Utrecht University, 2023

Faculty of Social and Behavioural Sciences

Department of Methodology and Statistics

Master Programme

Methodology and Statistics for the Behavioural, Biomedical and Social Sciences

Graduation Thesis

Applications of Survey Data in Digital Trace Data Total Error Reduction

Student:

Ilya Fominykh

Student Number 9495525

Supervisors:

Bella Struminskaya

Thijs Carrière

Ethics committee approval number: 23-1850

Contents

1 Introduction

1.1 Characteristics of Donated Digital Trace Data

The widespread use of digital devices and online platforms produces large volumes of data. This data, formed as a byproduct of life in modern society, is often referred to as digital trace data, or DTD (Boeschoten et al., 2020). Digital traces have shown the capability to measure dimensions of social life that are challenging or even impossible to assess through more conventional and classical sources (Pentland, 2010; Ledford, 2020). For example, survey experiments verified that differences in news reports can lead to the formation of different opinions based on the consumer's political motivations (Bolsen et al., 2014). But thanks only to digital trace data we can now observe the simultaneous effect of exposure to misinformation on opinion change (Ohme et al., 2023).

Digital trace data can be collected via different methods, which encompass application programming interfaces (API), data scraping, tracking, and data donation (e.g., Dongo et al., 2021; Ohme et al., 2023). Data donation is characterized as the voluntary act whereby an individual actively consents to contribute their personal data for research purposes (Skatova and Goulding, 2019). This approach relies on digital platforms' obligation to provide users with the data they have collected upon the user's request, guaranteed by General Data Protection Regulation (European Parliament and Council of the European Union, 2016). Participants in data donation studies receive their DTD in the form of 'data download packages' (DDPs) (Boeschoten, Ausloos, et al., 2022), compiled by companies such as Google and Meta (Bach et al., 2021).

For this recent data collection approach, uncertainties persist regarding the quality of donated DTD. In the context of Total Survey Error (Groves and Lyberg, 2010), challenges arise on both the representation and measurement sides. On the representation side, questions about data validity emerge due to factors like participant reluctance and lack of consent (Keusch and Kreuter, 2021; Boeschoten, Ausloos, et al., 2022; Boeschoten, Mendrik, et al., 2022). Measurement side concerns include low reliability and data extraction difficulties. In an adapted-to-big-data TSE called Total Error Framework (TEF) (Amaya et al., 2020), representation side problems are associated with nonresponse, compliance, consent, and extraction errors. Figure 1 illustrates the representation side of TEF and shows the steps where errors are introduced.

Common issues with digital trace data quality stem from researchers' challenges in addressing specific social groups, notably those less inclined to donate data due to lower technical skills (Vial, 2019; Ohme et al., 2021). The data request and upload process may be viewed as privacy-threatening and tiresome by respondents, potentially leading to imbalances in the available sample due to dropout. Since errors emerge in later stages of the data donation process, the issue may persist despite using a proper sampling frame. Thus, scholars warn that the quality of the donated data should be treated carefully since non-participation usually remains a large problem (Ohme et al., 2021; Silber et al., 2022). Our paper will center on the examination of how donated data is affected by non-participation and the potential avenues for enhancement.

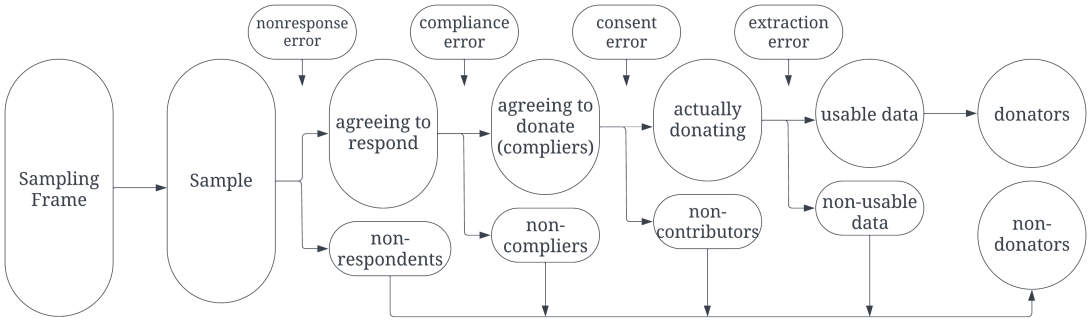

Figure 1. *Representation side of TEF in data donation process* [1]

