Public Opinion Quarterly (2023) **Vol 87 Special Issue,** 475–479

Augmenting Surveys with Paradata, Administrative Data, and Contextual Data

Joseph W. Sakshaug[1,*], Bella Struminskaya[2]

[1]Distinguished Researcher, Department of Statistical Methods, Institute for Employment Research, Nuremberg, Germany; and Professor, Department of Statistics, Ludwig-Maximilian University of Munich, Munich, Germany
[2]Associate Professor, Department of Methodology and Statistics, Utrecht University, Utrecht, The Netherlands

Introduction

Over the last decades, there has been growing interest in augmenting survey data with alternative data sources, such as paradata, administrative data, and contextual data. *Paradata*, for instance, refers to data related to the process of collecting survey data during the field period, which are not directly derived from respondents' answers to survey questions, but rather are a by-product of the data collection process (Couper 1998; Kreuter 2013). This may include data from call records, keystroke data in computer-administered surveys, interviewer observations, and more. *Administrative data* refers to externally created process data that are often linked to individual respondent records by matching personal information (Calderwood and Lessof 2009). For instance, social surveys may link respondents' interview data (conditional on consent) to tax, insurance, voter registration, and other government databases. Finally, *contextual data* comprises external sources of information that measure various aspects of the respondent's physical, social, or informational environment (Fortin-Rittberger et al. 2016). This could involve aggregate data on the demographic composition of a respondent's neighborhood or organizational characteristics of their place of work, as well as data on respondents' behaviors, environments, and social networks from wearables, sensors, apps, and digital traces from social media or web browsing.

Survey researchers are increasingly utilizing these data sources to enhance their substantive and methodological research and address complex research questions that are difficult (or impossible) to answer using survey data alone. Paradata, for instance, are employed in survey production to monitor fieldwork, increase data collection efficiencies, investigate measurement errors, and assess and correct for nonresponse errors (Biemer et al. 2013; Wagner et al. 2012; Yan and Olson 2013). Likewise, linked administrative data are

Corresponding author: Department of Statistical Methods, Institute for Employment Research, Regensburger Str. 104, Nuremberg 90478, Germany; email: joe.sakshaug@iab.de.

Advance Access publication June 10, 2023

https://doi.org/10.1093/poq/nfad026

frequently utilized in conjunction with survey data to bolster substantive analyses and assess and correct for nonresponse and measurement errors in methodological applications (Kreuter, Müller, and Trappmann 2010; Meyer and Mittag 2019). Contextual data can also supplement survey data in substantive research by facilitating the analysis of neighborhood effects, as well as proximity and access to services in respondents' geographic areas (Dick 2022). In addition, newer forms of contextual data, such as digital trace data obtained from respondents' smartphones, sensors, and social media, are being explored for several purposes in survey research, such as predicting voting behavior (Bach et al. 2021), measuring online media consumption (Cernat and Keusch 2022), and studying the effects of collecting passive data on panel survey retention (Trappmann et al. 2022).

Although interest in augmenting surveys with paradata, administrative data, and contextual data will likely continue to grow, significant challenges and unresolved issues remain. For instance, auxiliary data sources may be subject to selection issues, as they may not be observable or linkable to all survey respondents. Additionally, a nonrandom subset of respondents may be reluctant to provide consent to augment their interview data with these alternative sources (Jenkins et al. 2006; Sakshaug and Kreuter 2012; Struminskaya et al. 2020). A key question, therefore, is: How can researchers determine the extent of selection errors in the augmented data, and effectively minimize them? Furthermore, issues of measurement accuracy can arise. As these external data sources are often used to assess the validity of survey measurements, researchers typically assume that these data are error-free, or at least more accurate than the survey data. However, the validity of this assumption does not always hold in practice (Kapteyn and Ypma 2007). Thus, a critical question facing researchers is under what conditions this assumption is valid, and how to test its validity. Finally, although the use of digital trace data in survey research is still emerging and only a few surveys collect these data, what other potential uses and applications of these data could persuade survey programs to collect them routinely?

The Special Issue

This special issue features five papers that showcase innovative research using one or more of these survey enhancements to address important research questions in survey research, as well as a paper discussing ethical issues of collecting such data. The issue begins with two papers on paradata. Garbarski et al. (2023) investigate how interviewers' evaluations of respondents' performance relate to respondents' behaviors and response quality during interviews, as well as their relationship with sociodemographic characteristics of both interviewers and respondents. Their findings support

the notion that interviewer evaluations accurately reflect the interview situation, reinforcing their credibility as a measure of data quality. Gummer et al. (2023) use web survey client-side paradata on browser window and tab switching to examine the issue of respondents looking up answers to political knowledge questions in online surveys. Their results indicate that a nontrivial share of respondents look up answers, leading to higher rates of correct answers, but that providing explicit instructions not to look up answers can reduce this behavior.

