key choices that significantly improve or harm the reliability and validity of media exposure measurements, helping future researchers make better informed decisions.

In the present study, I explore the reliability and validity of the multiverse of measurements aimed at measuring one of the most explored constructs in the media exposure literature: the extent to which individuals are exposed to written news media in the online domain. Several important considerations merit discussion: Firstly, this paper's measurement scope excludes exposure to visual or aural media, such as videos or podcasts. Secondly, this examination is confined to the domain of news, broadly defined as accounts of recent, interesting, and significant events (Kershner, 2012) published in news media outlets. This conceptualization is intentionally broad, to account for the fluctuation in the definition of “news” as a design feature (Reinemann et al., 2011). Thirdly, the used definition excludes exposure to ”news” disseminated via social media, blogs, and other non-media platforms. Lastly, the focus is directed towards online exposure, reflecting the fact that web trackers can exclusively capture activities transpiring within the digital realm. All in all, as most past research studying media exposure with web trackers, I focus on written articles and news produced as an output of journalism and published by a news media outlet.

To perform the multiverse analysis for this specific construct, I first constructed its multiverse of measurements, encompassing all potential measurements resulting from the combination of various reasonable design choices. Subsequently, the reliability and validity of each measurement within this multiverse were independently computed, resulting in a multiverse of quality estimates. Reliability estimates were produced using the Quasi-Markov Simplex Model (QSM, Alwin 2007). Validity was estimated as the association between media exposure and political knowledge. Section 5 explains the analytical approach in more detail.

In this multiverse analysis, we considered design choices that other researchers have previously followed and those outlined by Bosch and Revilla (2022a,b) in their error framework

4.1. Constructing the multiverse of measurements

Table 1 presents a summary of the six identified design choices that were followed when creating measures for the concept of interest, along with the various reasonable options that were considered for each choice. Below, I provide detailed descriptions of the different options.

Metric: I measured media exposure using multiple metrics: the number of visits to news media outlets, time spent exposed to news, days of exposure, and the number of different media outlets exposed to.

List of traces: I adopted a list-based approach by defining a list of websites that publish ”written news” and then considering either all or part of their URLs as ”written news.” To create this list, I followed the stepwise approach proposed by Bosch and Revilla (2022a). Firstly, I identified websites on the Internet that publish ”written news” in Spain, Portugal, and Italy using four distinct top site rankings: Tranco, Alexa, Cisco, and Majestic. Secondly, from the aggregated lists of these ranking sites, I selected which websites to include in the list, based on their popularity. Hence, I chose different media outlets based on their ranking: the top 10, 20, 50, 100, and 200 most popular, as well as all identified news media outlets (up to 761 in Spain). Thirdly, I defined which URLs within each listed news media outlet's website I would consider as news. ”News” can encompass various meanings, often divided into categories such as ”hard” and ”soft” news (Reinemann et al., 2011). Past research has embraced different definitions, ranging from considering any URL posted by a news media outlet as news (including sports news or movie and theatre reviews) to manually or automatically identifying subsets of ”hard” or ”political” news only (e.g., Reiss 2022). I considered URLs as news in two distinct ways: 1) any URL published by a news media outlet, and 2) only URLs dealing with ”hard” news[2]. More in-depth insight into the approach used for identifying these URLs can be found in the official documentation of the TRI-POL dataset[3].

Table 5.1: Design choices and options for measuring media exposure

Choices	**Options**
Metric	Visits, Seconds, Days, No Media
List of traces	
List of media	Tranco, Alexa, Cisco, Majestic
Top media	10, 20, 50, 100, 200, All
Information	All URLs, only those identified as ”hard” news
Exposure	1-second, 30-seconds, 120-seconds threshold
App behaviour	Included, excluded
Tracking period	2, 5, 10, 15 days

Criteria to define exposure: Even if an individual visits one of the defined URLs and/or apps, the generated traces may not be relevant to the concept of interest. Exposure can involve merely encountering the content, irrespective of the duration, or also interacting with the content, such as reading it. To capture different levels of exposure, I defined three distinct categories based on the time spent on a specific URL or app. I assume that longer visits indicate a higher likelihood that the person read all or part of the content. For this study, I replicate Mangold et al.

[2] I considered as “hard” news any articles covering political, national, international, and regional affairs, as well as political opinion pieces.

[3] https://osf.io/38kt6?view$_{o}nly = 22e669dfd9a946d5b706e0efcd584d7c$

(2021) approach, considering that for a visit to lead to exposure a person should spend 1, 30, or 120 seconds or more on the defined URLs.

App behaviour: Some individuals get exposed to news through news media apps. For example, people can read news on the app of The Guardian, as well as on its webpage. Researchers, with their design decisions, can exclude or include the traces that researchers leave when using apps to consume news. Although in some cases this will be a by-product of the limitations of the tracking technologies that researchers used, it is a researcher's choice to use this information with by-design missing data to make inferences for people's general media exposure. In this case, I computed the measures with or without using app information.

Tracking period: Researchers generally aim to measure participants' ”typical” or ”normal” behaviours by calculating the average behaviour during the tracked period. The tracking period, however, can impact the prevalence of outliers and the skewness of the data, ultimately influencing the estimates. As such, I computed the average time individuals engage with written news media in the online domain using 2, 5, 10, and 15 days of tracking information.

Based on this tabulation of choices, the multiverse of measurements was constructed by considering all possible combinations of design choices, resulting in a unique measurement for each combination of choices. In total, there were 4 × 4 × 6 × 2 × 3 × 2 × 4= 4,608 choice combinations, although some of the combinations were not possible. For instance, when counting the minutes of exposure, I could not apply a “visit” threshold. Additionally, “hard” news were only identified for the top 10, 20 and 50 for each country since they were manually coded. After excluding impossible combinations, the study was left with 2,631 choice combinations.

5. Analyses

5.1. Reliability

To compute the true-score reliability of each variable, I used the Quasi-Markov Simplex Model (QSM) (Alwin, 2007), using the Wiley and Wiley (1970) approach. This can be summarised by the following system of equations:

$$Y_t = rT_t + E_t \tag{5.4}$$

$$T_t = \beta_{t,t-1} T_{t-1} + Z_t \tag{5.5}$$

where the observed score Y_t for a given wave t is the true score T_t for that wave plus a random error E. The true score at a given wave (T_t), in turn, reflects the true score at the previous wave (T_{t-1}) plus change over time (Z_t). Here, $\beta_{t,t-1}$ reflects the relationship between the true score at wave t and the true score at the prior wave. The association between the true score and the observed score (r), when standardised, is what we consider the parameter of reliability. Figure 1 shows this model.

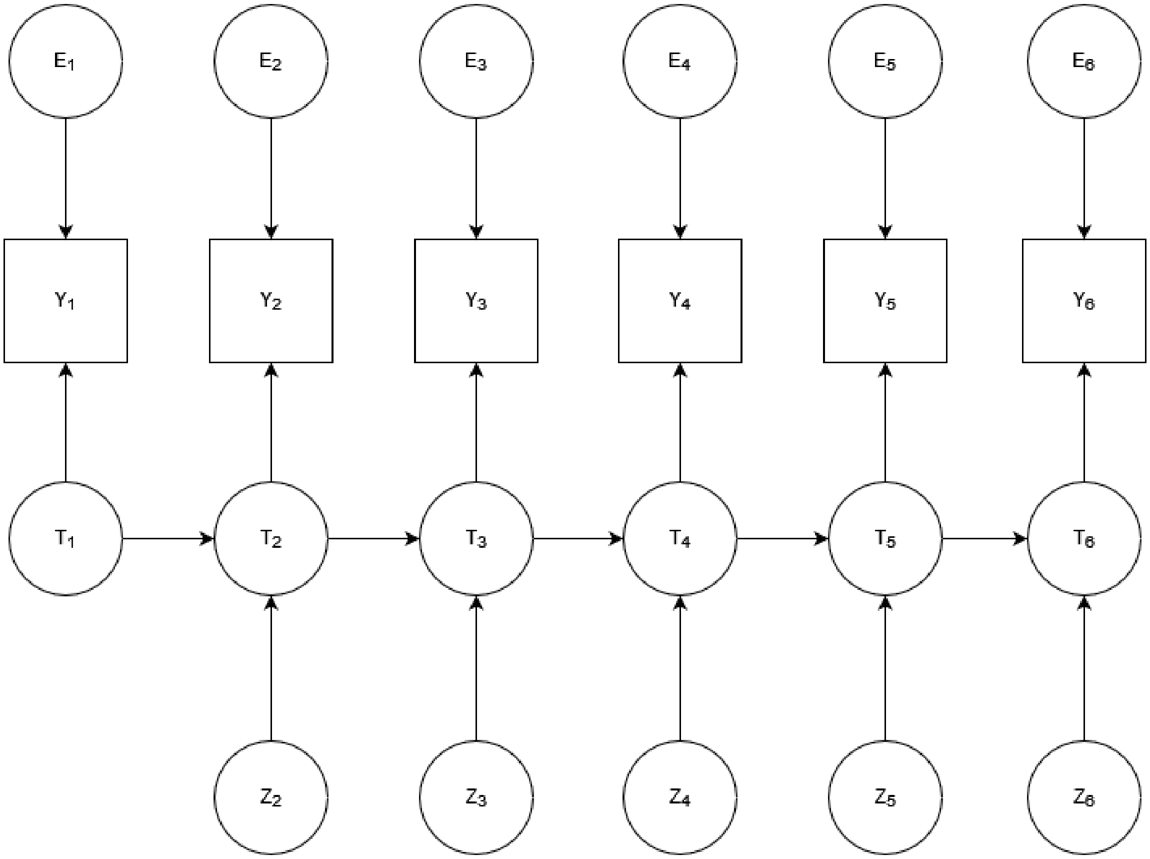

Figure 5.1: Quasi-simplex model with six waves

Wiley and Wiley's (1970) quasi-simplex model is only empirically testable when several assumptions and restrictions about the relations between the estimated parameters E_t, T_t and Z_t are made (see Cernat et al. 2021 for a more in-depth discussion of these). First, observations are assumed to be independent over time, hence, E_t and Z_t are uncorrelated across waves. Second, the mean of the observed score and the true score are 0. This implies that all variables are centred. Third, a Markovian process is assumed, in which the distribution of the true variables at time t are only dependant on the distribution at time $t-1$. Fourth, measurement errors are assumed to be equal over time. Hence, the error variances $V(E_t)$ must be equal at every time t. Fifth, the variance of the error term $V(E_t)$, and the stability term $V(Z_t)$ are constrained to follow a normal distribution with mean of 0. Finally, the covariances between the true scores, errors, and stability are zero.

With three waves of data, a QSM that implements the assumptions above will be just identified. Any more waves of data will lead to a degree of freedom greater than 0, and thus enable a test of the model fit. Additionally, if more waves are available, it is possible to test and relax some of these assumptions, which in some cases can improve the fit of the model and reduce the likelihood of models presenting convergence issues and implausible parameter estimates (Cernat et al., 2021).

In order to relax the assumptions of the model, and test whether by doing so models performed better, I subdivided the web tracking dataset into a six-wave panel dataset. Specifically, the raw dataset had data for three time periods of 30 days, amounting to a total of 93 days of data covering a span of six months. In order to compute the average measures of media exposure of participants, I subdivided this dataset into six distinct time periods within the 90-day span, to calculate the average news exposure for each period. The length of these time periods varied depending on how long the tracking period was determined to be for each of the 2,631 measurements created: 2, 5, 10 or 15 days (see Table 1). Figure 2 exemplifies this for measures computed using 15 days of tracking.

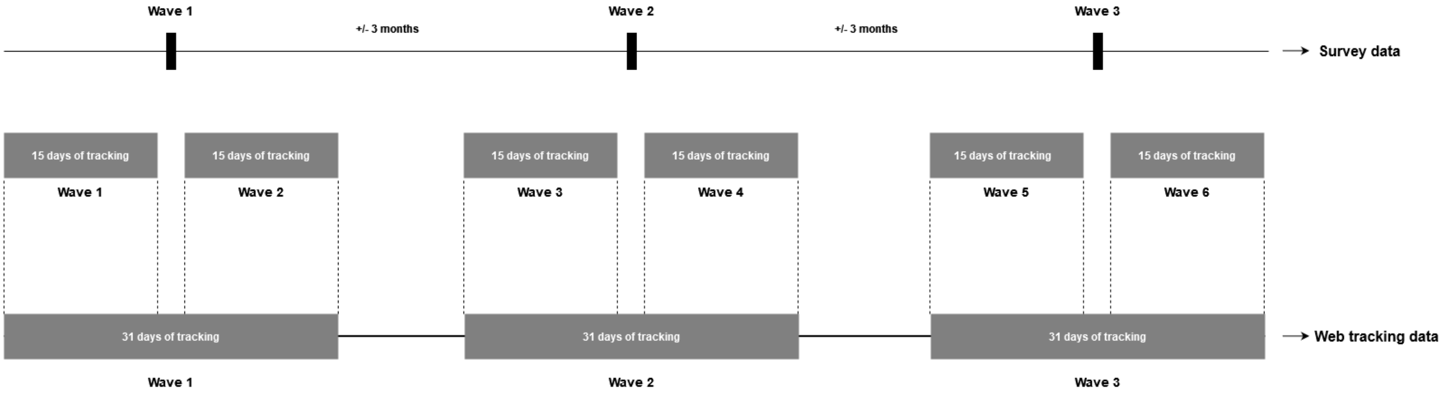

Figure 5.2: Exemplification of how web tracking data is operationalised into waves

Using these datasets, I computed the reliability for each variable using Confirmatory Factor Analysis (lavaan R package, Oberski 2014). To account for the truncated and zero inflated nature of media exposure variables, these were log transformed[4]. Following Cernat et al. (2021)'s recommendation, I computed a base model using Wiley and Wiley (1970)'s assumptions. Additionally, I tested four more models, each of them independently relaxing one of the previously presented assumptions.

After comparing the proportion of models leading to improper solutions, and the overall fit of the models, I determined that two models performed the best. On the one hand, a model relaxing the Markovian process assumption, which included four lag-2 effects between the true scores:

$$Y_t = rT_t + E_t \tag{5.6}$$

$$T_t = \beta_{t,t-1}T_{t-1} + \beta_{t,t-2}T_{t-2} + Z_t \tag{5.7}$$

Were $\beta_{t,t-2}T_{t-2}$ represents the coefficient that reflects the relationship between the true score at wave t and the true score at the wave two time points before (T_{t-2}). On the other hand, a model adding means to the baseline model by estimating the intercept of the observed scores (set to be equal across time), also outperformed the others:

$$Y_t = \alpha + rT_t + E_t \tag{5.8}$$

$$T_t = \beta_{t,t-1}T_{t-1} + Z_t \tag{5.9}$$

Where α represents the estimated intercept of the observed scores. I call these models "2-LAG" and "Equal means." SM 2 explains these comparisons in more detail and presents information about the proportion of improper models, and their average reliability. In the main text, I will only present the results for these two models.

Therefore, I obtained 2,631 true-score reliability coefficients (average reliability across the six waves), for each country and model. This information was used to measure the average reliability (RQ. 2), and its fluctuation across measurements (RQ. 3).

[4] To deal with 0s, all observations were added +1 before transformation.

5.2. Validity

To evaluate the validity of the 2,631 media exposure measures created with web tracking data, I focused on their predictive validity, a type of criterion validity. Predictive validity assesses whether the measurement predicts or correlates with an external criterion that is theoretically related to the construct being measured (Smith et al., 2019). To do so, researchers compute the (partial) correlation between the observed score of the measurement, and the observed score of the criterion. While in general the true relationship between the construct and the external criterion is often unknown, stronger associations between the observed scores are often considered to be an indication of better validity (Prior, 2009). Although I refrain from making this assumption, any variations across the multiverse of measurements in the predictive power of their variables would indicate differences in terms of predictive validity.

In the media effects literature, the common external criterion used to assess predictive validity is political knowledge. The underlying assumption is that exposure to news should lead to enhanced political information, resulting in higher political knowledge (e.g., Dilliplane et al. 2012). However, tests of validity typically focus on examining how well a measure correlates with political knowledge at a single point in time. This approach assumes that individuals who consume more news generally possess higher levels of political knowledge. To assess predictive validity, I employ two distinct approaches based on previous literature (Dilliplane et al., 2012; Prior, 2009).