1.2 DTD + Survey Data: Can It Provide a Solution?

Studies of survey non-participation remained a large field in social sciences for decades. An ongoing discussion led to formulation of several framework, such as leverage-salience theory (Groves et al., 2000; Groves and Couper, 2012) and social exchange theory (Cook et al., 2013). We adapt benefit-cost theory of participation (Singer, 2011), which is often seen as a synthesis of two mentioned theories. Benefit-cost theory is derived from a broader framework of rational choice theory (Singer and Presser, 2007): it identifies factors that may influence the respondent's trade-off by contributing to the costs or benefits of participation. These factors may be either objective, like financial incentives (benefit), or subjective, such as perceptions of disclosure risk (cost).

Particularly for donated DTD it is important to highlight that people stop participating on different stages of the data donation process. And while some factors are known for contributing to lower donation rates overall (Keusch et al., 2019; Gerdon et al., 2020), others are likely to be specific, being important for only one error source. Dropout varies across data donation steps, indicating a non-constant respondent utility function. Thus, it is not clear yet which mechanisms are responsible for error types described in Figure 1 and whether they can interact with each other, which calls for the further exploration.

A recently introduced approach enabling exploration and correction of specific error sources involves combining survey data with DTD (e.g., Stier et al., 2020; Silber et al., 2022; Ohme et al., 2023). However, papers on this topic only differentiated between donators and non-donators (Ohme et al., 2021), mixing all participant categories. Thus, it didn't allow for a thorough investigation into the specific reasons for dropout. This could pose a challenge, as comprehending these reasons is crucial for both increasing respondent participation rates and adjusting the available data.

A number of papers were recently published on DTD quality (Freelon, 2014; Vial, 2019; Stier

[1]The Representation side of the total error framework includes error sources contributing to data unrepresentativeness. By decomposing the data donation flow, TEF aids in identifying and mitigating specific errors at different process stages.

et al., 2020; Engel and Dahlhaus, 2021). In summary, these papers fall into two categories. Some employ exploratory analysis or predictive modeling to provide qualitative recommendations for optimizing the data collection process (Ohme et al., 2023). Others utilize various adjustment techniques on existing data to enhance its quality (Pak et al., 2022). Respectively, they can be called papers on **ex-ante** and **ex-post** error adjustment. However, few papers attempted to address both dimensions, potentially creating a gap in the literature as these strategies can be viewed as interdependent and mutually beneficial.

In this paper, we fill this gap by utilizing a strategy of combining survey data and DTD, taking the diversity of error sources into account. From it, two research questions can be formulated. First one aims to advance the theoretical understanding of non-donation by using the theories of survey participation: **Can different groups of non-donators be distinguished?** The second research question aims to provide technical solutions: **Can biases in donated DTD be reduced, given ex-ante and ex-post settings?**

1.3 Roots of Non-participation

We can hypothesize that the reasoning behind *nonresponse error* in data donation is similar to classical surveys: respondents were offered to participate in the survey, but they did not follow, ending up in the non-respondents group. Theoretically, it implies that they either weren't reached or had low interest in participation. Their Involvement was shaped by factors like rational reasoning (e.g., financial benefits), psychological traits (e.g., extraversion, agreeableness), and individual values (e.g., altruism) (Dillman et al., 2014). While individual factors may differ within this group, all non-respondents share a commonality: the function value wasn't sufficiently large to overcome the perceived survey burden (Marcus et al., 2007).

The situation is different for *compliance error* and the non-compliers group. Successfully contacted and added to the respondents, they later declined to participate in data donation. In studies combined with surveys, they completed the questionnaire but also refused to proceed. Hence, a higher degree of involvement is hypothesized for this group compared to non-respondents. Furthermore, we anticipate an emphasis on the role of trust in the data sharing process. It is probable that individuals were motivated to contribute but chose to discontinue due to privacy concerns, as observed in prior research (Ohme et al., 2021). Additionally, individuals in this category likely perceived themselves as incapable of uploading DDPs, indicating potentially low self-assessed technical skills.