Bollinger and Tasseva's article (2023), which uses data from a household panel survey linked to administrative unemployment benefits and earnings, examines misreporting in benefit programs and earnings. They identify evidence of underreporting of benefits and misreporting of the program from which benefits are received, as well as cases where benefits are reported as earnings. West and Andridge (2023) propose a measure to evaluate non-ignorable selection bias in pre-election polling estimates, which relies on aggregate information for the nonselected likely voter population. They evaluate this measure using pre-election polls conducted before the 2020 US presidential election and the 2015 general election in Great Britain. Henninger et al. (2023) investigate attitudes toward privacy in relation to mouse-tracking paradata collection. Using a vignette experiment, they study factors that influence willingness to participate in a survey that collects mouse-tracking data, finding that respondents were less willing to participate in a survey that included mouse-tracking data collection when the requests for the survey participation and paradata collection were posed sequentially versus when both requests were bundled together, and that explaining the purpose of the mouse-tracking data collection did not increase willingness to participate. These studies are part of a special issue that features innovative research using survey enhancements to address substantive and methodological questions. The special issue concludes with a paper on research ethics and challenges when augmenting surveys with paradata, administrative data, and contextual data, authored by the guest editors (Struminskaya and Sakshaug 2023).

Conclusion

To conclude, the above papers highlight the innovative possibilities that arise from augmenting surveys with these alternative data sources, but also point to the challenges involved in their acquisition and integration. As the survey profession moves forward, we anticipate witnessing continued advancements in this field, building on the progress made over the past decades. We wish to thank everyone who submitted an article for the special issue. We are grateful to all authors and reviewers for their time, efforts, and commitment.

Finally, we would like to thank the editors in chief of *POQ*, Eric Plutzer and Allyson Holbrook, for their unwavering support along the way.

References

Bach, Ruben L., Christoph Kern, Ashley Amaya, Florian Keusch, Frauke Kreuter, Jan Hecht, and Jonathan Heinemann. 2021. "Predicting Voting Behavior using Digital Trace Data." *Social Science Computer Review* 39:862–83. https://doi.org/10.1177/0894439319882896.

Biemer, Paul P., Patrick Chen, and Kevin Wang. 2013. "Using Level-of-Effort Paradata in Non-Response Adjustments with Application to Field Surveys." *Journal of the Royal Statistical Society Series A: Statistics in Society* 176:147–68. https://doi.org/10.1111/j.1467-985X.2012.01058.x.

Bollinger, Christopher Robert, and Iva Valentinova Tasseva. 2023. "Income Source Confusion using the SILC." *Public Opinion Quarterly* 87:542–74.

Calderwood, Lisa, and Carli Lessof. 2009. "Enhancing Longitudinal Surveys by Linking to Administrative Data." In *Methodology of Longitudinal Surveys*, edited by Peter Lynn, 55–72. Chichester, West Sussex, UK: Wiley. https://doi.org/10.1002/9780470743874.ch4.

Cernat, Alexandru, and Florian Keusch. 2022. "Do Surveys Change Behaviour? Insights from Digital Trace Data." *International Journal of Social Research Methodology* 25:79–90. https://doi.org/10.1080/13645579.2020.1853878.

Couper, Mick P. 1998. "Measuring Survey Quality in a CASIC Environment." Invited paper presented at the Joint Statistical Meetings of the American Statistical Association, Dallas, August. http://www.asasrms.org/Proceedings/papers/1998_006.pdf.

Dick, Christopher. 2022. "The Health and Retirement Study: Contextual Data Augmentation." *Forum for Health Economics & Policy* 25:29–40. https://doi.org/10.1515/fhep-2021-0068.

Fortin-Rittberger, Jessica, David Howell, Stephen Quinlan, and Bojan Todosijevic. 2016. "Supplementing Cross-National Survey Data with Contextual Data." In *The Sage Handbook of Survey Methodology*, edited by Christof Wolf, Dominique Joye, Tom W. Smith, and Yang-Chih Fu, 670–80. London: Sage Publications. https://www.researchgate.net/profile/Bojan-Todosijevic/publication/303381858_Supplementing_cross-national_surveys_with_contextual_data/links/5a5a1c31aca2727d6085e3ee/Supplementing-cross-national-surveys-with-contextual-data.pdf.