First, for each country, I conducted a fixed-effects regression model examining within-person changes in political knowledge between waves 1 and 3:

$$X_{it} = \beta_1 Y_{ijt} + \alpha_i + u_t + e_i \tag{5.10}$$

Here, X_{it} represents the observed political knowledge for participant i at time t, and Y_{ijt} the observed media exposure score for participant i at time t, measured with measurement j. The political knowledge variable was an additive index ranging from 0 to 4, with 0 indicating no correct answers to political knowledge questions and 4 indicating all answered correctly. Questions revolved around basic knowledge about how institutions work, and the composition of the current government (specific questions for each country are detailed in SM 3). The media exposure measures were computed using web tracking data collected immediately before participants answered each survey wave. Additionally, the model included an individual- (α_i) and a wave-specific (u_t) fixed effect, allowing to account for unobserved time-invariant heterogeneity among individuals and time-specific effects. From this equation, hence, predictive validity is assumed to be represented by the partial regression coefficient of Y (β_1). To facilitate the comparison of the coefficients across the multiverse of measurements, the coefficients used were standardised. Considering the limited number of waves available and the short time span between them (three months), there was limited margin for substantial fluctuations in people's knowledge over time. Hence, I also conducted separate Ordinary Least Squares (OLS) regressions for each country using data from the first wave of TRI-POL:

$$X_i = \beta_1 Y_{ij} + \beta_2 + C_i + e_i \tag{5.11}$$

The dependent variable (X_i) was the index of political knowledge, and the main independent variable was our measures of media exposure (Y_{ij}). Again, the political knowledge variable ranged from 0 to 4. C_i represents the vector of the different common control variables included in the model, accounting for participants' self-reported sex (Male/Female), age (continuous), education (completed tertiary education/did not complete), left-right orientation (partially labelled on a 0 to 10 scale), and political interest (fully labelled on a 1 to 5 scale, five representing the highest interest). Further details on these variables can be found in SM 4. As for the previous model, predictive validity is assumed to be represented by the partial (standardised) regression coefficient of Y (β_1).

Both models were executed for each computed measure, resulting in 2,631 standardized partial regression coefficients for each country, and for each model. This comprehensive analysis allowed us to measure the average predictive validity and assess its variation across different measurements.

5.3. Predicting the effect of each design choice

In addition to examining how the validity of online news media exposure measurements created with metered data varies across different measurement approaches, I also aimed to understand how these choices impact the reliability and validity of the measurements.

Inspired by the Survey Quality Predictor (SQP, Saris et al. 2011), the study took a predictive approach to uncover the connection between each design choice and the reliability and validity of media exposure measurements from web tracking data. To do this, I compiled a new dataset with the multiverse of quality assessments. In this dataset, I used the total of 7,893 variables computed across countries as observations, their validity and reliability coefficients as dependent variables, and the seven design choices of the measurements (see Table 1) plus the country of the variables as the group of predictors.

For predicting the impact of each design choice, I used random forests of regression trees (using the R package randomForest, Liaw and Wiener 2002). Random forests have been shown to work best for similar endeavours (e.g., SQP 3.0, Felderer et al. 2023), due to their ability to handle complex relationships, interactions, and outliers effectively. A random forest is a machine learning algorithm that combines the predictions of multiple decision trees to make more accurate and robust predictions. While OLS regression tries to find a single linear equation that best fits the data, a random forest uses an ensemble of regression trees, making it capable of handling complex, nonlinear relationships. Hence, a random forest builds not just one but many decision trees. Each regression tree is created through a Classification and Regression Tree (CART) algorithm (Breiman et al., 1984). Each tree is trained on a subset of the measurements (some in-bag, some out-of-bag), and a random subset of predictors (design choices). This randomness introduces diversity among the trees. I employed 500 trees for each of the four models: two for the validity coefficients and two for the reliability coefficients. On average, each measurement was ”out-of-bag” about 346 to 374 times, which means it was absent in around 69-75% of the trees in the forest. Additionally, I randomly selected five out of the eight variables without replacement for each analysis.

To gain a comprehensive understanding of how individual design options (e.g., using 15 days of data, instead of 10) impact the reliability and validity of measurements, while accounting for the effect of other design choices, I computed Partial Dependence Plots (PDPs) for each design choice. PDPs allow to compute the predicted reliability and validity as it systematically varies the values of each feature while keeping all other design choices averaged, capturing the average effect across those choices. Building upon the PDP analysis, I extended the investigation to predict the reliability and validity that each design option would achieve. This prediction involved averaging the predictions across the categories of all other features, effectively estimating the expected reliability and validity for each design choice under varying configurations. This step enables providing quantitative insights into the anticipated performance of different design options.

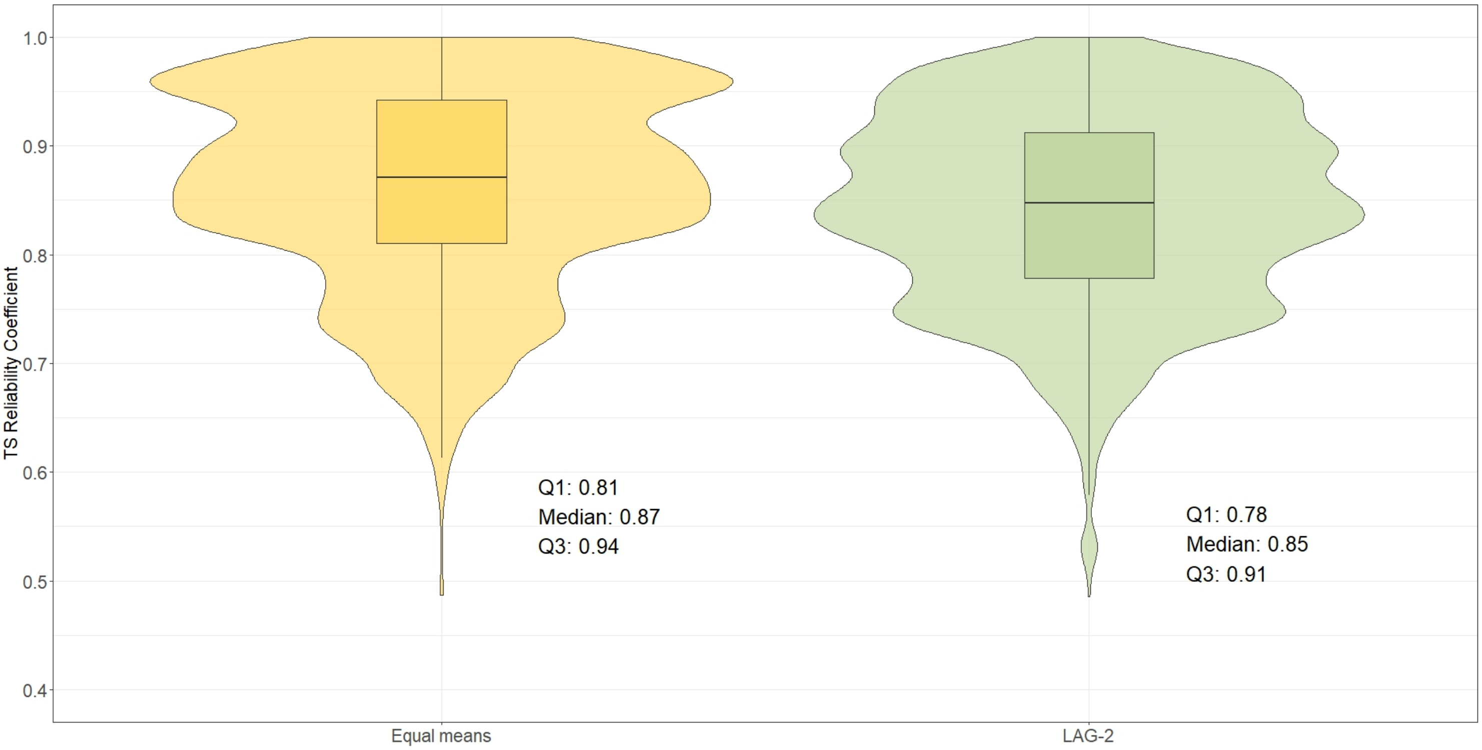

Figure 5.3: Descriptive results of the 7,893 reliability coefficients obtained, for each method

6. Results

6.1. True-score reliability

Figure 3 presents an overview of the findings on the multiverse of reliability estimates. The diagram illustrates the distribution of the 7,893 reliability coefficients for all countries combined, separately for each model. Additionally, the graph shows the first, second, and third quartiles.

Focusing on the 2-LAG model, Figure 3 shows that the median reliability coefficient obtained across countries is of 0.85. Hence, the median explored measure of media exposure captures 85% of the variance of its true score. Although Figure

2 shows that variation exists across measurements, and following DeCastellarnau et al. (2017), 24.9% of the measurements present values deemed as acceptable (0.70-0.80), 39.5% as good (0.80 – 0.90), and 29.6% excellent (> 0.90). Only 0.9% of the measurements yielded measurements deemed poor or unacceptable (< 0.60).

Figure 3 also shows the results for the model with equal means. This model, overall, yields slightly higher reliability coefficients. The median reliability is of 0.87. Again, although variability is observed across measurements, 17.2% of the measurements present acceptable reliability coefficients, 37.6% good, and 39.4% excellent. Out of all measurements tested, only 0.4% presented measurements deemed poor or unacceptable.

Considering these results, the measurements explored generally present high reliability coefficients. Not only so, many of them present excellent reliability estimates, which capture close to all the variance of their true score.

6.2. Predictive validity

Next, I focus on predictive validity. Figure 4 illustrates the distribution of standardized regression coefficients for all countries combined, separately for each model. These coefficients represent the association between media exposure and (gains in) political knowledge.

When examining the average values derived from the fixed effect regression model, Figure 4 shows that the median standardized association between changes in media exposure and political knowledge is of 0.01. Essentially, this suggests that, overall, a one standard deviation increase in media exposure corresponds to a minimal 0.01 standard deviation increase in political knowledge. Although results exhibit fluctuations across measurements, the overall impact of all coefficients remains notably diminutive, denoting a near-null to very weak association. The highest standardized coefficients attained is 0.13. Notably, it is interesting to observe that 38.7% of coefficients across countries manifest negative values, implying that augmented media exposure corresponds to marginal declines in political knowledge across different waves. For comparison purposes, across countries, the average standardised coefficient yielded by a survey self-report was of 0.02 (see SM 5 for the details of this analysis).

Figure 4 also shows the outcomes stemming from the cross-sectional OLS regression model. In this context, the median standardized association is substantially larger, at 0.09. Although the distribution of coefficient estimates diverges among these outcomes, both models yield comparably modest standard deviations (approximately 0.03). The peak standardized coefficients reached if of 0.17. Across countries, 36.2% of coefficients surpass values of 0.10, with 1.7% even exceed 0.15. Nevertheless, despite the relatively larger coefficients presented by the cross-sectional OLS regression model, most of the examined measurements still show notably feeble associations. In comparison, across countries, a survey self-report yielded an average standardised coefficient of 0.10 (also see SM 5).

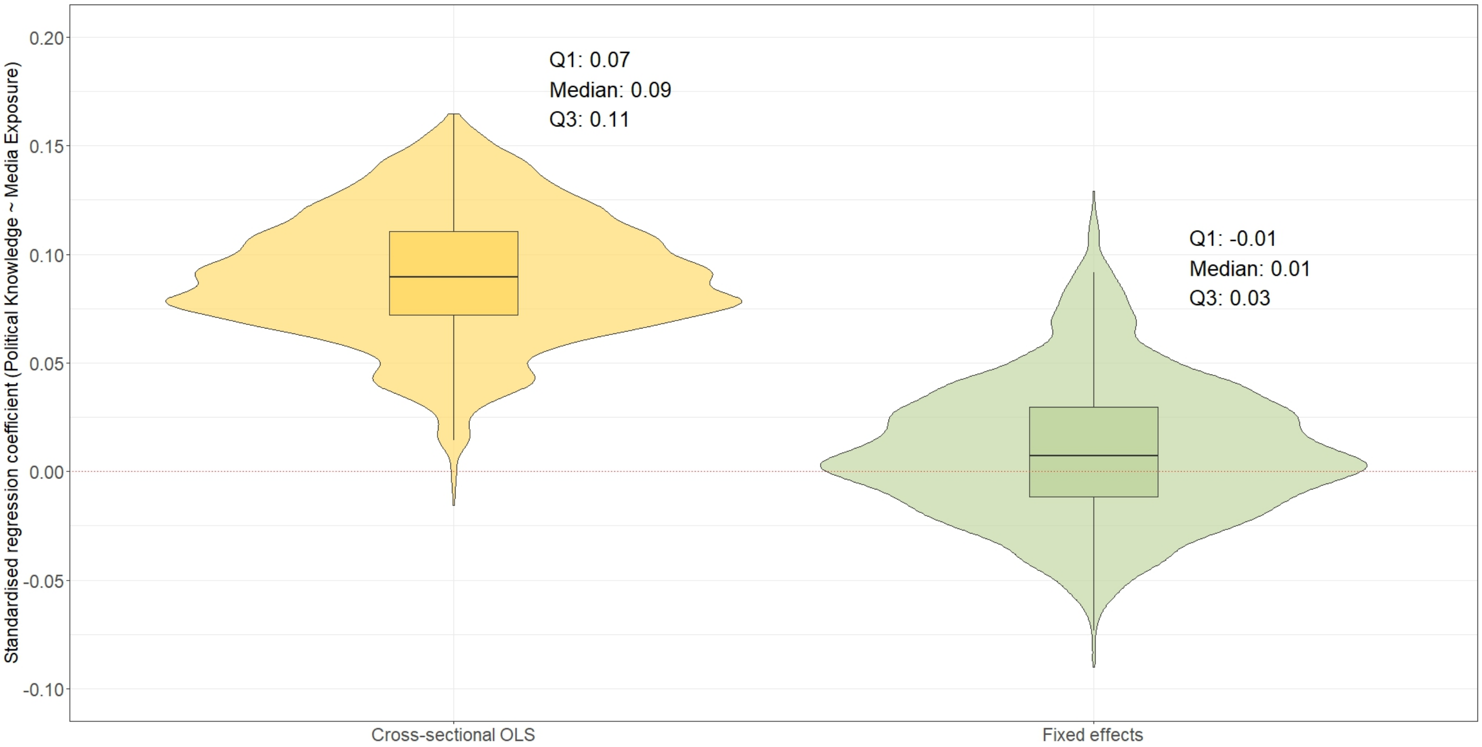

Figure 5.4: Descriptive results of the 7,893 standardised regression coefficients obtained, for each method

On the whole, while the theoretical expectation is that media exposure would be associated to a gain in people's political knowledge, or at least that people that are exposed to media should show higher political knowledge, these results hardly suggest so. Most of the measures show that media exposure, in the time explored, hardly lead to any increase in political knowledge. Additionally, while there is an association between media exposure and political knowledge at the cross-sectional level, it is modest if not low.

If we follow what past research has done, this suggests that the explored measures of media exposure derived from web tracking data exhibit low predictive validity, irrespective of the regression model employed. While this is hard to prove with this data, given that the true association between media exposure and political knowledge is unknown, results do show that validity fluctuates to some extent across measurements, suggesting that how measurements are designed has an impact on validity.

6.3. The effect of each design choice

In the following section, we consider the connection between various design options and the validity and reliability coefficients of the explored measurements related to media exposure. Overall, the models yielded strong results, with R-squared values of 0.97 for both reliability models, and 0.89 and 0.95 for the fixed and cross-sectional validity coefficients, respectively. The squared correlations between predicted and observed coefficients were 0.97-0.98 for the reliability coefficients, and 0.91-0.90 for the validity ones. SM 6 graphically presents the predicted coefficients against the observed ones.

Presented in Figures 5 and 6 are the adjusted predicted reliability and validity coefficients for each design option across the different models I explored. These results encompass data from all countries, with the effects unique to each country being accounted for through three distinct dummy variables.

Figure 5 presents the average predicted reliability coefficients, for both models. Both the 2-LAG and the equal means model show almost identical results. Based on a variable importance test, the design choices with a higher importance in the prediction model (i.e., higher increase in the RMSE when excluded) are the length of the tracking period, the metric used, whether metrics are computed considering all URLs as news or only those dealing with politics, and the country (see SM 7 for the exact values). This can also be appreciated in Figure 5.