A group of non-contributors includes those who agreed to donate but did not do so, resulting in *consent error*. This group may have high involvement, with non-contributors possibly feeling the need to appear socially desirable, prompting their initial agreement to donate. However, we also anticipate that many individuals in the non-contributors group, while facing fewer privacy concerns, simultaneously lacked the technical skills to donate data - skills that they may have overestimated before. (Ohme et al., 2021).

Among donators, we anticipate higher technical skills (Keusch et al., 2019). Also, device usage

type	factor	non-respondents	non-compliers	non-contributors	donators
Knowledge	Privacy concerns	Medium	High	Low	Low
	Technical skills	Medium	Medium	Low	High
	Device usage	Low	Medium	Medium	High
Values	Altruism	Low	Medium	High	High
	Civic Duty	Low	Medium	High	High
	Generalized trust	Low	Medium	High	High
Traits	Neuroticism	High	High	Medium	Low
	Extraversion	Low	Medium	High	High
	Agreeableness	Low	Medium	High	High

Table 1. *theoretically expected estimate of a participation-related factor, by group*[2]

time is considered not as a proxy for technical skills reducing participation costs, but as a factor increasing the likelihood of respondents viewing notifications - a proxy for availability. Moreover, we assume that donators tend to show more prosocial behavior and differ in altruism and civic duty from non-donators (Skatova and Goulding, 2019), similarly to blood donors. Furthermore, those with higher levels of generalized trust (e.g., Bjørnskov, 2007) are likely to have a propensity to cooperate more frequently in similar scenarios (Acedo-Carmona and Gomila, 2014). Psychologically, donators are expected to be more extroverted, as extraversion may heighten emotional passion for participation (Huber et al., 2021). We anticipate high agreeableness in donors, as donating data can be seen as a form of cooperative action.

We distinguish between three types of factors that contribute to involvement: knowledge-related, personal values and psychological traits. Knowledge-related factors, such as high familiarity with the devices, are usually responsible for lower costs of participation (Keusch et al., 2019; Silber et al., 2022). Personal values and psychological traits can generate non-material stimulus for participation. Thus, high altruism can motivate people to donate due to warm glow effect (Crumpler and Grossman, 2008), while high neuroticism, on the contrary, can contribute to increased stress, stimulating dropout (Mohiyeddini et al., 2015). While values are usually shaped by society, psychological traits remain more consistent across generations (Jang et al., 2001; Dawes et al., 2014). We find it important to distinguish between them because, in some cases (e.g., cross-national comparisons), values are likely to show a higher variance in comparison with psychological traits (Schmitt et al., 2007; Schwartz et al., 2014), which also may affect non-participation differently.

Thus, three groups of non-donators may differ significantly from each other, implying variations in social, economic, and demographic characteristics. Non-donators are not a monolithic group, and unexplored differences pose risks to data representativeness. Recognizing these distinctions can aid in addressing non-response types, reducing overall data error for empirical researchers. Mentioned factors that are hypothesised to be different among the groups are summarized in Table 1. Based on it, the following hypotheses are formulated:

H_0 : *The non-respondents group can be distinguished from other groups, based on differences in factors from Table 1*

H_1 : *The non-compliers group can be distinguished from other groups, based on differences in factors from Table 1*

H_2 : *The non-contributors group can be distinguished from other groups, based on differences in factors from Table 1*

The project unfolds in two parts. In the first part, we explore the differences and similarities between **non-respondents, non-compliers, non-contributors and donators**, using survey data. We also account for differences in socio-demographic profiles. In the second part, based on categorization of the participants and data adjustment, recommendations are provided that can allow to reduce total error both ex-post and ex-ante. Overall, we anticipate observing bias in the sample resulting from participants' self-selection. Our research endeavors to assist scholars planning to incorporate donated data into their projects, offering insights into potential sources of biases and corrective measures.

2 Analytic Strategy

2.1 Design and Research Procedure

In this study, we employ a combination of established methods and innovative approaches (Ohme et al., 2021; Silber et al., 2022) to integrate survey and digital trace data. Working with survey data, we are constrained to non-parametric approaches suited for processing ordinal and nominal variables. The investigation will address research questions by conducting between-group comparisons, developing a participation prediction model, categorizing respondents, and adjusting for potential biases.