Garbarski, Dana, Jennifer Dykema, Nora Cate Schaeffer, Cameron Jones, Tiffany S. Neman, and Dorothy Farrer-Edwards. 2023. "Factors Associated with Interviewers' Evaluations of Respondents' Performance in Telephone Interviews: Behavior, Response Quality Indicators, and Characteristics of Respondents and Interviewers." *Public Opinion Quarterly* 87:480–506.

Gummer, Tobias, Tanja Kunz, Tobias Rettig, and Jan Karem Höhne. 2023. "How to Detect and Influence Looking up Answers to Political Knowledge Questions in Web Surveys." *Public Opinion Quarterly* 87:507–41.

Henninger, Felix, Pascal J. Kieslich, Amanda Fernández-Fontelo, Sonja Greven, and Frauke Kreuter. 2023. "Privacy Attitudes Toward Mouse-Tracking Paradata Collection." *Public Opinion Quarterly* 87:602–18.

Jenkins, Stephen P., Lorenzo Cappellari, Peter Lynn, Annette Jäckle, and Emanuela Sala. 2006. "Patterns of Consent: Evidence from a General Household Survey." *Journal of the Royal Statistical Society Series A: Statistics in Society* 169:701–22. https://doi.org/10.1111/j.1467-985X.2006.00417.x.

Kapteyn, Arie, and Jelmer Y. Ypma. 2007. "Measurement Error and Misclassification: A Comparison of Survey and Administrative Data." *Journal of Labor Economics* 25:513–51. http://dx.doi.org/10.1086/513298.

Kreuter, Frauke, ed. 2013. *Improving Surveys with Paradata: Analytic Uses of Process Information*. Hoboken, NJ: Wiley.

Kreuter, Frauke, Gerrit Müller, and Mark Trappmann. 2010. "Nonresponse and Measurement Error in Employment Research: Making Use of Administrative Data." *Public Opinion Quarterly* 74:880–906. https://doi.org/10.1093/poq/nfq060.

Meyer, Bruce D., and Nikolas Mittag. 2019. "Using Linked Survey and Administrative Data to Better Measure Income: Implications for Poverty, Program Effectiveness, and Holes in the Safety Net." *American Economic Journal: Applied Economics* 11:176–204. https://doi.org/10.1257/app.20170478.

Sakshaug, Joseph W., and Frauke Kreuter. 2012. "Assessing the Magnitude of Non-Consent Biases in Linked Survey and Administrative Data." *Survey Research Methods* 6:113–22. https://doi.org/10.18148/srm/2012.v6i2.5094

Struminskaya, Bella, Vera Toepoel, Peter Lugtig, Marieke Haan, Annemieke Luiten, and Barry Schouten. 2020. "Understanding Willingness to Share Smartphone-Sensor Data." *Public Opinion Quarterly* 84:725–59. https://doi.org/10.1093/poq/nfaa044.

Struminskaya, Bella, and Joseph W. Sakshaug. 2023. "Ethical Considerations for Augmenting Surveys with Auxiliary Data Sources." *Public Opinion Quarterly* 87:619–33.

Trappmann, Mark, Georg-Christoph Haas, Sonja Malich, Florian Keusch, Sebastian Bähr, Frauke Kreuter, and Stefan Schwarz. 2022. "Augmenting Survey Data with Digital Trace Data: Is There a Threat to Panel Retention?" *Journal of Survey Statistics and Methodology*. https://doi.org/10.1093/jssam/smac023.

Wagner, James, Brady T. West, Nicole Kirgis, James M. Lepkowski, William G. Axinn, and Shonda Kruger Ndiaye. 2012. "Use of Paradata in a Responsive Design Framework to Manage a Field Data Collection." *Journal of Official Statistics* 28:477–99. https://www.scb.se/contentassets/f6bcee6f397c4fd68db6452fc9643e68/use-of-paradata-in-a-responsive-design-framework-to-manage-a-field-data-collection.pdf.

West, Brady T., and Rebecca R. Andridge. 2023. "Evaluating Pre-Election Polling Estimates Using a New Measure of Non-Ignorable Selection Bias." *Public Opinion Quarterly* 87:575–601.

Yan, Ting, and Kristen Olson. 2013. "Analyzing Paradata to Investigate Measurement Error." In *Improving Surveys with Paradata: Analytic Uses of Process Information*, edited by Frauke Kreuter, 73–95. Hoboken, NJ: Wiley. https://doi.org/10.1002/9781118596869.ch4.