The reliability varies greatly depending on the length of the tracking period. While, on average, we should expect a predicted reliability of 0.75/0.78 when computing a measure of media exposure only with two days of data, this value goes up to 0.90/0.91 when using 15 days of data. Hence, results seem to suggest that the longer

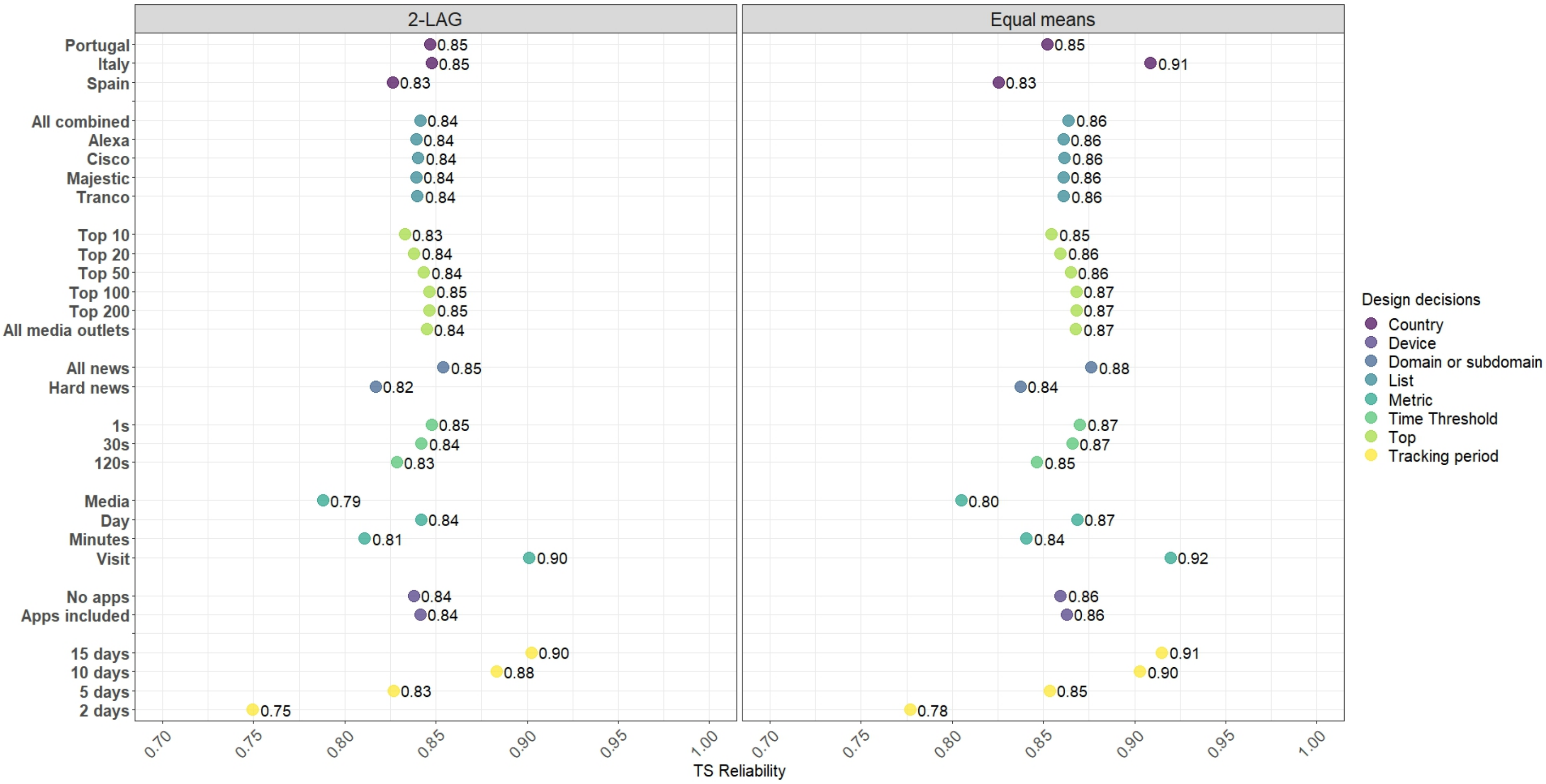

Figure 5.5: Average predicted reliability coefficients

the tracking period, the higher the reliability of the methods. However, this increase does not follow a strictly linear trajectory, as evidenced by the present yet not overly substantial difference between the measures computed with 15 and 10 days of tracking data. Furthermore, the choice of metric significantly influences the reliability of measurements. Computing the average number of media outlets consumed leads to the lowest average predicted reliability, while using the average number of visits clearly appears to be the best option. Interestingly, this finding contrasts with the advocacy for a list-based approach in surveys, often touted as superior to visit or time-based counts. In addition, considering any URL published by a media outlet as news leads to an average increase in the percentage of the measure explained by the true score of 3 to 4 percentage points, compared to only using those URLs manually identified as hard news. Furthermore, the predicted reliability of measures seems to be, on average, associated with the country. Specifically for the equal means model, Italy seems to present substantially higher reliability coefficients, with Spain under-performing. This implies that cross-national comparisons of standardized coefficients might exhibit some bias if not corrected for measurement errors (Bosch and Revilla, 2021). Besides these design decisions, Figure 5 shows that all other design options do not have a strong effect on the size of the measurement errors of media exposure measures computed with web tracking data.

Focusing on predictive validity, a key observation stemming from Figure 6 is that although minor fluctuations are evident, the size of all the predicted coefficients is small. However, beneath these variations lie intriguing patterns. Across the array of models, Figure 6 demonstrates that the predicted validity coefficients exhibit fluctuations across countries, and across the metrics and length of the tracking periods employed. A more in-depth analysis of the variable importance within the model (outlined in SM 7) provides similar insights, thus reinforcing these observations.

While the outcomes from the two models differ slightly, they both suggest a consistent trend: the magnitude of the association between media exposure and political knowledge is greater, on average, when employing a metric based on average visits (with the minutes metric demonstrating notably weaker performance). Furthermore, both models indicate that extended tracking periods do not inherently result in higher predicted regression coefficients. Surprisingly, either the 10-day or 5-day tracking measures surpass those computed using a 15-day tracking dataset.

Variations are also evident across the models. The fixed effects model indicates that the projected association between media exposure and political knowledge is, on average, least pronounced in Spain. Conversely, the cross-sectional model presents a contrary scenario. Additionally, the cross-sectional model uncovers a connection between other design choices and the average predicted coefficient of the measurements. Notably, considering any URL published by a media outlet as news yields a higher standardized regression coefficient compared to exclusively utilizing URLs manually identified as ”hard” news or political in nature. Furthermore, utilizing information from the top 50 to 200 most popular news media outlets in each country corresponds to higher regression coefficients compared to the usage of only the top 10 or 20 outlets.

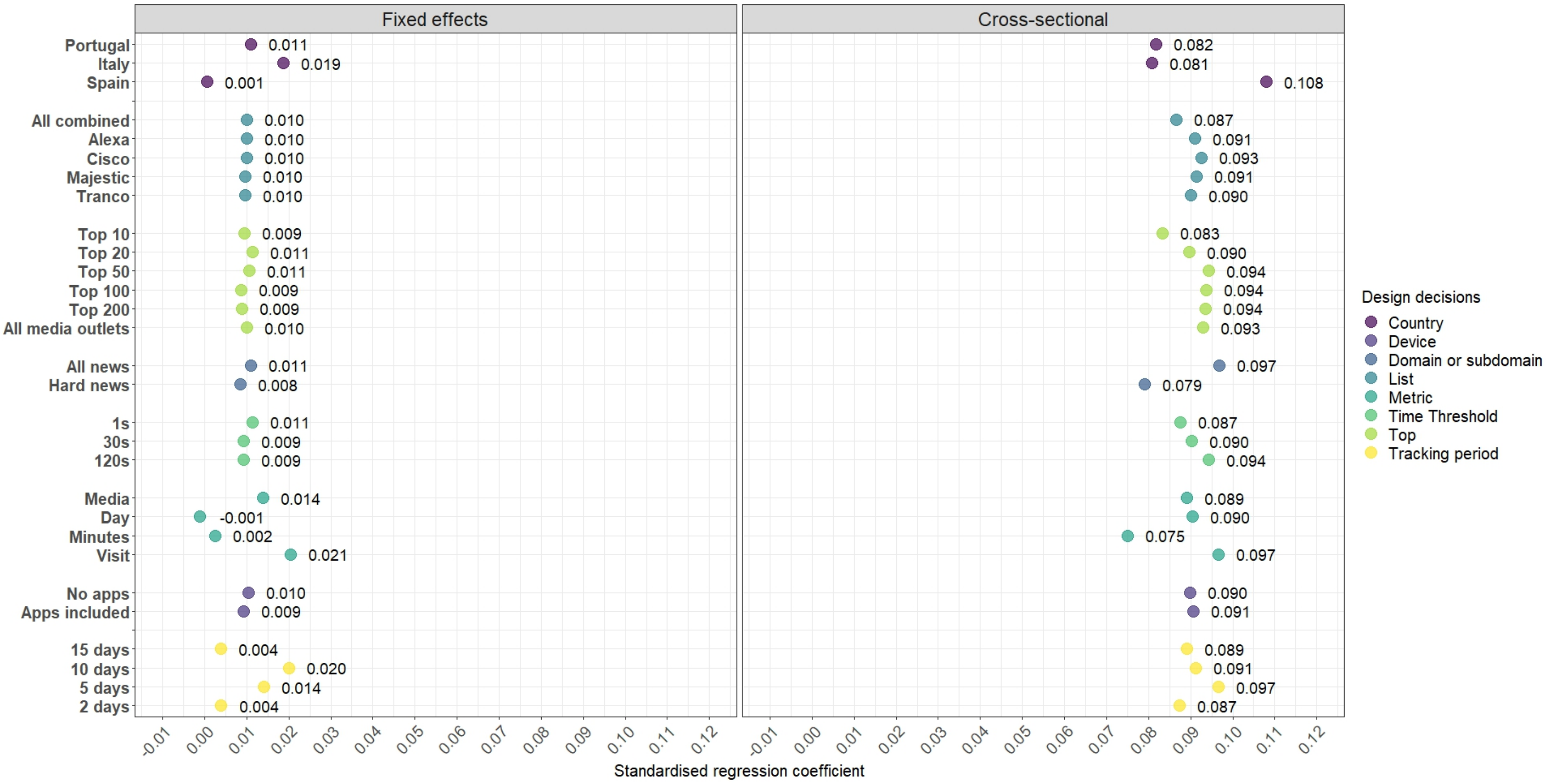

Figure 5.6: Average predicted validity coefficients

7. Discussion

The measurement of online media exposure is of paramount importance for studying the uses and effects of online media. In the recent years, there has been much excitement about the possibility of obtaining unbiased measure of online media exposure using web tracking data. Nonetheless, considering that past research has shown that web tracking data can be affected by errors, in this paper I explored to what extent media exposure measures computed with web tracking data are valid and reliable. In addition, through an adaptation of the multiverse analysis approach, the paper also explored the association that several key design decisions have on the reliability and validity of media exposure measurements.

7.1. Main results

First, results show that the reliability of the multiverse of measurements explored is, in general, quite high. Across the different models studied, the median reliability of the multiverse of measurements ranged between 0.85 and 0.87. Hence, overall, the explored measures of media exposure capture around 86% of the variance of their true score. Although variation exists across measurements, with some performing below this average, most of the measurements present reliability coefficients deemed as good or excellent (DeCastellarnau et al., 2017). Considering these results, the measurements explored generally present high reliability coefficients. To put it in context, Bartels (1993) reported a reliability of .75 for a traditional television exposure measure, while Dilliplane et al. (2012) found that a list-based measure assessing the number of political programs viewed on TV yielded a reliability of .83. Hence, on average the media exposure measurements computed with web tracking data explored in the multiverse analysis show, in general, higher reliability estimates.

Second, regarding the predictive validity of the multiverse of measurements, results show that overall, the association between the measures of media exposure and political knowledge is weak. Specifically, for the fixed effect model, the median standardised coefficient is of 0.01, with a substantial proportion of the measurements yielding small but negative coefficients. The cross-sectional model, additionally, produced a slightly higher but still small median coefficient of 0.09. These results are almost identical than the ones obtained using a survey self-report. Past research has considered that, for a media exposure measure to be valid, it should show a significant and substantial association with political knowledge. Based on this assumption, most self-reported measures of media exposure have been criticized because of their low predictive power. Results suggest that this lack of predictive power is not limited to self-reports, but it is also found when using web tracking data. Based on past research's approach, these results would suggest that the explored measures of media exposure derived from web tracking data exhibit low predictive validity, irrespective of the regression model employed. Nonetheless, this could also mean that the common approach used to measure the predictive validity of media exposure measures is, to some extent, flawed. As Prior (2009) already said, we do not know the "true" association between media exposure and political knowledge. We assume that there should exist one, and that its size should not be small. However, there is a chance that consuming media does not necessarily lead to an increase of political knowledge.

Or even more feasible, that small fluctuations of media exposure are not related to a change in the knowledge that people have about how their basic political system works. These results could be an indication that the true relationship is indeed small, indicating that the lack of predictive power of surveys is not an indictment of their validity.

Besides describing the overall validity and reliability of the multiverse of measurements, results also help understand the key design choices that significantly influence the validity and reliability of the explored measurements. Results from the random forest of regression trees models show that some design choices do affect both the reliability and validity coefficients obtained. Based on these results, we can extract some interesting patters:

1. The length of the tracking periods is significantly associated with the reliability of the measurements, and their validity. While the average predicted reliability of a measure computed with two days of data is of 0.75/0.78, this value goes up to 0.90/0.91 when using 15 days of data. This relationship, however, is not observed when focusing on the validity coefficients. In this case, an extended tracking period does not result in higher predicted coefficients: five days of data can yield a stronger predictive power than 15 days.

2. The choice of metric significantly influences the reliability and validity of measurements. Using an average count of the number of visits to media outlets seems to yield the highest reliability and validity coefficients.

3. The approach used to decide what URL should be considered as "written news" is also significantly linked to the reliability and validity of measures. Specifically, considering any URL published by a media outlet as news leads to an average increase in the percentage of the measure explained by the true score, and an increase on the cross-sectional association between media exposure and political knowledge.

4. Both the reliability and validity of the explored measurements are affected by the country where measurements were computed. This indicates that the measurement properties of media exposure measures computed with web tracking data vary across countries. Hence, any cross-national comparison of standardised coefficients must take this into account or run the risk of being flawed (Bosch and Revilla, 2021).

5. Other design choices present significant differences across their available options, but most of them are rather small. For instance, focusing only on the top 10 or 20 most popular media outlets leads to slightly smaller reliability and validity estimates. And using stricter time thresholds when defining what can constitute as a visit is associated with smaller reliability coefficients, and a higher cross-sectional association between media exposure and political knowledge.

6. The list used to identify and rank news media outlets, as well as the inclusion (or not) of app data, seem to be mostly irrelevant choices.

7.2. Limitations and future research

These results present some limitations. First, participants were recruited using an opt-in online panel of already tracked participants. Although this is the most common approach in the literature to recruit participants for web tracking studies, and the most popular tracking provider in the market was used, it is unclear whether these results would replicate when conducted on another opt-in panel using a different recruiting and tracking approach, as well as on a probability-based sample of less experienced respondents. Second, the limited number of survey waves, and the reduced spacing between waves (three months), mean that there might not be enough margin to capture any kind of political knowledge formation. Additionally, the used measure of political knowledge, albeit common, might not be the best suited one to assess the type of political knowledge that might be gained through the consumption of news. Future research could explore whether these results hold when asking participants about knowledge of current political events, instead of only focusing on basic questions about the political system. Third, the Quasi-Markov Simplex Model depends on some potentially unrealistic and strict assumptions. Although in this study I have tested the effect of relaxing some of these assumptions on the fit and overall performance of the models, following Cernat et al. (2021) approach, this approach has been applied on the macro-level. Considering that I have run the same models for the 2,631 measurements, it is feasible to assume that the two models used (2-LAG and equal means) might not be the best performing models for all measurements. Ideally, I would have manually tested, for each measurement, the best model to use. Nonetheless, given the large volume of measurements, I applied the two models performing the best on average, for all measurements. Fourth, the reliability estimated by the QSM measures the percentage of variance due to the true score as opposed to random error. Therefore, even if the results suggest that the reliability of the explored measurements is very high, there is still the chance that these measurements are affected by systematic errors and, hence, severely biased. This could partially explain the weak association between media exposure and political knowledge. If systematic errors are present, measures could be very consistent across time, while still being severally biased. Future research should consider testing the measurement quality of media exposure measurement created with web tracking data using methods that allow disentangling the systematic and random components of measurement errors, such as MultiTrait-MultiMethod models (e.g., Bosch and Revilla 2021). Fifth, although the multiverse of measurements has considered many of the most important design decisions that researchers face when designing media exposure measurements, I have not been able to cover all of them. Specifically, much research has measured media exposure by tracking participants only on their desktop devices, excluding mobile devices. I was not able to test this design feature, since the sample was composed by people tracked in both mobile and desktop devices and excluding those tracked in only one device would have led to a substantial drop in the size of the sample. Sixth, interactions were not directly considered when fitting the random forest of regression trees. This is the common approach, given the random forests model's nonparametric nature, lack of assumption of independence between variables, and sequential approach. Nonetheless, new research suggests that interaction forests can deliver better predictions than conventional random forests (Hornung and Boulesteix, 2022). Lastly, the media exposure measure explored in this study (i.e., "written news media") is very narrow. Research has suggested that most news and political content that people are exposed online is not encountered on news media outlets, but on other webpages (Wojcieszak et al., 2023). Additionally, news are increasingly consumed via videos and podcasts (Newman et al., 2021), which are not considered in this study. A more comprehensive definition of media exposure should be explored in future research, to test the performance of web tracking measures when focusing on a richer but more complex concept of interest.