First, we compare the differences between the groups of non-respondents, non-compliers, non-contributors, and donators. Non-parametric Mann-Whitney U-tests are implemented to check whether differences among groups can be seen. Then, we utilize a binary logistic model to estimate which variables are related with actual participation, distinguishing between donators and non-donators. After it, multinomial logistic regression is used. We will model the assumed relationship between response categories, exploring factors influencing transitions between them. This part allows us to explore traits of those who share and do not share their data. As a result, it allows us to **test our hypotheses**.

In the next step, we categorize participants by performing suitable non-parametric clustering approaches. They include K-Modes Clustering, Agglomerative Hierarchical Clustering (AHC),

[2]Names for different groups of non-donators were either taken from previous research in the field (Ohme et al., 2021; Boeschoten, Ausloos, et al., 2022) or formulated based on a relevant error type.

and Latent Class Analysis (LCA) (Tatiana et al., 2009). Several clustering algorithms are used to compare their performance (Rodriguez et al., 2019). K-Modes Clustering is an extension of the K-Means algorithm designed for categorical data. It aims to find cluster centroids that represent modes for each categorical attribute. The elbow method (Syakur et al., 2018) is used to choose number of clusters. AHC is adapted to handle categorical data by using appropriate dissimilarity measures designed for categorical variables. We use dendrogram analysis for determining the optimal number of clusters. Hyperparameters such as linkage method and distance metric are chosen based on the resulting cluster structure and goodness-of-fit measures. LCA allows to model the probability of observing certain patterns of categorical responses within clusters. To determine the optimal number of classes, we fit LCA with different class counts and compare models. This part expands our understanding of people's motives to participate by describing related socio-demographic profiles, thus allowing us to access unique behavioral strategies and providing insights on **ex-ante data quality** improvement.

Finally, we use inverse probability weighting (IPW) and censored regression (Type II Tobit Model) to adjust for biases (Pak et al., 2022). The purpose of IPW here is to balance the groups with respect to covariates, and Type II Tobit Model can be used similarly. For IPW, the quality of weights is addressed, and stepwise selection is employed to optimise the algorithm. For the Tobit model, errors are assumed to have a logistic distribution. We also use Least Squares Weighted Regression (LSWR), which is a strategy not previously used on combined data. For LSWR, cross-validation is performed to determine regularization strength optimal to minimize prediction error. The performance of different strategies is compared using a close-to-population sampling frame dataset. Overall, this step addresses **ex-post data quality**.

All described statistical procedures were performed in Rstudio (RStudio Team, 2020), with usages of such packages as *mixtools* (Tatiana et al., 2009), *ipw* (van der Wal and Geskus, 2011), *glmnet* (Friedman et al., 2010), *cluster* (Maechler et al., 2023), *poLCA* (Linzer and Lewis, 2011), *survival* (Therneau, 2023), *MASS*, and *nnet* (Venables and Ripley, 2002). Package *WhatsR* which allows to process WhatsApp data, is also used (Kohne et al., 2022; Pak et al., 2022).

2.2 Data and Measures

Data from the LISS panel, probability-based online panel of the general population in the Netherlands, Core Studies Wave 15 (2023) and WhatsApp data donation project (2023) is used in the research (Scherpenzeel, 2018; Corten and Boeschoten, 2022). The LISS panel is a longitudinal study that collects data through surveys yearly since 2007. The content of the surveys may cover a wide range of subjects, including health, work, education, economics, values, and political attitudes. Socio-demographic data on each respondent is also included as background variables. By using identification tokens, it becomes possible to associate each participant in the WhatsApp data study with prior information gathered in LISS studies, offering a unique opportunity to find measurements for all concepts mentioned in the theoretical part.

The WhatsApp data donation project incorporated a sample of 4,800 randomly selected

panel members aged 16 years and above (Corten and Boeschoten, 2022), gathered in February and April of 2023. Participants were asked to hand in DDP's (Boeschoten, Ausloos, et al., 2022) on their first group chat and/or account information. The packages underwent local processing on the respondent's device using Eyra's PORT to safeguard the respondent's privacy (Boeschoten, Mendrik, et al., 2022). If respondents did not have the device or WhatsApp, they were excluded prior to data collection. Besides donation, respondents completed a questionnaire, including inquiries about device usage, language settings, and smartphone activities. By default, questions were asked before data donation, but some were posed only after the data donation phase concluded. [3]. Out of 4800 (100%) participants, 3220 (67.1%) fully completed the survey. Non-response amounted to 1202 (25%) panel members, while 378 (7.9%) left it to some extent incomplete. In summary, 809 (16.6%) DDP's were donated: 460 group chat packages and 349 account information packages. Thus, a comparatively small share of the panel members actually donated data by the end of the study, while most joined the ranks of non-respondents, non-compliers, and non-contributors.

Measurements for knowledge-related factors such as privacy concerns, technical skills and device usage are taken from *the WhatsApp data donation* questionnaire. As a part of it, respondents were asked to indicate their level of familiarity with smartphones, whether they are concerned about their privacy, and how often they use smartphones. For other independent variables, indicators are chosen from the LISS panel core studies based on their conceptual relevance. In utilized core modules, sample size vary from 6518 to 7056 participants, and response rates vary from 80.3% to 86.3 %. *LISS Core Study 4: Social Integration and Leisure*, which includes a number of measurements on volunteering and helping others, is used to estimate altruism (Enelamah and Tran, 2020). Measurements of psychological traits such as extraversion, neuroticism, and agreeableness are taken from the *LISS Core Study 7: Personality*, which included original questions on the big five personality traits (Soldz and Vaillant, 1999). Generalised trust "standard question" (Uslaner, 2012) from this study is also utilized. Question on contribution to society, which measures civic duty (Blais and Galais, 2016) is taken from *LISS Core Study 8: Politics and Values*. Background socio-demographic variables are also utilized as control variables.

References

Acedo-Carmona, C., & Gomila, A. (2014). Personal trust increases cooperation beyond general trust. *PLoS One*, *9*(8), e105559.

Amaya, A., Biemer, P. P., & Kinyon, D. (2020). Total error in a big data world: Adapting the tse framework to big data. *Journal of Survey Statistics and Methodology*, *8*(1), 89–119.

[3]Questions asked after data donation mostly covered donation process. For example, participants were asked whether they were "able to share WhatsApp account information with the researchers" and "What went wrong throughout data donation"

Bach, R. L., Kern, C., Amaya, A., Keusch, F., Kreuter, F., Hecht, J., & Heinemann, J. (2021). Predicting voting behavior using digital trace data. *Social Science Computer Review*, *39*(5), 862–883.

Bjørnskov, C. (2007). Determinants of generalized trust: A cross-country comparison. *Public choice*, *130*(1-2), 1–21.

Blais, A., & Galais, C. (2016). Measuring the civic duty to vote: A proposal. *Electoral Studies*, *41*, 60–69.

Boeschoten, L., Ausloos, J., Moeller, J., Araujo, T., & Oberski, D. L. (2020). Digital trace data collection through data donation. *arXiv preprint arXiv:2011.09851*.

Boeschoten, L., Ausloos, J., Möller, J. E., Araujo, T., & Oberski, D. L. (2022). A framework for privacy preserving digital trace data collection through data donation. *Computational Communication Research*, *4*(2), 388–423.

Boeschoten, L., Mendrik, A., van der Veen, E., Vloothuis, J., Hu, H., Voorvaart, R., & Oberski, D. L. (2022). Privacy-preserving local analysis of digital trace data: A proof-of-concept. *Patterns*, *3*(3).

Bolsen, T., Druckman, J. N., & Cook, F. L. (2014). The influence of partisan motivated reasoning on public opinion. *Political Behavior*, *36*, 235–262.

Cook, K. S., Cheshire, C., Rice, E. R., & Nakagawa, S. (2013). Social exchange theory. *Handbook of social psychology*, 61–88.

Corten, R., & Boeschoten, L. (2022). *Whatsapp data donation. a study integrated in the liss panel: Collecting digital trace data through eyra port*. Centerdata.

Crumpler, H., & Grossman, P. J. (2008). An experimental test of warm glow giving. *Journal of public Economics*, *92*(5-6), 1011–1021.

Dawes, C., Cesarini, D., Fowler, J. H., Johannesson, M., Magnusson, P. K., & Oskarsson, S. (2014). The relationship between genes, psychological traits, and political participation. *American Journal of Political Science*, *58*(4), 888–903.