7.3. Conclusions and practical recommendations

First, web tracking measures of media exposure exhibit, on average, a very high reliability. Most of the explored measures show very small measurements errors, capturing most of the variance of the measurement's true score. Maybe even more encouraging, only a very small minority of the explored measurements presents reliabilities below 0.70, meaning that it is very unlikely for web tracking measures of media exposure to be affected by sizable random measurement errors. Although this is an overall improvement to self-reports, the main critique of survey measures of media exposure was never their reliability. It has been established that these measures present acceptable or even good reliabilities (see Bartels 1993; Dilliplane et al. 2012). Collecting web tracking data is more complex, expensive, and ethically challenging than collecting survey data. Hence, researchers must consider the trade-off between these drawbacks, and the potential reduction of random measurement errors. This is not to say that web tracking data is not worth it, it might still be. But the gains need to be considered more carefully than simply assuming that this new source of data is the gold standard.

Compared with the high reliability found, the association between media exposure measures computed with web tracking data and political knowledge is weak. This can mean that web tracking measures are not an improvement to self-reports, and/or that the critiques to self-reports based on this commonly used approach to measure predictive validity have been overstated all along. Whatever of the hypothesis leads to the same conclusion, nonetheless: we need to improve the way that we think about validity, and the approaches used to assess it. Changing self-reported measures for web tracking ones is not justified based on this paper's results, either because both are equally bad, or because we have too little understanding on the validity of both approaches to establish any evidence-based comparison. This is in line with what other researchers have found, and advocated, when using data donations to measure concepts related to mobile usage (Bosch et al., 2023).

Finally, I recommend researchers to embrace the multiverse of measurements approach. Results show that the reliability and validity of measurements substantially fluctuate across design choices. It is to be expected, hence, that substantive results might also considerably vary. An advantage of web tracking data over survey self-reports is the feasibility of creating a large multiverse of measurements. Once the raw data from the web trackers is collected, it is mostly inexpensive and effortless to go from one measurement to thousands of them. In the current context of the social sciences, with many disciplines experiencing replication crises, the multiverse of measurements approach allows researchers to embrace uncertainty, improve the transparency of their research, and provides readers with a rich amount of data to form their own opinions and conclusions. Beyond this, and for those researchers unable or unwilling to apply the multiverse of measurements approach, the results provide practical evidence of potential best practices when designing web tracking measures of media exposure: use at least 10 to 15 days of tracking data, use a count of visits as the metric, and consider all URLs published by new media outlets as "written news", especially if the only alternative is to manually identify the URLs that publish "hard" news.

Acknowledgements: I would like to thank Patrick Sturgis, Jouni Kuha, Melanie Revilla, and Mariano Torcal for their comments and support

Funding: Fundación BBVA; H2020 European Research Council, Grant/Award Number: 849165; Ministerio de Ciencia e Innovación, Grant/Award Number: PID2019-106867RB-I00/AEI/10.13039/501100011033

Data: The data that support the findings of this study is openly available in OSF at https://osf.io/3t7jz/ (DOI: 10.17605/OSF.IO/3T7JZ). More information can be found in the following paper

Bibliography

Alwin, Duane F. 2007. *Margins of error a study of reliability in survey measurement*. Wiley.

Araujo, Theo, Anke Wonneberger, Peter Neijens, and Claes de Vreese. 2017. "How Much Time Do You Spend Online? Understanding and Improving the Accuracy of Self-Reported Measures of Internet Use." *Communication Methods and Measures* 11:173–190.

Bach, Ruben L., Christoph Kern, Denis Bonnay, and Luc Kalaora. 2022. "Understanding political news media consumption with digital trace data and Natural Language Processing." *Journal of the Royal Statistical Society Series A: Statistics in Society* 185.

Bartels, Larry M. 1993. "Messages received: The political impact of media exposure." *American Political Science Review* 87:267–285.

Biemer, Paul P. 2010. "Total survey error: Design, implementation, and evaluation." *Public Opinion Quarterly* .

Bosch, Oriol, Marc Asensio, and Caroline Roberts. 2023. *Data donations, are they worth the effort? The accuracy and validity of smartphone usage measures computed with self-reports and data donations*.

Bosch, Oriol J. and Melanie Revilla. 2021. “The Quality of Survey Questions in Spain: A Cross-National Comparison.” *Revista Española de Investigaciones Sociológicas* 175:3–26.

Bosch, Oriol J. and Melanie Revilla. 2022a. “The challenges of using digital trace data to measure online behaviors: lessons from a study combining surveys and metered data to investigate affective polarization.” *SAGE Research Methods Cases* .

Bosch, Oriol J. and Melanie Revilla. 2022b. “When survey science met web tracking: Presenting an error framework for metered data.” *Journal of the Royal Statistical Society Series A: Statistics in Society* 185.

Bosch, Oriol J., Melanie Revilla, Anna DeCastellarnau, and Wiebke Weber. 2018. “Measurement reliability, validity, and quality of slider versus radio button scales in an online probability-based panel in Norway.” *Social Science Computer Review* 37:119–132.

Breiman, Leo, Jerome H Friedman, Richard A Olshen, and Charles J Stone. 1984. *Classification And Regression Trees*. Routledge.

Cernat, Alexandru, Peter Lugtig, Nicole Watson, and S.C. Noah Uhrig. 2021. “Assessing and relaxing assumptions in quasi-simplex models.” *Measurement Error in Longitudinal Data* p. 155–172.

Chang, Lin Chiat and Jon A. Krosnick. 2003. “Measuring the frequency of regular behaviors: Comparing the “Typical week” to the “past week”.” *Sociological Methodology* 33:55–80.

Cleary, T. Anne, Robert L. Linn, and G. William Walster. 1970. “Effect of reliability and validity on power of statistical tests.” *Sociological Methodology* 2:130.

DeCastellarnau, Anna. 2017. “A classification of response scale characteristics that affect data quality: a literature review.” *Quality & Quantity* 52:1523–1559.

DeCastellarnau, Anna, Anna DeCastellarnau, and Melanie Revilla. 2017. “Two Approaches to Evaluate Measurement Quality in Online Surveys: An Application Using the Norwegian Citizen Panel.” *Survey Research Methods* 11:415–433.

Dilliplane, Susanna, Seth K. Goldman, and Diana C. Mutz. 2012. “Televised exposure to politics: New measures for a fragmented media environment.” *American Journal of Political Science* 57:236–248.

Felderer, Barbara, Ludwig Bothmann, Lydia Repke, Jonas Schweisthal, and Wiebke Weber. 2023. *European Survey Research Association*. European Survey Research Association.

Golder, Scott A. and Michael W. Macy. 2014. “Digital Footprints: Opportunities and challenges for online social research.” *Annual Review of Sociology* 40:129–152.

Goldman, Seth K. and Stephen M. Warren. 2019. “Debating how to measure media exposure in surveys.” *The Oxford Handbook of Electoral Persuasion* p. 997–1015.

González-Bailón, Sandra and Michael A. Xenos. 2022. “The blind spots of measuring online news exposure: A comparison of self-reported and observational data in nine countries.” *Information, Communication amp;amp; Society* 26:2088–2106.

Guess, Andrew M. 2015. “Measure for Measure: An Experimental Test of Online Political

Media Exposure." *Political Analysis* 23:59–75.

Guess, Andrew M., Brendan Nyhan, Zachary O'Keeffe, and Jason Reifler. 2020. "The sources and correlates of exposure to vaccine-related (mis)information online." *Vaccine* 38:7799–7805.

Hornung, Roman and Anne-Laure Boulesteix. 2022. "Interaction forests: Identifying and exploiting interpretable quantitative and qualitative interaction effects." *Computational Statistics amp;amp; Data Analysis* 171:107460.

Howison, James, Andrea Wiggins, and Kevin Crowston. 2011. "Validity issues in the use of social network analysis with Digital Trace Data." *Journal of the Association for Information Systems* 12:767–797.

Jungherr, Andreas. 2019. "Normalizing Digital Trace Data." In *Digital Discussions*.

Kershner, James W. 2012. *The elements of news writing*. Pearson Allyn amp; Bacon.

Liaw, Andy and Matthew Wiener. 2002. "Classification and regression by randomForest." *R news* pp. 18–22.

Lord, F. M., M. R. Novick, and Allan Birnbaum. 1968. *Statistical theories of mental test scores*. Addison-Wesley.

Mangold, Frank, Sebastian Stier, Johannes Breuer, and Michael Scharkow. 2021. "The overstated generational gap in online news use? A consolidated infrastructural perspective." *New Media amp;amp; Society* 24:2207–2226.

Michaud, Agnalys, Oriol J. Bosch, and Nicolas Sauger. 2023. "Can survey scales affect what people report as a fair income? evidence from the cross-national probability-based online panel Cronos." *Social Justice Research* 36:225–262.

Newman, N, R Fletcher, A Schulz, S Andi, C T Robertson, and R K Nielsen. 2021. *Reuters Institute digital news report 2021. Reuters Institute for the study of Journalism*.

Oberski, Daniel. 2014. "lavaan.survey: An r package for complex survey analysis of structural equation models." *Journal of Statistical Software* 57.

Parry, Douglas A., Brittany I. Davidson, Craig J. Sewall, Jacob T. Fisher, Hannah Mieczkowski, and Daniel S. Quintana. 2021. "A systematic review and meta-analysis of discrepancies between logged and self-reported digital media use." *Nature Human Behaviour* 5:1535–1547.

Peterson, Erik, Sharad Goel, and Shanto Iyengar. 2018. "Echo Chambers and Partisan Polarization: Evidence from the 2016 Presidential Campaign."

Pew Research Center. 2020. "Measuring News Consumption in a Digital Era." Technical report.

Price, Vincent and John Zaller. 1993. "Who gets the news? alternative measures of news reception and their implications for research." *Public Opinion Quarterly* 57:133.

Prior, Markus. 2009. "Improving Media Effects Research through Better Measurement of News Exposure." *The Journal of Politics* 71:893–908.

Reinemann, Carsten, James Stanyer, Sebastian Scherr, and Guido Legnante. 2011. "Hard and soft news: A review of concepts, operationalizations and key findings." *Journalism*

13:221–239.

Reiss, Michael V. 2022. "Dissecting non-use of online news – systematic evidence from combining tracking and automated text classification." *Digital Journalism* 11:363–383.

Revilla, Melanie, Carlos Ochoa, and Germán Loewe. 2017. "Using Passive Data From a Meter to Complement Survey Data in Order to Study Online Behavior." *Social Science Computer Review* 35:521–536.

Saris, Willem E. and Irmtraud N Gallhofer. 2007. *Design, Evaluation, and Analysis of Questionnaires for Survey Research*. Hoboken, NJ, US: John Wiley Sons, Inc.

Saris, Willem E. and Irmtraud N. Gallhofer. 2014. *Design, Evaluation, and Analysis of Questionnaires for Survey Research*. Hoboken, NJ, US: John Wiley Sons, Inc., second edi edition.

Saris, Willem E., Daniel Oberski, Melanie Revilla, Diana Zavalla, Laur Lilleoja, Irmtraud Gallhofer, and Tom Gruner. 2011. "The development of the program SQP 2.0 for the prediction of the quality of survey questions."

Scharkow, Michael. 2016. "The Accuracy of Self-Reported Internet Use—A Validation Study Using Client Log Data." *Communication Methods and Measures* 10:13–27.

Scharkow, Michael, Frank Mangold, Sebastian Stier, and Johannes Breuer. 2020. "How social network sites and other online intermediaries increase exposure to news." *Proceedings of the National Academy of Sciences* 117:2761–2763.

Schwarz, N. 2001. "Asking questions about behavior: Cognition, Communication, and questionnaire construction." *The American Journal of Evaluation* 22:127–160.

Slater, Michael D. 2004. "Operationalizing and analyzing exposure: The foundation of Media Effects Research." *Journalism amp;amp; Mass Communication Quarterly* 81:168–183.

Smith, Brianna, Scott Clifford, and Jennifer Jerit. 2019. "TRENDS: How internet search undermines the validity of political knowledge measures." *Political Research Quarterly* 73:141–155.

Steegen, Sara, Francis Tuerlinckx, Andrew Gelman, and Wolf Vanpaemel. 2016. "Increasing transparency through a multiverse analysis." *Perspectives on Psychological Science* 11:702–712.

Stier, Sebastian, Johannes Breuer, Pascal Siegers, and Kjerstin Thorson. 2019. "Integrating Survey Data and Digital Trace Data: Key Issues in Developing an Emerging Field." *Social Science Computer Review* p. 089443931984366.

Torcal, Mariano, Emily Carty, Josep Maria Comellas, Oriol J. Bosch, Zoe Thomson, and Danilo Serani. 2023. "The dynamics of political and affective polarisation: Datasets for Spain, Portugal, Italy, Argentina, and Chile (2019-2022)." *Data in Brief* 48:109219.

Wiley, David E. and James A. Wiley. 1970. "The estimation of measurement error in panel data." *American Sociological Review* 35:112.

Wojcieszak, Magdalena, Ericka Menchen-Trevino, Bernhard Clemm von Hohenberg, Sjifra de Leeuw, João Gonçalves, Sam Davidson, and Alexandre Gonçalves. 2023. "Non-News websites expose people to more political content than news websites: Evidence from browsing data in three countries." *Political Communication* p. 1–23.

Zaller, John. 2002. "The statistical power of election studies to detect media exposure effects in political campaigns." *Electoral Studies* 21:297–329.

Chapter 6

Conclusion

We are currently in an exhilarating era for the social sciences. The proliferation of digital data has revolutionized the research landscape, expanding the horizons of what social scientists can investigate and how they conduct their studies. This transformative wave has disrupted various aspects of the field, reshaping curricula, and influencing research funding priorities. The enthusiasm surrounding digital trace data, driven by its granularity and objectivity, has been undeniably warranted. However, this enthusiasm has also, to some extent, obscured the potential challenges associated with leveraging this data for rigorous social scientific research.

To fully harness the transformative potential of this ”measurement revolution” (Watts, 2012) offered by digital trace data in the social sciences, we must subject this new data source to the same rigorous scrutiny in terms of conceptualization and methodology as we do with traditional data sources (Gerring, 2016; Hand, 2004). While digital trace data holds tremendous promise, it still has a considerable distance to cover before becoming a widely accepted and credible mainstream data source, firmly entrenched in the core of social science discourse and funding priorities.

This thesis makes a significant contribution to this ongoing effort. Specifically, it provides a comprehensive theoretical and empirical examination of the quality of web tracking data, shedding light on its associated errors and their ramifications. As a methodological thesis, this work underscores a crucial point: the initial promises of web tracking data must be viewed through a lens of scepticism. Like any other data source, web tracking data is imperfect. This observation is not intended to dampen enthusiasm or convey negativity; instead, it represents a realistic perspective. This thesis does not aim to criticize; rather, by acknowledging the imperfections of web tracking data, it paves the way for the development of approaches that enhance our understanding of this data source and establish best practices for its utilization. By aligning web tracking research more closely with the established practices of imperfect yet well-understood data collection methodologies such as surveys, this thesis contributes to the integration of digital trace data into the social science toolkit and workflow.

Paper 1, in Chapter 3, serves as the foundation for a more mature and improved use of web tracking data in the social sciences. It addresses a historical gap where researchers utilizing web tracking data often lacked a clear understanding of the entire process, from data collection design to the computation of statistical estimates. This knowledge gap was understandable, as social researchers typically lacked familiarity with the intricacies of digital trace production processes, while computer scientists lacked training in social research methodology. Consequently, few researchers employed strategies to measure, mitigate, or transparently report errors. Moreover, methodological research quantifying these errors and assessing biases in web tracking research was exceptionally scarce.

In this context, there was an urgent need to establish a systematic description of the data generation and analysis processes for web tracking data, along with a classification of potential sources of bias and variance at each stage of this process. Drawing on the extensive scholarship in survey methodology, psychometrics, and social statistics, Paper 1 adopts a survey science perspective to examine web tracking data. The result is the Total Error framework for digital traces collected with Meters (TEM). The TEM elucidates that making statistical inferences with web tracking data is an intricate endeavour, involving numerous understudied design decisions that often go unacknowledged. It also underscores that, theoretically, web tracking data can be vulnerable to a plethora of errors that might bias research outcomes and conclusions. Consequently, the TEM validates the early critiques of digital trace data voiced by researchers like Jungherr (2019) and highlights the imperative need for a more rigorous measurement approach when employing this data source for scientific research.

A glaring oversimplification in the existing literature is the assumption that web tracking data measures individual-level behaviours. This assumption is fundamentally incorrect. As elaborated in Paper 2 (Chapter 4), web tracking data is captured at the device level, not the individual level. Therefore, we can only consider web tracking data as representative of a person's complete online behaviours if two conditions are met: 1) the device is used exclusively by the person of interest, and 2) all the devices they use for online activities are being tracked. Focusing on the second condition, achieving a comprehensive understanding of individuals' online behaviour necessitates collecting data from all the devices they use. Nevertheless, this is a complex task. As demonstrated in the first and second papers of this thesis, achieving perfect device coverage is highly challenging. Tracking technologies are imperfect, and there is no universal option that can be installed and configured across all types of devices and operating systems. Furthermore, participants might not be willing to cooperate, and even when they are, the installation of tracking technologies on all their devices can be logistically challenging. Failure to track participants on some of their online devices leads to incomplete measurements, which can introduce bias by deviating observed behaviours from the participants' true actions. Despite its conceptual simplicity, tracking undercoverage has not been acknowledged in the field of web tracking data research.