Dillman, D. A., Smyth, J. D., & Christian, L. M. (2014). *Internet, phone, mail, and mixed-mode surveys: The tailored design method*. John Wiley & Sons.

Dongo, I., Cardinale, Y., Aguilera, A., Martinez, F., Quintero, Y., Robayo, G., & Cabeza, D. (2021). A qualitative and quantitative comparison between web scraping and api methods for twitter credibility analysis. *International Journal of Web Information Systems*, *17*(6), 580–606.

Enelamah, N. V., & Tran, T. (2020). Dimensions of altruism behaviors among americans in the general social survey. *Journal of Human Behavior in the Social Environment*, *30*(2), 213–227.

Engel, U., & Dahlhaus, L. (2021). Data quality and privacy concerns in digital trace data: Insights from a delphi study on machine learning and robots in human life. In *Handbook of computational social science, vol 1*. Taylor & Francis.

European Parliament & Council of the European Union. (2016, May 4). *Regulation (EU) 2016/679 of the European Parliament and of the Council* [Of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)]. Retrieved April 13, 2023, from https://data.europa.eu/eli/reg/2016/679/oj

Freelon, D. (2014). On the interpretation of digital trace data in communication and social computing research. *Journal of Broadcasting & Electronic Media*, *58*(1), 59–75.

Friedman, J., Tibshirani, R., & Hastie, T. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, *33*(1), 1–22. https://doi.org/10.18637/jss.v033.i01

Gerdon, F., Nissenbaum, H., Bach, R. L., Kreuter, F., & Zins, S. (2020). Individual acceptance of using health data for private and public benefit: Changes during the covid-19 pandemic. *Harvard Data Science Review*.

Groves, R. M., & Couper, M. P. (2012). *Nonresponse in household interview surveys*. John Wiley & Sons.

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

Groves, R. M., Singer, E., & Corning, A. (2000). Leverage-saliency theory of survey participation: Description and an illustration. *The Public Opinion Quarterly*, *64*(3), 299–308.

Huber, B., Goyanes, M., & Gil de Zúñiga, H. (2021). Linking extraversion to collective and individual forms of political participation: The mediating role of political discussion. *Social Science Quarterly*, *102*(4), 1289–1310.

Jang, K. L., Livesley, W. J., Riemann, R., Vernon, P. A., Hu, S., Angleitner, A., Ando, J., Ono, Y., & Hamer, D. H. (2001). Covariance structure of neuroticism and agreeableness: A twin and molecular genetic analysis of the role of the serotonin transporter gene. *Journal of Personality and Social Psychology*, *81*(2), 295.

Keusch, F., & Kreuter, F. (2021). Digital trace data: Modes of data collection, applications, and errors at a glance. In *Handbook of computational social science, vol 1*. Taylor & Francis.

Keusch, F., Struminskaya, B., Antoun, C., Couper, M. P., & Kreuter, F. (2019). Willingness to Participate in Passive Mobile Data Collection. *Public Opinion Quarterly*, *83*(S1), 210–235. https://doi.org/10.1093/poq/nfz007

Kohne, J., Elhai, J. D., & Montag, C. (2022). A practical guide to whatsapp data in social science research. In *Digital phenotyping and mobile sensing: New developments in psychoinformatics* (pp. 171–205). Springer.

Ledford, H. (2020). How facebook, twitter and other data troves are revolutionizing social science. *Nature*, *582*(7812), 328–331.

Linzer, D. A., & Lewis, J. B. (2011). poLCA: An R package for polytomous variable latent class analysis. *Journal of Statistical Software*, *42*(10), 1–29. https://www.jstatsoft.org/v42/i10/

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., & Hornik, K. (2023). *Cluster: Cluster analysis basics and extensions* [R package version 2.1.6 — For new features, see the 'NEWS' and the 'Changelog' file in the package source)]. https://CRAN.R-project.org/package=cluster

Marcus, B., Bosnjak, M., Lindner, S., Pilischenko, S., & Schütz, A. (2007). Compensating for low topic interest and long surveys: A field experiment on nonresponse in web surveys. *Social Science Computer Review*, *25*(3), 372–383.

Mohiyeddini, C., Bauer, S., & Semple, S. (2015). Neuroticism and stress: The role of displacement behavior. *Anxiety, stress, & coping*, *28*(4), 391–407.