Paper 2 rectifies this oversight. It demonstrates that tracking undercoverage is widespread and consequential, and addressing this source of error is imperative for conducting sound research. The paper reveals that across the three countries examined, 74% of people had at least one device they used for online activities that went untracked. Consequently, given that this thesis collected data using the most common online panels and web tracking technologies prevalent in the literature, this implies that a significant portion of past research may have been based on data significantly affected by tracking undercoverage. Is this problematic? According to the results from the simulations conducted in Paper 2, it is indeed. Across a range of univariate and multivariate statistics typically computed in media effects research, Paper 2 demonstrates that tracking undercoverage introduces substantial biases. For instance, considering the level of undercoverage observed in the TRI-POL dataset, it reveals that the estimated average time participants spend on the Internet is underestimated by approximately 30%.

These findings pose significant challenges for the field of web tracking research. Firstly, they debunk the assumption that web tracking data is largely unbiased, calling into question the legitimacy of many claims made using this data. Phenomena that have been downplayed based on web tracking data, such as the prevalence of Fake News exposure, may actually be more widespread. If these issues are not addressed, the credibility of a substantial portion of claims made with web tracking data may be rightfully challenged. Additionally, much of the research asserting that surveys are biased when measuring online behaviours has been built on comparisons between self-reports and web tracking data, or other forms of digital traces. These studies typically suggest that survey participants "overreport" their behaviours, such as news exposure, compared to web tracking data. However, based on the findings of Paper 2, it is plausible that the web tracking estimates used in these comparisons were themselves underestimated. If we continue to believe that self-reports are somewhat overestimated, it is possible that the ”true” behaviours of participants fall somewhere between the underestimated web tracking measures and the overestimated self-reports. This implies that self-reports may be less biased than previously assumed.

Secondly, the evidence presented in Paper 2 clearly exposes deficiencies in the practices and procedures followed by both online fieldwork companies and researchers. Companies offering metered panels frequently fall short in capturing a comprehensive view of their panellists' online activities. This shortcoming often stems from limitations in their tracking technologies, which are seldom addressed. Alarmingly, these companies typically do not provide information about the level of tracking coverage within their panels and rarely disclose how they address these issues. On the researcher's side, the prevailing practice has been to accept the data provided by these companies uncritically. No research paper has ever reported the prevalence of tracking undercoverage in web tracking samples or discussed the potential limitations stemming from this issue in their results. If we compare this with the standard practice when using surveys, it would be akin to not reporting response or participation rates—a practice widely recognized as suboptimal.

The simplistic measurement approach used with web tracking data, it seems, has far-reaching negative consequences. These results, coupled with the theoretical contributions of Paper 1, underscore the need for a comprehensive examination of the measurement properties of web tracking data. In survey research, substantial time and resources have been devoted to understanding the validity and reliability of measures employed. This endeavour is essential for ascertaining whether the measures accurately capture the concept of interest, as well as to understand the extent random errors might inflate variance, potentially distorting statistical estimates. Additionally, assessing the validity and reliability of measures allows for standardized comparisons between different measurement methods. When working with imperfect data sources, it is key to understand what design choices maximise the validity and reliability of the measurements used. Having a good understanding of the measurement properties of measures designed with different combinations of design choices can aid researchers in understanding the expected quality of published research and designing their own measures.

Paper 3, in Chapter 5, marks a significant step towards narrowing the existing gap in our understanding of the measurement properties of web tracking data in comparison to well-established data sources, particularly surveys. Its primary focus centres on media exposure measures. Firstly, the paper challenges the prevailing assumption that translating a construct into a web tracking measurement is a straightforward process. It demonstrates that creating these measurements—deciding which traces to employ, how to transform them, and in what manner to amalgamate them—involves a multitude of uncharted decisions. In fact, the paper reveals that there are potentially thousands of different measures for assessing media exposure within web tracking data. However, what becomes evident is a conspicuous dearth of empirical evidence to guide the selection of appropriate design decisions in this context. Within this landscape of uncertainty, it becomes challenging to make informed decisions regarding measurement approaches when working with web tracking data. The lack of established guidelines and empirical evidence surrounding these critical decisions envelops the results derived from web tracking data with a level of uncertainty that distinguishes it from the use of traditional data sources like surveys.

Secondly, Paper 3 raises questions about the presumption that web tracking measures of media exposure inherently possess greater validity than survey self-reports. Prior research has contended that, for a media exposure measure to be deemed valid, it must exhibit a significant and substantial association with political knowledge. Building upon this assumption, most self-reported measures of media exposure have faced criticism due to their limited predictive power. Our findings challenge this belief by demonstrating that this lack of predictive power is not unique to self-reports; it is also prevalent when utilizing web tracking data. This suggests that web tracking measures might not represent a substantial improvement over self-reports. Alternatively, it implies that critiques of self-reports, premised on the "gold standard" approach to predictive validity, may have been overstated. Regardless of which hypothesis holds true, the shift from self-reports to web tracking data seems unjustifiable based on the validity of surveys and web tracking, either because both approaches are equally suboptimal, or because our understanding of their validity is insufficient to establish an evidence-based comparison.

Lastly, Paper 3 reveals that web tracking measures of media exposure generally demonstrate very high reliability. Thus, while these results acknowledge that web tracking measures of media exposure are subject to random measurement errors, the magnitude of these errors is minimal. Furthermore, among the multitude of measurements explored, only a small fraction exhibits low reliability estimates, underscoring the inherent stability of web tracking measures to some extent. Although this represents an overall improvement over self-reports, it is important to note that the primary critique of survey-based measures of media exposure has never revolved around their reliability. It has been well-established that these measures typically demonstrate acceptable or even commendable levels of reliability. Consequently, it remains uncertain whether this slight improvement in reliability alone justifies a complete shift in how social scientists collect data regarding individuals' online media consumption behaviours.

An optimistic way forward

This thesis may appear critical and, at times, even pessimistic in its tone. However, its findings are fundamentally optimistic. The notion that digital trace data could serve as a flawless and unbiased data source was, from its inception, more of an illusion than a reality. The pioneering work of those who ventured into the realm of digital data opened doors and ignited excitement within the social sciences, but their efforts were never intended to create a perfect solution. Just as digital trace data was never meant to remain unquestionably perched upon a ”gold standard” pedestal indefinitely, this thesis reaffirms that there is ample room for improvement and maturation in this field.

Indeed, this thesis embodies optimism by demonstrating that we can move the field forward. It reveals that we have the capacity to approach digital trace data, and by extension, computational social sciences, with greater maturity and discernment. Optimistically, the thesis underscores the adaptability of the extensive methodological and statistical knowledge developed for traditional data sources, such as surveys, to enhance our comprehension of the quality and errors inherent in web tracking data. As a result, researchers and practitioners are encouraged to view digital trace data through the lens of survey science and psychometrics, enhancing their utilization of digital traces for social scientific research.

Furthermore, the thesis offers practical tools to aid researchers in this endeavour. Firstly, the Total Error Framework presented can serve as a valuable resource for planning the design of web tracking projects, as well as for devising strategies to identify and address various sources of error. Secondly, the thesis introduces a method to recognize, report, and simulate the bias stemming from tracking undercoverage, a critical error source. Lastly, the last paper introduces a guide to measure the predictive validity and true-score reliability of web tracking measures in a manner comparable to surveys.

With this thesis as a foundation for future progress, researchers can consider the adaptation of other methodologies. A notable example is MultiTrait-MultiMethod (MTMM) models, which have empowered survey scientists to refine their understanding of the measurement quality of survey questions and identify the methods least susceptible to errors. MTMMs have even been used to predict the quality of prospective survey questions, as demonstrated by the Survey Quality Predictor (SQP). Extending and adapting MTMMs could offer valuable insights into the measurement quality of web tracking data. These models might even evolve to simultaneously estimate the quality of survey and web tracking data for the same concepts of interest, akin to the work conducted by Oberski et al. (2017) for administrative data.

The inherent nature of web tracking data also presents an intriguing opportunity for methodological research. This thesis highlights how the flexibility and granularity of web tracking data can be harnessed for methodological purposes, offering solutions to overcome some of the limitations inherent in surveys. Leveraging these characteristics, however, requires a different set of tools and methods than the ones that survey methodologists are normally used to use. By looking at methods normally used in the data sciences and more computationally demanding fields, the field can find creative ways to go beyond what has been possible for surveys. For instance, Paper 2 demonstrates how the sheer volume of data generated by web trackers, coupled with insights into certain types of errors, enables the use of simulation techniques to estimate the biases affecting web tracking data. In the realm of surveys, we face the challenge of attempting to measure something inside people's minds, making it inherently elusive. Unlike surveys, web tracking data is produced through known mechanisms. Simulations can manipulate these mechanisms to quantify the impact of errors on data production. An illustration of this approach is presented in Paper 2, which focuses on tracking undercoverage. By identifying individuals with all their devices tracked and collecting data separately for mobile devices and PCs, we were able to simulate the effects on various statistics when certain devices were not tracked. This approach could similarly be applied to simulate errors stemming from other sources discussed in the TEM, such as misclassification, shared devices, or technological errors.

Paper 3 further underscores the opportunities that stem from the granularity of web tracking data. In traditional surveys, generating multiple measures for a single concept of interest is a resource-intensive and complex undertaking. Surveys are bound by constraints on the number of questions that can be included in a questionnaire, and participants' memory limitations can affect the independence of responses to repeated measures. Typically, surveys include only a limited number of measures for select concepts, often with no more than three repetitions. In contrast, web tracking data is unrestricted by such limitations. When researchers have access to the raw web tracking dataset, they can craft hundreds, if not thousands, of distinct measures for a single concept of interest. This flexibility empowers researchers to conduct analyses across a multiverse of measurements within a single project. This can serve various purposes, such as reporting result uncertainties in web tracking research or assessing the measurement properties of diverse measures and focusing on those demonstrating the highest quality. In essence, the capability to perform analyses using not just one arbitrary measurement, but the entire multiverse, positions web tracking research to establish a more extensive corpus of methodological evidence than surveys have ever provided. To illustrate, decades of MTMM experiments in numerous countries have paved the way for the creation of the SQP program. SQP can predict the quality of survey questions based solely on their characteristics and uses a database comprising 3,483 variables, complete with question attributes and quality evaluations. Notably, Paper 3 alone assembled a database of 7,893 web tracking measures, including their respective characteristics, as well as estimates of validity and reliability. Therefore, it becomes evident that as methodological research increasingly adopts a multiverse approach to web tracking data, the volume of available evidence will swiftly surpass what resource-intensive traditional data sources like surveys can offer. With this abundance of data, it should be feasible to promptly identify patterns in the characteristics of web tracking measures that hold relevance, thereby aiding future research in predicting data quality even before collecting it.

Optimistically, most of the errors and problems identified seem to be fixable. Contrary to surveys, which will always have to deal with the limits of human cognition, there is no reason to believe that most of the errors of web tracking data are unsolvable. By identifying these challenges, companies and researchers can start working on them. In the next section, I present an optimistic agenda of improvements for the next few years.

An agenda for an improved use of web tracking data

This thesis, as a methodological endeavour, aims to empower both researchers and practitioners to enhance their utilization of web tracking data. The findings presented here illuminate significant areas for improvement, with each paper offering evidence that can be applied by both data providers and users.

The insights within this thesis imply that online fieldwork companies have substantial room for advancement in their practices. While much of web tracking research has been facilitated by the availability of metered panels from companies like Netquest and YouGov, it is imperative that we collectively encourage them to elevate their standards. A key aspect demanding enhancement is ensuring that participants are comprehensively tracked. As elucidated in Paper 2, the pervasive issue of tracking undercoverage significantly taints the statistical estimates derived, introducing substantial biases. Companies should intensify their efforts to address this issue. This may entail refining their array of tracking technologies or providing stronger incentives, potentially linked to panellists' coverage. Failure to address this issue might render it challenging for academic researchers to justify allocating public funding for the acquisition of data that is evidently biased.

Furthermore, these companies must prioritize transparency in their practices. While the confidentiality surrounding the tracking technologies they employ, and the constraints they face, might align with their business interests, it contradicts the principles of open science. Researchers require transparent information regarding the tools utilized in data production to establish trust in the data. Additionally, it should be required for online fieldwork companies to collect information about the prevalence of errors such as tracking undercoverage within their panels, and to make this information accessible to researchers purchasing their data. Additionally, this information should be available to enable researchers to tailor their sampling approaches more effectively. Moreover, companies must diligently work on improving tracking practices for Apple users. As Paper 2 reveals, iOS and MAC devices are disproportionately undercovered. While it is acknowledged that tracking iOS devices poses greater challenges than other devices, heightened efforts should be directed towards minimizing the systematic loss of data from these widely adopted devices.

Funders and major research institutions must reevaluate their practices in light of the challenges posed by web tracking data. Presently, many of these issues lie beyond the control of researchers. When utilizing data already collected by private companies, researchers cede control over how the data is generated and its source. This has far-reaching implications for data representativeness and measurement quality in social research. The most effective way for researchers to implement the recommendations from this thesis is to gain control over the entire data collection process. This encompasses everything from the development of tracking technologies to participant recruitment and convincing them to install tracking tools across all their devices, to data processing. However, given the complexity and substantial cost associated with this approach for individual web tracking studies, resources should be allocated to establish publicly funded, open web tracking infrastructures.

These infrastructures could take various forms, such as probability-based or high-quality nonprobability metered panels constructed and maintained by public research institutions. Alternatively, they might manifest as a versatile toolkit available to researchers for a nominal fee or through open-call competitions, offering open-access tracking technologies with easily configurable servers. Cumulatively, the resources directed towards private companies for data acquisition could be redirected to initiate these infrastructures. Nonetheless, their success hinges on adhering to established best practices, including those outlined in this thesis. If data from private companies is still perceived as of higher quality, it may hinder the full transition from private to public data.

Researchers and practitioners using web tracking data should also reconsider their practices. Firstly, there is considerable room for improvement in the reporting practices of already published papers. If researchers opt to employ a single measurement instrument (or a few), they should be explicitly defined, similar to the way survey questions are typically reported in papers. This involves disclosing how concepts were translated into measures, including details such as the list of URLs, criteria for their selection, which visits to those URLs were considered, and the duration of participant tracking. Additionally, journals should require researchers to transparently disclose the mix of tracking technologies employed, document them, and elucidate the limitations associated with each. If this data is not made available by the private companies used, it raises concerns about the suitability of this data for scientific research.

Secondly, researchers should implement strategies to identify and, if possible, mitigate some of the error sources identified in the TEM. For example, they can include a battery of questions in their questionnaires to ascertain the devices individuals use for online activities and who uses those devices. This aids in identifying the extent to which tracking undercoverage and shared devices may affect the data. All web tracking studies should include this information to allow for the assessment of data quality.

Thirdly, researchers should embrace uncertainty. Unlike traditional social research, where the use of a single measure is commonplace, digital trace data research does not necessitate such limitations. The design of these measures is often arbitrary, given the lack of substantial evidence to determine the best approach. Rather than concealing this uncertainty, researchers using web tracking data should acknowledge it and incorporate it into their analyses. In psychology, multiverse analysis has been defended as a valuable and legitimate approach to address the replication crisis in the social sciences. This thesis demonstrates that adopting a multiverse of measurements approach can inject much-needed transparency into web tracking research.

Lastly, one of the key takeaways from this thesis is that partial observations of online behaviours can result in significantly biased measures. Given that most research to date has solely tracked individuals on desktop devices, and contemporary projects still adhere to this practice, it is essential to encourage researchers to reconsider this approach. Desktop-only approaches provide insights only into what individuals do on their desktops. Extrapolating beyond this, particularly considering that half of online activities occur on mobile devices, can lead to erroneous conclusions. While tracking individuals on mobile phones presents greater complexity and raises concerns about data quality, it is insufficient reason to entirely disregard mobile browsing. This is particularly relevant for phenomena primarily occurring on mobile devices, such as those associated with social media usage.

Limitations

This thesis clearly contributes to an enhanced utilization of web tracking data in the realm of social sciences, yet it does possess some inherent limitations. While each individual study within this thesis underscores specific limitations, there are overarching constraints that warrant consideration.

First and foremost, it is crucial to recognize that this thesis does not endeavour to present a comprehensive and exhaustive understanding of the quality of web tracking data. As elucidated in Paper 1, there exists a multitude of errors that theoretically could affect web tracking data. This thesis, however, narrows its focus to those errors which, based on my experience with this data, appeared to demand more immediate scrutiny. A comprehensive comprehension of the quality of web tracking data will only emerge through cumulative research that exhaustively explores all potential sources of error.

Secondly, while Paper 3 delves into the validity and reliability of media exposure measures computed with web tracking data, it is essential to acknowledge that alternative approaches exist for exploring this dimension. For example, this thesis only concentrates on predictive validity, in line with the traditional research practices in political communication, neglecting other dimensions of validity traditionally explored in psychometrics. Moreover, the simplex model utilized enables the estimation of random measurement errors but falls short of addressing systematic ones. Consequently, even though the thesis asserts that web tracking measures are generally highly reliable, it does not rule out the possibility that measures are severely biased. To attain a comprehensive understanding of the measurement properties of web tracking data, it is imperative to supplement the findings in this thesis with other approaches to measuring validity and reliability. The application MTMM models, as previously mentioned, could offer insights into the systematic errors of web tracking data measures and their measurement validity.

Thirdly, this thesis predominantly centres on the use of web tracking data to replace survey self-reports. Consequently, it emphasizes the creation of simple measures that gauge behaviours at the URL or domain level, aligning with the prevailing research approach. However, it is essential to recognize that web tracking data holds the potential for more innovative applications. For instance, it can be leveraged to extract information from the content participants visit, subsequently generating variables for analysis. This includes quantifying the time individuals are exposed to news content, irrespective of the URL source. Furthermore, web tracking data can be employed to measure people's attitudes, opinions, or their interactions with webpage content. It can even serve purposes beyond statistical inferences, such as identifying individuals to invite to real-time surveys. While many of the errors identified in Paper 1 remain relevant to these applications, and the issues scrutinized in Papers 2 and 3 will inevitably impact any form of web tracking data application, the conclusions drawn from this thesis can be directly applied primarily to studies utilizing web tracking data in a manner akin to the approach taken here.

Lastly, this thesis predominantly centres on the use of web tracking data for measuring media exposure and related constructs. This focus is unsurprising, given that a majority of prior studies employing web tracking data have adopted a similar approach. However, it is pertinent to highlight that measuring other constructs might present varying degrees of complexity and challenge. For instance, assessing online behaviours like TV and movie streaming, which often transpire on digital devices typically not tracked by web trackers (e.g., smart TVs) or occur on devices from non-tracked individuals (e.g., partners), might entail significantly greater complexity. Similarly, behaviours that unfold in the background, such as listening to music or podcasts, rather than on active browser tabs, may also pose distinct challenges. Consequently, further research is imperative to ascertain the magnitude of errors associated with web tracking data for other constructs, enabling a comprehensive understanding of when web tracking data is suitable or unsuitable for specific research purposes.

Conclusions

This thesis collectively challenges the prevailing notion that web tracking data stands as the gold standard for measuring online behaviours. To identify and quantify the errors inherent in web tracking data, this thesis draws upon traditional survey methodology and psychometrics theories and methods. These established methods and theories are adapted to the realm of digital trace data through the use of approaches and frameworks commonly employed in the field of computational social sciences. While the promises of digital data and computational methods hold immense potential for advancing our understanding of society, it is imperative to acknowledge the persistent issues surrounding measurement and representativeness in web tracking research.

Throughout this thesis, I have provided both theoretical and empirical evidence of the errors and biases embedded within web tracking data. I have also illuminated the complexities and uncertainties surrounding the use of this data for statistical inferences. However, I have demonstrated that these challenges can be effectively addressed by applying a mature measurement theory tailored specifically for web tracking data.

Since the inception of this thesis, there has been a notable shift in the perception of web tracking data. Many researchers are now coming to terms with the limitations of this emerging data source. This shift has led to the production of empirical examinations of these errors, the formulation of best practices, and the development of guidelines to navigate this challenging terrain. Moreover, initiatives are underway to establish publicly funded web tracking infrastructures, with the work of organizations like GESIS serving as a commendable example.

Researchers have also started to exhibit greater creativity in their utilization of web tracking data. Unlike in the past, where web tracking data was predominantly employed as a substitute for survey data, researchers are beginning to recognize that web tracking data possesses unique and valuable characteristics that can enhance, rather than replace, surveys. For instance, web tracking data allows for the measurement of the content individuals are exposed to. Leveraging advanced computational tools, researchers can extract more nuanced insights from this data than from conventional aggregated lists of URLs.

In sum, these developments signal that the field is finally maturing. Collectively, we can progress toward a new era characterized by open and publicly accessible web tracking data, healthy scepticism and methodological rigor, and marked by collaboration rather than competition between traditional and novel data sources.

It is important to note that no data source is flawless. Surveys have provided society with unprecedented insights into itself. Meanwhile, digital trace data has the potential to enhance our understanding of the human experience, especially within the digital domain, in ways that may have otherwise been unattainable through surveys alone. Realizing this potential hinges on our ability to combine both new and traditional data sources, viewing digital trace data through the lens of survey science, and leveraging digital data and computational methods to enhance the use and analysis of surveys. With a commitment to realism and rigor, we can usher the social sciences into a new measurement revolution.

Bibliography

Gerring, John. 2016. *Social Science Methodology A Unified Framework*. Cambridge University Press.

Hand, D. J. 2004. *Measurement theory and practice: The world through quantification*. Wiley.

Jungherr, Andreas. 2019. “Normalizing Digital Trace Data.” In *Digital Discussions*.

Oberski, D. L., A. Kirchner, S. Eckman, and F. Kreuter. 2017. “Evaluating the Quality of Survey and Administrative Data with Generalized Multitrait-Multimethod Models.” *Journal of the American Statistical Association* .

Watts, Duncan J. 2012. *Everything is obvious: Once you know the answer*. Crown Business.

Supplementary Material

Supplementary Material: Paper 1

SM1: Tracking solutions offered by Wakoopa, and their limitations

Table 1 shows the different technologies that the company Wakoopa (currently the main provider of tracking solutions, and the one used by Netquest) uses, as well as the type of information that these collect and for which devices they are used.

Table 6.1: Data Collectable by Tracking Technology and Target for Wakoopa

Data Category	**Data Type**	**PC App**	**PC Plug-ins**	**Android SDK**	**iOS Proxy**	**Chrome**	**Firefox**
Online Tracking	URLs - HTTP Traffic	Yes	Yes	Yes	Yes	Yes	Yes
	URLs - HTTPS Traffic	No	Yes	Yes	Yes	Yes	No
	Incognito Sessions	No	Yes	Yes	Yes	Yes	No
	HTML	No	Yes	Yes	Yes	No	No
	Time Stamps	Yes	Yes	Yes	Yes	Yes	Yes
Apps	App Name	-	-	-	-	Yes	Yes
	App Usage Start Time	-	-	-	-	Yes	Yes
	App Usage Duration	-	-	-	-	Yes	Estimated
	Offline Apps	-	-	-	-	Yes	No
Search Terms	Search Terms	Yes	Yes	Yes	Yes	Yes	No
Device Information	Device Type (e.g., desktop)	Yes	Yes	Yes	Yes	Yes	Yes
	Device Brand (e.g., Xiaomi)	-	-	-	Yes	Yes	-
	Device Model (e.g., S9)	-	-	-	-	Yes	Yes
	Operating System (e.g., iOS)	Yes	Yes	Yes	Yes	-	Yes
	OS Version (e.g., 10.1.2)	-	-	-	-	Yes	Yes
	Internet Provider (e.g., Voxi)	-	-	-	-	Yes	Yes

SM2: A discussion of why we could not minimize or quantify some error sources for TRI-POL

This SM briefly discusses why, in the TRI-POL project, we could not design strategies to minimize or quantify some of the error sources affecting metered data not mentioned in the main manuscript's text.

Noncontacts and Non-consent: This error source cannot be measured nor prevented for samples obtained through metered online panels since the invitation to install a meter is outside researchers' control.

Social desirability: Currently, the best approach to assess the potential effect of social desirability on people's behaviour is to use or combine tracking approaches that allow to collect traces from before participants installed the tracking solutions and compare these with post-installation traces. This should allow assessing whether installing the meter changed their behaviours. Unfortunately, this cannot be done with a metered panel, since sampled participants installed their technologies before being sampled.

Shared devices: Several approaches can potentially reduce the effects of shared devices and/or allow to assess the associated errors. First, depending on whether a fresh sample is used or not, participants can be asked to only install tracking solutions on non-shared devices, or we can only sample participants who use non-shared devices. This, nonetheless, could potentially introduce tracking undercoverage and/or selection errors. A more appropriate strategy would be to use algorithms to differentiate between the participant's behaviour and third-person's behaviours . This approach helps estimate the bias introduced by shared devices and select only participant's traces to create the desired variables. Nonetheless, we could not apply this strategy since these classification algorithms are trained using the raw data of participants, which we did not have access to. Hence, we expect TRI-POL data to be affected by shared-device-related measurement errors but do not know to what extent.

Data aggregation: Although we did not have access to the raw dataset, we were able to use all the tracking information to compute our desired variables. Hence, we do not expect data aggregation errors.

Data anonymization: Since we did not have access to the raw data, but to a structured dataset composed of pre-defined variables, there was no need to anonymize pieces of information from the raw dataset. Hence, we do not expect data anonymization errors.

Non-trackable individuals: Opt-in online panels are affected by coverage errors, but they are unmeasurable per se. The problem of non-trackable individuals might make this problem worse. Given that we used a metered opt-in online panel, the prevalence and added potential bias of non-tracked individuals was not measurable.

SM3: Information about the TRI-POL data

For the empirical analyses, we used web survey answers, metered data and paradata from the first wave of the TRI-POL panel dataset. Specifically, we focus on the data collected in Italy, Portugal and Spain. Data was collected through the metered panels of Netquest. The Netquest metered panels provide a pool of individuals with the meter installed, who can also be contacted to conduct surveys and, hence, link their online behaviour with their survey answers. When the panellists agree to join the metered panel, they must install the meter on at least one device (PC, tablet, or smartphone) and start sending information (passively) to Netquest to become part of the metered panel.

Cross quotas for age, gender, educational level and region were used to guarantee that the samples were similar on these variables to the general Internet population between 18 and 70 years, for each country. Data collection for the three waves took place between September 2021 and April 2022. In total, for the first wave, 3,548 respondents completed the survey until the end, 1,289 in Spain, 1,028 in Portugal and 1,231 in Italy. Challenges were faced when filling some of the specific cross-quotas with participants from the metered panel. This required supplementing in some cases with non-metered panellists. For the first wave, 993, 818 and 842 participants in Spain, Portugal and Italy respectively had the meter installed in at least one mobile or PC device. We will focus on this subsample of metered individuals.

Metered data was collected for the 15 days prior and posterior of participants starting each survey wave. The meter captured each URL (or app for mobile devices) accessed by the panellists, with timestamps for when the panellists first visited the URL, and the number of seconds in which the URL remained active in the browser. Participants were tracked on iOS and Android mobile devices, and Windows and MAC computers. Besides, we collected paradata from participants. For all individuals we were able to identify the technology with which they were being tracked, the type of device, the OS, whether it was a tablet or smartphone and, for plug-ins, the browser in which they were installed. Finally, the questionnaire focused on measuring, among others, political trust, participation and polarization, as well as several sociodemographic variables

SM4: Quantifying the validity of metered data measures

As an example, SM 4 briefly discusses the approach followed to quantify the validity of “online news media exposure.”

Measuring online news media exposure

Our first step was defining how to create the measurement of “online news media exposure.” The goal was to get a measure of participants average exposure during the project’s period of interest. Hence, we considered all the potential design choices that we could make. We identified eight design questions, with several choices to be made within them. Table 2 summarises these.

Table 6.2: Design Characteristics and Choices for the Concept ”Online News Media Exposure”

Questions	**Choices**
What list of news media domains to use?	Own list, Alexa, Tranco, Cisco, Majestic
How many news media domains to use?	All domains, top 200, 100, 50, 20, 10 most visited
What information to use within those domains?	All URLs, only those identified as political
What do we consider as being exposed to a URL?	Visits equal or longer than 1 second, 30 seconds, or 120 seconds
What is the level of interest?	Number of visits, number of minutes
Should we use information from all devices?	Mobile and PC, only PC, only mobile
How many days of tracking should we use?	2, 5, 10, 15, 31
Should we use information from before or after the survey?	Before, after, both

Since it was unclear what design choices would be the ones yielding the most valid measures, we decided to apply our proposed approach to measure the validity of metered data measures (presented in section 6.1.1 of the paper). Hence, we created a variable for each potential combination, which resulted in 3,573 variables to measure the concept of online news media exposure.

Analysing convergent validity

We first explored the convergent validity of the 3,573 variables computed. Convergent validity describes the fit between independent measures of the same underlying concept. Hence, if different variables were measuring the same underlying concept, they should highly correlate with each other. To explore whether this is the case or not, we computed one correlation for each potential pair of variables. Therefore, we obtained 6,349,266 unique Pearson’s correlation coefficients for each country.

Analysing predictive validity

Next, we explored the predictive validity of the different computed variables. Predictive validity refers to the degree to which a measurement instrument is related to a gold standard measurement. Measures closer to the theorised true relationship should be preferred. Often, when the true relationship is unknown, it is assumed that the higher the predictive power, the better. Although we do not make this assumption, any fluctuation in the predictive power of the variables would indicate differences in terms of predictive validity. Specifically, we use political knowledge as gold measure, since it is the accepted practice in the media effects literature. To measure political knowledge, we use an additive political knowledge index that ranges from 0 to 4. Hence, for each of the 3,573 variables we ran an OLS regression model, with political knowledge as the dependant variable, the media exposure variable as the main independent variable, and some common control variables (age, gender, and educational level). Consequently, we obtained 3,573 partial regression coefficients for media exposure, in each country.

Analysing the impact of each design choice on predictive validity

Next, to understand the extent to which these choices affect the validity of measurements, drawing inspiration from the Survey Quality Predictor (SQP), we created a new dataset, in which the 3,573 variables were used as the observations, their associated partial regression coefficients as the dependant variables and the characteristics of the variables as the predictors. To predict the impact of each design choice, we used random forests of regression trees (R package randomForest) to extract the following information:

- the variable importance, measured as the percentage increase of Mean Squared Error (MSE) of the model if a specific variable had not been included in the trees used.
- the marginal effect of each design choices, understood as the adjusted predictions when holding all predictors constant.

SM5: Quantifying the prevalence and bias of tracking undercoverage

This SM discusses the approach followed to compute the prevalence of tracking undercoverage and simulate the potential bias it can introduce, focusing on the "average time spent on the Internet."

Analysing the prevalence of tracking undercoverage

The prevalence of tracking undercoverage was estimated by combining survey questions and paradata Netquest collects from their metered panellists about the technology with which they were tracked, the type of device, the OS, whether it was a tablet or smartphone and, for plug-ins, the browser in which they were installed.

In terms of survey questions, we measured the number of devices used to access the Internet by asking the following: "During the last 15 days, from how many of these different types of devices have you accessed the Internet (including using apps such as Facebook, Twitter or YouTube)? Please, type the number of devices in the respective boxes." The list of devices provided was designed to match the information available from the paradata. Specifically, we asked for: (1) Computer(s) with Windows OS; (2) Apple computer(s) (MAC); (3) Smartphone or tablet with Android OS; (4) Apple smartphone or tablet (iPhone or iPad); (5) Others.

Moreover, to assess the types of browsers used to access the Internet, we asked participants a maximum of three questions (depending on which devices were tracked, and the technology used according to the available paradata), as follows: "During the last 15 days, have you used any of the following web browsers to access the Internet through [a computer with Windows operating system/ an Apple computer (MAC)/ a smartphone or tablet with Android operating system]?" The list of browsers varied depending on the ones available in each OS.

Combining both sources of data, we created a variable indicating whether a participant had at least one device or browser not covered. (i.e., the number of devices and/or browsers tracked was lower than the self-reported one). Although we are mindful that self-reports might themselves be affected by measurement errors, we computed the proportion of individuals being undercovered, for each country in this way due to the lack of better information.

Simulating the bias introduced by tracking undercoverage

We used data from those participants being fully covered to run Monte Carlo simulations. Given that metered data was collected separately for computers and mobile devices, for fully covered participants it is possible to compute their estimates using all tracked targets, as well as only some of the tracked targets, simulating their estimates for specific undercoverage scenarios. Differences between the fully tracked behaviours and those affected by undercoverage can be considered as tracking undercoverage bias. Hence, in our simulations we modified the targets used to compute the estimates and simulated the effect on those estimates for different tracking undercoverage scenarios. Specifically, to run the Monte Carlo simulations, we developed the following steps:

1. We identified the participants fully covered and tracked in both mobile and PC devices.

2. We defined seven tracking undercoverage scenarios to simulate varying the targets being omitted when computing the estimates (PCs or mobiles), and the prevalence in the sample. Specifically, we tested the effect of having 25%, 50% and 75% of participants with no PC or mobile device tracked. In addition, we conducted the simulation of the actual undercoverage scenario in our samples, together for PC and mobile.