Ohme, J., Araujo, T., Boeschoten, L., Freelon, D., Ram, N., Reeves, B. B., & Robinson, T. N. (2023). Digital trace data collection for social media effects research: Apis, data donation, and (screen) tracking. *Communication Methods and Measures*, 1–18.

Ohme, J., Araujo, T., de Vreese, C. H., & Piotrowski, J. T. (2021). Mobile data donations: Assessing self-report accuracy and sample biases with the ios screen time function. *Mobile Media & Communication*, *9*(2), 293–313.

Pak, C., Cotter, K., & Thorson, K. (2022). Correcting sample selection bias of historical digital trace data: Inverse probability weighting (ipw) and type ii tobit model. *Communication Methods and Measures*, *16*(2), 134–155.

Pentland, A. (2010). *Honest signals: How they shape our world*. MIT press.

Rodriguez, M. Z., Comin, C. H., Casanova, D., Bruno, O. M., Amancio, D. R., Costa, L. d. F., & Rodrigues, F. A. (2019). Clustering algorithms: A comparative approach. *PloS one*, *14*(1), e0210236.

RStudio Team. (2020). *Rstudio: Integrated development environment for r*. RStudio, PBC. Boston, MA. http://www.rstudio.com/

Scherpenzeel, A. C. (2018). "true" longitudinal and probability-based internet panels: Evidence from the netherlands. In *Social and behavioral research and the internet* (pp. 77–104). Routledge.

Schmitt, D. P., Allik, J., McCrae, R. R., & Benet-Martínez, V. (2007). The geographic distribution of big five personality traits: Patterns and profiles of human self-description across 56 nations. *Journal of cross-cultural psychology*, *38*(2), 173–212.

Schwartz, S. H., Caprara, G. V., Vecchione, M., Bain, P., Bianchi, G., Caprara, M. G., Cieciuch, J., Kirmanoglu, H., Baslevent, C., Lönnqvist, J.-E., et al. (2014). Basic personal values underlie and give coherence to political values: A cross national study in 15 countries. *Political Behavior*, *36*, 899–930.

Silber, H., Breuer, J., Beuthner, C., Gummer, T., Keusch, F., Siegers, P., Stier, S., & Weiß, B. (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*(Supplement_2), S387–S407.

Singer, E. (2011). Toward a benefit-cost theory of survey participation: Evidence, further tests, and implications. *Journal of Official Statistics*, *27*(2), 379–392.

Singer, E., & Presser, S. (2007). Privacy, confidentiality, and respondent burden as factors in telephone survey nonresponse. *Advances in telephone survey methodology*, 447–470.

Skatova, A., & Goulding, J. (2019). Psychology of personal data donation. *PloS one*, *14*(11), e0224240.

Soldz, S., & Vaillant, G. E. (1999). The big five personality traits and the life course: A 45-year longitudinal study. *Journal of research in personality*, *33*(2), 208–232.

Stier, S., Breuer, J., Siegers, P., & Thorson, K. (2020). Integrating survey data and digital trace data: Key issues in developing an emerging field.

Syakur, M., Khotimah, B., Rochman, E., & Satoto, B. D. (2018). Integration k-means clustering method and elbow method for identification of the best customer profile cluster. *IOP conference series: materials science and engineering*, *336*, 012017.

Tatiana, B., Didier, C., David, R., & Derek, Y. (2009). Mixtools: An r package for analyzing finite mixture models. *Journal of Statistical Software*, *32*(6), 1–29.

Therneau, T. M. (2023). *A package for survival analysis in r* [R package version 3.5-7]. https://CRAN.R-project.org/package=survival

Uslaner, E. M. (2012). Measuring generalized trust: In defense of the 'standard'question. *Handbook of research methods on trust*, *72*.

van der Wal, W. M., & Geskus, R. B. (2011). Ipw: An r package for inverse probability weighting. *Journal of Statistical Software*, *43*(13), 1–23. https://doi.org/10.18637/jss.v043.i13

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with s* (Fourth) [ISBN 0-387-95457-0]. Springer. https://www.stats.ox.ac.uk/pub/MASS4/

Vial, G. (2019). Reflections on quality requirements for digital trace data in is research. *Decision Support Systems*, *126*, 113133.