3. For each of the coverage scenarios, we randomly created 1,000 allocation scenarios, in which all participants had the same chance to be undercovered. For instance, for the scenario with 25% of participants without any mobile device tracked, an individual would have .25 probability of having all the mobile devices untracked, and a .75 probability of being fully covered.

4. For each specific variable, in our case the “average time spent on the Internet”, 1,000 estimates were created for all seven tracking undercoverage scenarios. Participants selected as undercovered got part of their tracking data removed when computing the estimates of interest. As an example, if Individuali was tracked on a PC and a mobile device and Individuali was randomly selected to be non-tracked for all their mobile devices, all data from these devices would be considered as 0 when computing the estimates of interest. For the sake of simplicity, all complete losses of information were set to zero.

5. All estimates were computed using inverse probability weights created with the random forest relative frequency method, to account for differences between the subsample of fully covered and the full sample of metered participants. Finally, for each tracking undercoverage scenario, the average of the 1,000 simulations was considered as the average undercovered estimates. The difference between the average undercovered estimate and the fully covered estimate was considered as the average bias estimate.

As illustration, we focus on the "average time spent on the Internet". First, we added the duration of the visits to all URLs and apps across all tracked devices and browsers to compute the total time spent on the Internet in a day. Then we computed the average time for the 15 days prior to the survey being answered.

SM 6: Quantifying the prevalence of technology limitations

Since we knew exactly what information could and could not be collected with each tracking technology (see SM 1), as well as the specific technologies used to track each participant, we could compute the proportion of participants affected by the following technology limitations: 1) unobserved behaviours in incognito mode, 2) impossibility of tracking subdomain information, and 3) in-app information.

We computed the prevalence of all these limitations separately, for each country. We considered a participant to be affected by any of these limitations if at least one of their targets was tracked with a technology suffering from the limitation of interest.

SM 7: Exploring whether participants tracked on iOS devices present different measurement properties

We designed an approach to identify differential measurement properties for participants tracked on iOS devices. First, we asked participants the following: "Approximately, how much time do you spend on a typical day on the Internet (including using apps such as Facebook, Twitter or YouTube)? Please, type the number of hours and minutes in the respective boxes."

We then combined this self-reported information with the observational data from the meter, to compute the absolute difference in minutes (Absolute difference: —Self-reported time – Tracked time—). Next, we ran a model exploring which variables were associated with this absolute difference. The main independent variable was "Tracked on iOS", which indicates whether someone was tracked on an iOS or not (0 = No, 1 = Yes). The model also included other variables to control for potential confounders: general undercoverage (1= undercovered, 0 = fully covered); a self-reported measure of Internet use (continuous, minutes spent on the Internet on a typical day); the number of months that a participant had been part of the Netquest panels (continuous), as a proxy for panel loyalty; whether the person self-reported using mobile devices to access the Internet (0 = No, 1 = Yes); finally, we introduced age (continuous), gender (women= 0, male= 1) and whether a participant had completed higher education (0 = no, 1= yes). Although we expect both the survey and metered measures to be affected by errors, a significant effect of "Tracked on iOS" could indicate that participants tracked on an iOS present different measurement properties.

SM 8: Quantifying the prevalence of participants with undercoverage-induced non-observations.

We asked participants whether they had visited Twitter, Facebook, and the top 10 most visited news media domains in each country (according to Tranco: https://tranco-list.eu/) with non-tracked targets. Specifically, the questions asked: "During the last 15 days, have you used another device or browser apart from [INSERT DEVICE(S)] to visit the following web pages or apps." For each individual, the devices inserted where those targets which we knew, thanks to the paradata, that participants were being tracked with. After that, the list of web pages / apps was presented, with a yes/no scale.

For each specific web pages and/or apps, participants were identified of having undercoverage-induced non-observations when they self-reported having visited those web pages and/or apps, but no behaviour had been tracked with the meters. Although we are mindful that the self-reports might also be affected by measurement errors, since we did not had access to other sources of information, after identifying those participants, we computed the proportion for each country.

Supplementary Material: Paper 2

SM1: Comparison between full sample and subsample tracked participants.

Table 3 presents an analysis comparing the full sample to the subsample of tracked participants, considering a range of demographic, political, and technological variables. To assess these differences, I employed t-tests for continuous variables (age, left-right, self-reported minutes online, and number of devices), as well as chi-square tests for categorical variables (sex, education, and interest in politics).

Importantly, the results displayed in Table 1 indicate that no statistically significant differences were detected between the two samples across the selected variables. Furthermore, it is noteworthy that these differences, even when present, are of negligible magnitude. Consequently, both the full sample and the subsample of tracked participants exhibit remarkable similarity in terms of the explored variables.

Table 6.3: Comparison Between Full Sample and Subsample Tracked Participants

Variables	**Full Sample**	**Tracked Participants**
% Female	53.1	51.7
Avg. age	45	45
% Tertiary education	39.5	40.4
Avg. left-right	4.8	4.8
% Interested in politics	11.6	12.3
Self-reported minutes online	195	197
Avg. number of devices	3.2	3.3

SM2: Questions and paradata used to identify tracking undercoverage

The prevalence of tracking undercoverage was estimated by combining survey questions and paradata Netquest collects from their metered panellists about the technology with which they were tracked, the type of device, the OS, whether it was a tablet or smartphone and, for plug-ins, the browser in which they were installed. Specifically, they collect:

- Number of tracked PCs with Windows OS.
- Number of tracked PCs with MAC OS.
- Number of tracked Android devices.
- Number of tracked iOS devices.
- Number of chrome browsers tracked on Windows devices.
- Number of chrome browsers tracked on MAC devices.
- Number of Firefox browsers tracked on Windows devices.
- Number of Firefox browsers tracked on MAC devices.
- Number of Safari browsers tracked on MAC devices.
- Whether a participant is tracked with a desktop app tracker (can track all browsers in device)
- OS version of Android device (will have an effect on what browsers can be tracked with technology used)

In terms of survey questions, we measured the number of devices used to access the Internet by asking the following: "During the last 15 days, from how many of these different types of devices have you accessed the Internet (including using apps such as Facebook, Twitter or YouTube)? Please, type the number of devices in the respective boxes." The list of devices provided was designed to match the information available from the paradata. Specifically, we asked for: (1) Computer(s) with Windows OS; (2) Apple computer(s) (MAC); (3) Smartphone or tablet with Android OS; (4) Apple smartphone or tablet (iPhone or iPad); (5) Others.

Moreover, to assess the types of browsers used to access the Internet, we asked participants a maximum of three questions (depending on which devices were tracked, and the technology used according to the available paradata), as follows: "During the last 15 days, have you used any of the following web browsers to access the Internet through [a computer with Windows operating system/ an Apple computer (MAC)/ a smartphone or tablet with Android operating system]?" The list of browsers varied depending on the ones available in each OS.

SM3: Comparison full sample and subsample of fully tracked participants.

Table 4, akin to SOM 1, presents a comparative analysis between the full sample and the subsample comprising participants who were comprehensively tracked on their devices, a subset used for the simulation analyses. This analysis encompasses the same array of demographic variables and employs identical statistical methodologies for assessment.

However, in contrast to the findings of SM 1, Table 4 reveals a noteworthy difference between the two samples. Specifically, individuals in the subsample of fully tracked participants tend to utilize a fewer number of devices for online activities on average. This distinction aligns with our analytical results, highlighting that device usage significantly influences the likelihood of being fully tracked.

It is imperative to note that apart from this variable, no other statistically significant differences were observed between the two samples. Furthermore, the magnitude of these differences, where present, remains minimal, underscoring the overall similarity between the full sample and the subsample concerning the examined variables.

Table 6.4: Comparison Between Full Sample and Subsample of Fully Tracked Participants

Variables	**Full Sample**	**Tracked Participants**
% Female	53.1	52.8
Avg. age	45	46
% Tertiary education	39.5	36.5
Avg. left-right	4.8	4.7
% Interested in politics	11.6	12.4
Self-reported minutes online	195	191
Avg. number of devices	3.2	1.8*

* Note: The value in this column represents an average with a different scale or unit than the corresponding value in the "Full Sample" column.

Supplementary Material: Paper 3

SM1: Comparison between full sample and subsample tracked participants.

As for SM 1, Paper 2, Table 3 presents a comparison between the full sample and the subsample of tracked participants across various demographic, political, and technological variables. This analysis employed t-tests for continuous variables (such as age, left-right political orientation, and self-reported minutes spent online) and chi-square tests for categorical variables (including sex, education level, and interest in politics). It is noteworthy that not only were the observed differences not statistically significant, but Table 5 also underscores the negligible nature of these differences. In essence, both samples exhibit an almost identical profile in terms of the examined variables.

SM2: Comparison between Quasi-Markov Simplex Models

This SM briefly discusses the differences found between the different Quasi-Markov Simplex Models tested to estimate the reliability of the measures. The tested models are the following:

1. **Basiline model:** the one described in the equations 4 and 5.
2. **2-LAG:** the one presented in equations 6 and 7.
3. **Equal means:** presented in equations 8 and 9.
4. **Correlations between errors:** This model adds four lag-1 correlations between random errors to the baseline model. They are freely estimated.
5. **Unequal variances in time:** This model relaxes the assumption of equal variances in time by constraining the variance of the measurement errors to be equal only at waves one and six. The other measurement error variances are freely estimated.

These models were run for the 2,631 measures, in all three countries. Table 5 shows the proportion of improper models that these approaches yielded, and their average reliability. Results are presented for the three countries separately.

Table 6.5: Performance Comparisons Between Models

Models	**% Improper**			**Reliability**		
	Spain	Italy	Portugal	Spain	Italy	Portugal
Baseline	75	73	72	0.77	0.89	0.76
2-LAG	35	13	18	0.83	0.84	0.86
Equal means	34	12	20	0.83	0.92	0.86
Correlated errors	77	67	63	0.78	0.89	0.84
Unequal variances	100	100	100	NA	NA	NA

Note: ”NA” indicates not applicable.

As Table 5 shows, both the LAG-2 and Equal means models perform significantly better than the other models, with substantially lower proportions of models being improper. The model with unequal variances performs so badly that no reliability coefficient could be obtained from any of the measurements. Although not always, in general these two models yield higher average reliability coefficients. Nonetheless, these differences are not very large.

SM3: Political knowledge questions

In this SM I present the political knowledge questions asked to build the index of political knowledge used to compute the predictive validity of the explored measures. Every participant was asked four questions about their country’s political system, and their current cabinet. These questions asked them whether a specific sentence was true, or not true. They had a limit of 30 seconds to answer the question, to avoid participants looking for the answers online. After 30 seconds, they were moved to the next question. The order of the questions was randomized for each participant. The general request for an answer was the following: “Now you will read some statements about politics in [Insert country]. These questions are not a personal ”test”, it’s just a matter of finding out how much knowledge people have about certain topics that are considered somewhat complicated. For each one, could you please indicate if you think it is true or false? If you don’t know, just select ”I don’t know” and move on to the next one.”

The specific questions for each of the countries are the following:

Questions in Spain

- The Spanish Congress has 525 deputies: True / False / Don't Know
- A person must be 25 years of age or older to stand as a candidate in the Spanish general: True / False / Don't Know
- Salvador Illa is a member of the Spanish Government: True / False / Don't Know
- The current government is a coalition government formed by the PSOE, Unidas Podemos, and ERC: True / False / Don't Know

Questions in Portugal

- The Portuguese Parliament has 175 deputies: True / False / Don't Know
- A person must be 25 years of age or older to stand as a candidate in the Portuguese general election: True / False / Don't Know
- Pedro Marques is a member of the Portuguese Government: True / False / Don't Know
- The current government is a minority government formed by the PS: True / False / Don't Know

Questions in Italy

- The Chamber of Deputies currently has 630 members: True / False / Don't Know
- A person must be 35 years or older to stand as a candidate in the Italian Senate: True / False / Don't Know
- Vincenzo Spadafora is a minister in the Italian government: True / False / Don't Know
- The current Italian government is supported in Parliament by Fratelli d'Italia, Lega, Partito Democratico and Movimento 5 Stelle: True / False / Don't Know

SM4: Control variables for the predictive validity models

In accordance with the main text's description, the cross-sectional OLS regression model used to evaluate the predictive validity of the media exposure variables included a set of common control variables. This section provides a brief overview of the phrasing employed to gauge these variables and outlines their respective measurement scales.

1. **Sex and Age:** The first and second control variables encompassed participants' self-reported sex and age. The sex question featured two response options: "male" and "female." Meanwhile, the age query required participants to input their age as numerical values.

2. **Education Level:** The third control variable pertained to participants' self-reported education. Respondents were presented with the following question: "What is the highest level of education you have completed?" Numerous country-specific response options were made available in each participating country. For example, Spain had 27 education options. For the complete list of categories, please refer to the translated questionnaires, accessible here: Link to Questionnaires. To simplify this variable, I recategorized it into a binary variable representing tertiary education status, with the following criteria for each country: Spain (>20), Italy (>7), Portugal (>10).

3. **Left-Right Ideology:** The fourth control variable involved participants' self-reported left-right political ideology. It was assessed using the widely utilized question: "When discussing politics, people often refer to the terms 'left' and 'right.' Can you please indicate where you position yourself on a scale from 0 to 10, where 0 signifies 'left' and 10 signifies 'right'?" This scale was a partially labeled 10-point scale, featuring "left" at 0 and "right" at 10. Participants also had the option to select "I prefer not to answer." The scale was displayed in a vertical orientation across all devices.

4. **Interest in Politics:** Finally, the model included a variable measuring participants' interest in politics. This was assessed through the following question: "To begin with, to what extent are you interested in politics?" The scale, presented in an inverted format, featured four response options: Not at all, A little, A fair amount, A lot

SM5: Predictive validifty for self-reports

To compare the predictive validity results obtained using web tracking data, I conducted parallel analyses by substituting the web tracking measures with a self-reported measure of media exposure. In this alternative approach, the main independent variable was derived from the following question: "Could you please indicate how often you keep yourself informed about current political issues, news, or opinions through online newspapers?" Respondents could select from the following scale of responses: never, less than once a month, once a month, several times a month, once a week, several times a week, every day, and several times a day.

Using this self-reported question, I performed both the fixed-effects regression model and the cross-sectional OLS model, which are presented in equations 10 and 11. Across the different countries considered, the standardized regression coefficients are as follows, presented in descending order: 0.12 (Spain-OLS), 0.11 (Portugal-OLS), 0.07 (Italy-OLS), 0.05 (Portugal-Fixed), 0.03 (Spain-Fixed), and -0.03 (Italy-Fixed).

SM6: Predicted coefficients versus observed ones

This SM includes two graphs that display scatter plots of predicted and observed values for both reliability and validity. These figures also incorporate multiple lines, illustrating the OLS regression line, the lowess smoother line, and the 45-degree line of unbiasedness. Lowess, an acronym for "Locally Weighted Scatterplot Smoothing," represents a non-parametric regression technique that fits a smooth curve to the data points. It offers a way to visualize data trends without making explicit assumptions about the functional form, such as assuming a linear relationship. In this context, the red dotted line represents a smoothed curve that approximates the relationship between the predicted and observed values. The 45-degree line serves as a reference line, denoting a perfect one-to-one correspondence between predicted and observed values. In an ideal scenario where predicted values perfectly align with observed values, all data points would fall precisely on this line. Departures from this line indicate the extent of bias or inaccuracy present in the predictions.

As detailed in the main text, the correlation between the squared correlations of predicted and observed coefficients ranged from 0.97 to 0.98 for the reliability coefficients and 0.90 to 0.91 for the validity coefficients. The figures visually demonstrate that, on the whole, the predictions perform admirably. However, two potential observations regarding model biases are worth noting. First, in the case of the 2-LAG reliability model (Figure 1), it appears to generate inflated predictions for questions with low reliability, evident from the deviation of the blue and red lines from the red 45-degree line. Second, the fixed-effects validity model (Figure 3) consistently inflates lower validity coefficients while deflating higher ones. Nevertheless, it is important to emphasize that, overall, the models provide more than satisfactory predictions of the observed values.

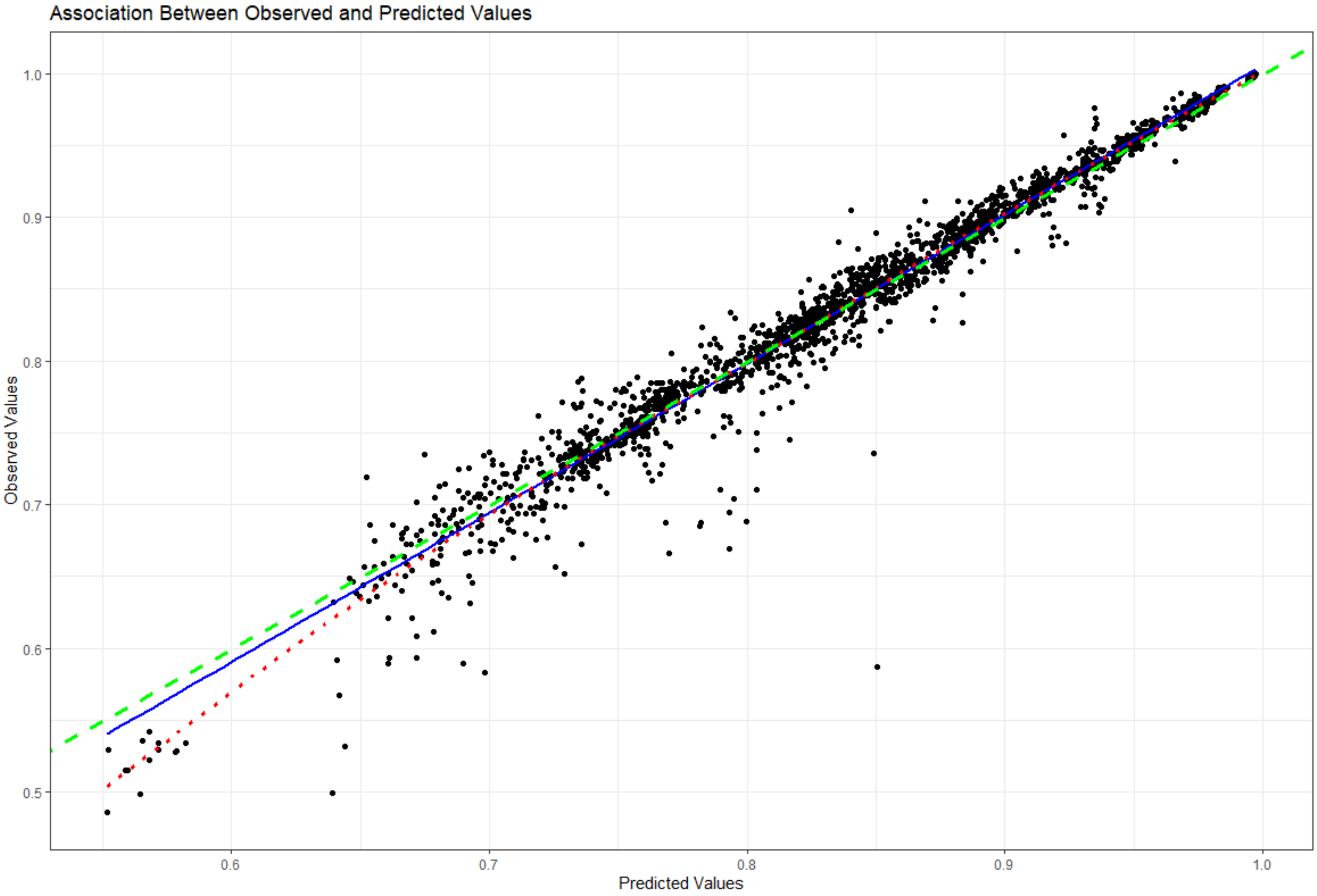

Figure 6.1: Predicted reliability versus observed reliability, 2-LAG model. Blue line represents an OLS regression line, the red line a lowess smoother line, and the green one the 45-degree line of unbiasedness.

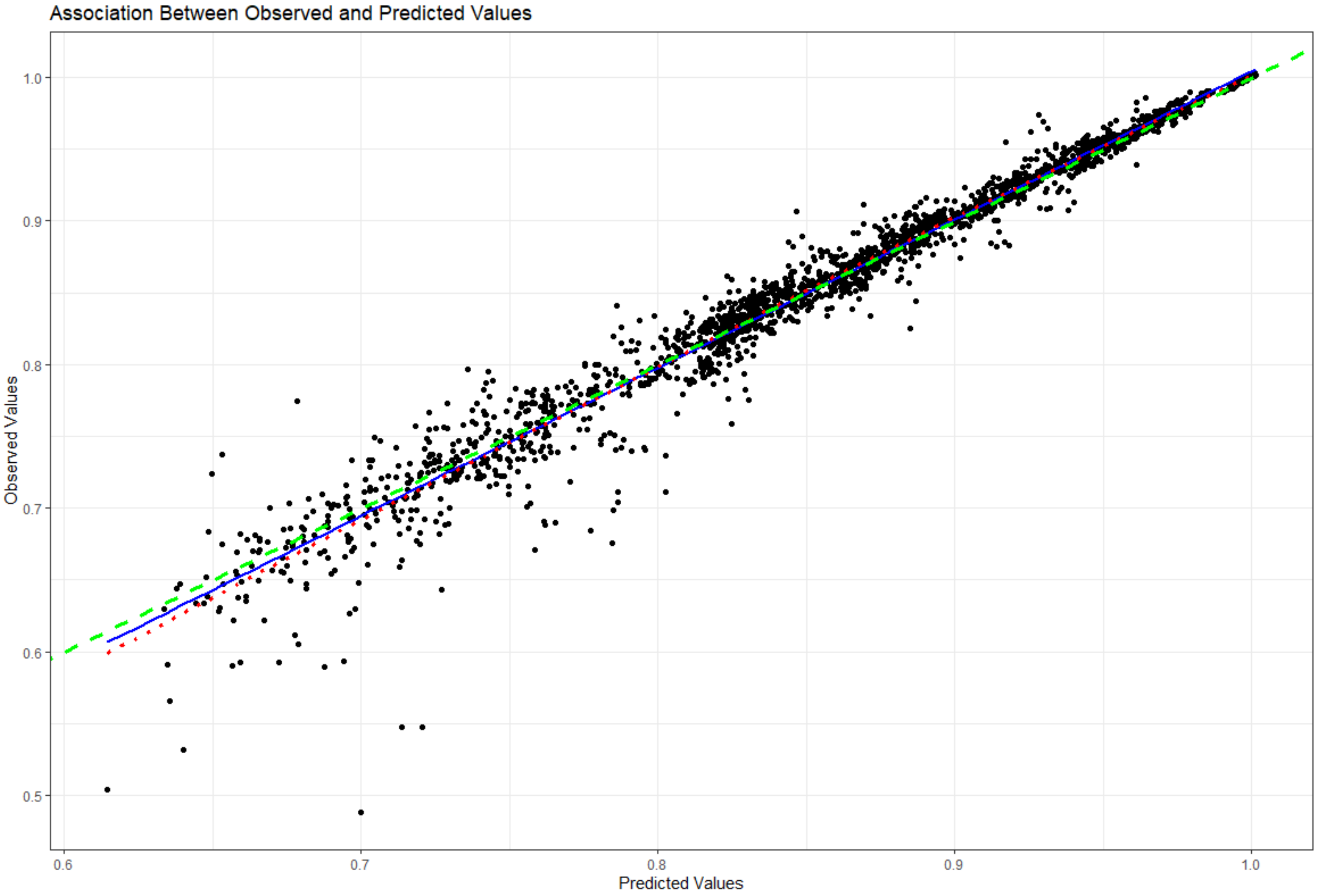

Figure 6.2: Predicted reliability versus observed reliability, equal means model. Blue line represents an OLS regression line, the red line a lowess smoother line, and the green one the 45-degree line of unbiasedness.

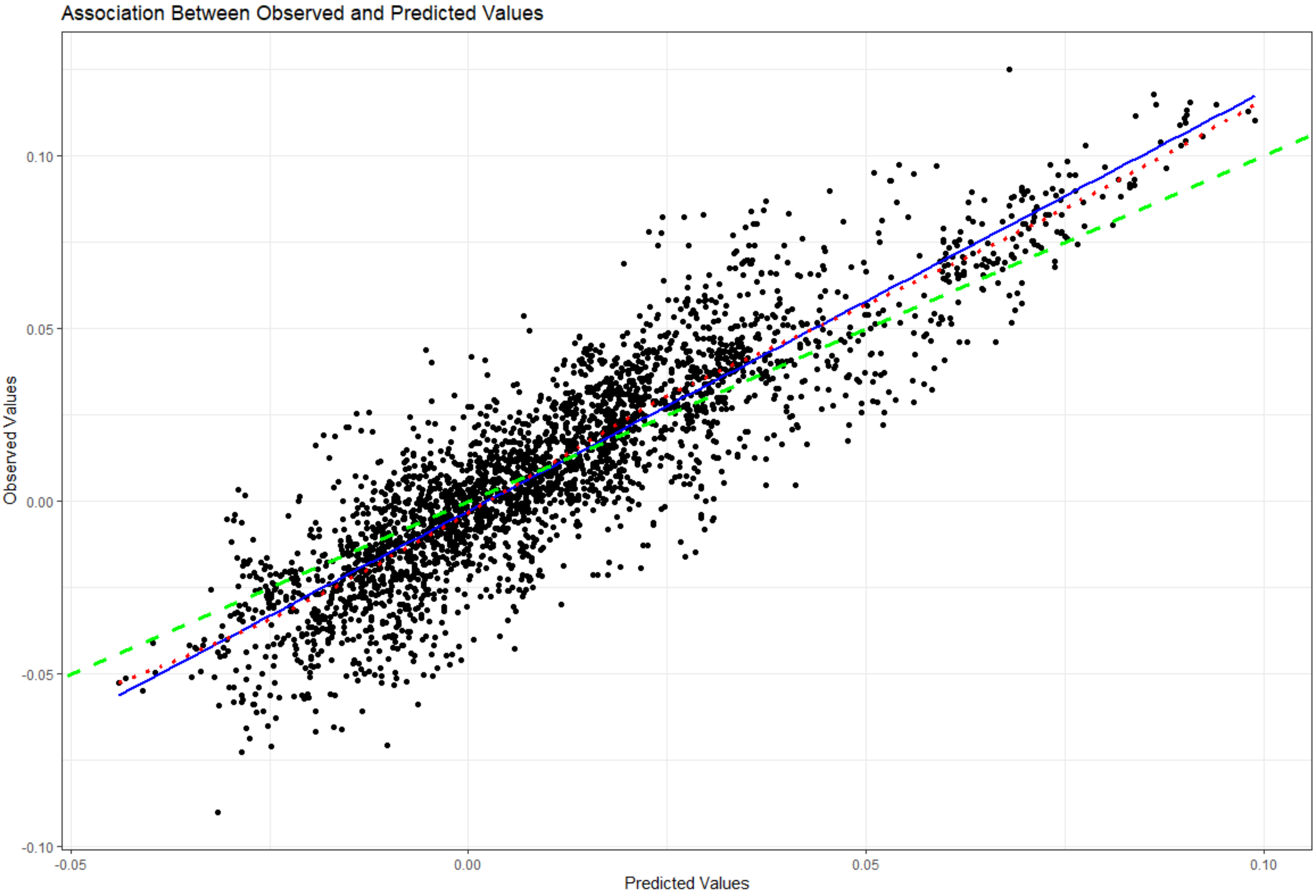

Figure 6.3: Predicted validity versus observed validity, fixed-effects model. Blue line represents an OLS regression line, the red line a lowess smoother line, and the green one the 45-degree line of unbiasedness.

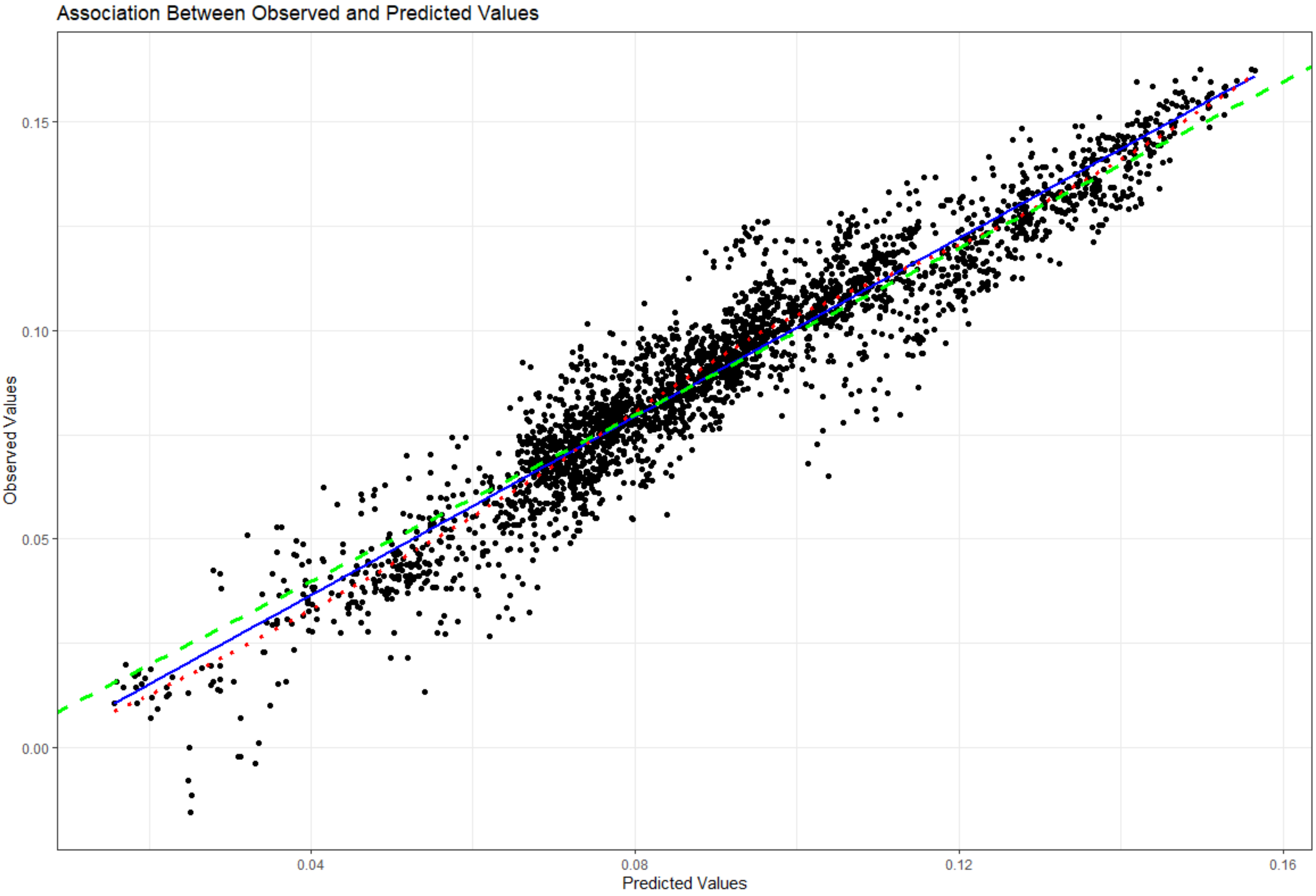

Figure 6.4: Predicted validity versus observed validity, cross-sectional model. Blue line represents an OLS regression line, the red line a lowess smoother line, and the green one the 45-degree line of unbiasedness.

SM6: Variable importance

The random forest procedure also provides what are commonly referred to as ”variable importance” measures. These measures gauge the marginal deterioration in mean square prediction error when a specific variable is omitted. To determine a variable’s importance, the algorithm randomly permutes the observed values of that variable and then recalculates the out-of-bag mean square error of predictions. If this reduction in mean square error is substantial, the variable is considered to have high importance. To evaluate the significance of each design choice, I calculated Breiman’s variable importance. However, it’s important to note that this approach has limitations, primarily related to its inability to account for correlations among different predictive variables. To address this concern, I also calculated conditional (unbiased) importance, using the r package ”cforests.” The results obtained from both approaches align, and thus, I present the traditional variable importance here.

The following table displays the variable importance scores for the four random forest models examined in this study. Focusing on the outcomes for the reliability models, for the 2-LAG model, the design choices with the highest importance in the prediction model include the length of the tracking period, the metric used, the country, and whether metrics are computed based on all URLs or only those related to ”hard” news. Similar results are observed for the equal means model. Concerning the validity models, for the fixed-effects model, the most influential design choices are the country of analysis, the metric employed, the length of the tracking period, and the number of media outlets used. These findings are consistent with those of the cross-sectional model, with the primary difference being the emphasis on the number of outlets versus the type of URLs categorized as news.

Table 6.6: Percentage Increase of RMSE when Variable is Excluded from the Model

	2-LAG	**Equal means**	**Fixed-effects**	**Cross-sectional**
Metric	0.006	0.005	0.0009	0.0004
List of media	0.00	0.00	0.00	0.00008
Top media	0.0003	0.0003	0.0005	0.0001
Information	0.001	0.002	0.0003	0.0002
Exposure	0.0005	0.0007	0.0004	0.0001
App behaviour	0.00003	0.00004	0.00003	0.000003
Tracking period	0.008	0.007	0.0005	0.0002
Country	0.002	0.005	0.0012	0.0009

Funding

The completion of this thesis has been possible thanks to funding from: Fundación BBVA; H2020 European Research Council, Grant/Award Number: 849165; Ministerio de Ciencia e Innovación, Grant/Award Number: PID2019-106867RB- I00/AEI/10.13039/501100011033.