

Learning Through Failures. What Unsuccessful Survey Experiments Teach Us About Design, Measurement, and Inference

Lena Hipp^{1,2}  · Sebastian E. Wenz³ 

¹WZB Berlin Social Science Center

²University of Potsdam

³GESIS Leibniz Institute for the Social Sciences

Survey experiments have become an indispensable tool in the social sciences for estimating causal effects in diverse populations. Yet, the apparent simplicity of experimental designs can be deceptive, as weak treatments, noncompliance, or implementation breakdowns may lead to statistically nonsignificant or otherwise unexpected results that leave researchers wondering whether to conclude that their studies have failed or show true null effects. This special issue of *Survey Research Methods* presents empirical and methodological contributions that examine survey experiments that did not work as planned. Rather than dismissing unexpected or nonsignificant results, the papers address key challenges in treatment validation, context dependency, treatment-related nonresponse, precisely articulated causal states, field implementation, and indirect questioning techniques. Collectively, the contributions illustrate that unexpected findings reveal critical assumptions connecting experimental design to causal inference. By emphasizing pretesting, field monitoring, paradata, manipulation checks and other diagnostic measures, this special issue highlights the value of publishing carefully diagnosed ‘failures’ to advance cumulative methodological knowledge.

Keywords: survey experiments; implementation; treatment failure; list experiments; interviewer; inclusive language

Survey experiments have become an indispensable part of the methodological toolkit in the social sciences for studying attitudes, beliefs, preferences, and sensitive behaviors (Mize & Manago, 2022; Rauf et al., 2025; Schachter & Weisshaar, 2025; Treischl & Wolbring, 2022). By combining random assignment with the reach and flexibility of surveys, they provide a powerful framework for estimating causal effects in diverse populations and settings. Increasingly sophisticated designs—including factorial surveys, choice experiments, list experiments, and survey-embedded field interventions—have extended this promise even further.

Yet the apparent simplicity of survey experiments can be deceptive. Randomly assigning respondents to conditions does not, by itself, ensure that respondents comply with the

experimental task and interpret it as expected, nor does it ensure that the intended treatment is effectively delivered and translates into a meaningful difference in the outcome of interest. When any of these links breaks down and the resulting statistical estimate does not answer the substantive question of interest (Lundberg et al., 2021), researchers may therefore conclude that their experiment “failed”.

The SRM special issue *Learning Through Failures* originated from two sessions at the 2023 ESRA Conference in Milan and focuses on precisely these complications. Its contributions examine survey experiments that did not work as planned: Treatments that failed to shift the intended construct, manipulations that respondents interpreted differently from researchers’ expectations, indirect-questioning techniques that generated implausible estimates, interventions that were undermined by nonresponse or noncompliance, and field implementations that produced little evidence of their intended effects. Rather than treating these experiences as unfortunate deviations from otherwise suc-

Corresponding author: Lena Hipp, WZB Berlin Social Science Center, Berlin, Germany (Email: hipp@wzb.eu)

successful research programs and keeping them in the file drawer (Chopra et al., 2024; Rosenthal, 1979), this special issue treats them as opportunities for methodological diagnosis and, ultimately, for learning how to design, measure, and draw inferences from survey experiments more effectively.

While neither the call for the sessions at ESRA 2023 nor the call for this special issue provided an explicit definition of failure, we edited this special issue on the premise that neither a failure to reject the null hypothesis nor an unexpected result implies that an experiment has failed (Kane, 2024). A precisely estimated null effect can be empirically and theoretically as relevant as a statistically significant effect (Abadie, 2020). Conversely, survey experiments that “failed” can nonetheless yield statistically significant outcomes. Therefore, what is and what is not a “failure” must be understood to the research question of the experiment, and the assumptions connecting the intervention to the observed data. Accordingly, many of the contributions in the special issue are better understood as studies of unexpected findings that illuminate the limits of survey experiments.

The first set of papers assembled in the special issue reports unexpected findings due to experimental treatments failing to produce the intended manipulation: They report failures of non-deceptive, real-world treatments that were presumably too weak, insufficiently tested, or rarely occur in reality (Hipp & Leumann, 2026; Hofstetter et al., 2026; Kunz et al., 2026). They also show that treatment effectiveness is context-dependent and may not travel across countries, historical moments, survey modes, or populations (Jungkunz & Mayer, 2026; Schwindt-Bayer et al., 2026).

The second set of papers deals with unexpected findings that likely resulted from treatment-related nonresponse and selective participation: Response behavior may itself become part of the causal process when treatments influence who responds, who complies, or who is included in the analysis (Hensgen & Zimmermann, 2026; Waldendorf et al., 2026). Differential noncompliance can undermine comparability, while item nonresponse may constitute post-treatment selection.

Closely related is the next group of papers that addresses the often-overlooked role of control conditions in survey experiments: Seemingly neutral control conditions may fail to provide a valid counterfactual if respondents infer omitted information (Theodoridis & Goggin, 2026) or comply differently across conditions (Zhang & Baldassarri, 2026). Or, put differently, omitted information is still information that may influence the outcome of a study.

A fourth group of papers highlights implementation challenges: Missing information about how respondents access a survey can make it impossible to determine whether an intervention is ineffective or simply not used (Maslovskaya et al., 2026), and interviewer behavior can alter treatment

delivery itself (Litchfield & Douarin, 2026). These examples of potential failure show that implementation is part of the experimental design and causal process. Detailed paradata on what respondents actually saw, did, and experienced are therefore essential for diagnosing unexpected findings. Vendors, programmers, interviewers, survey platforms, and field conditions all become part of the experimental design. Researchers should therefore anticipate plausible implementation failures before fieldwork and collect the information needed to diagnose them.

Finally, the last two papers illustrate the challenges associated with indirect questioning techniques: These methods depend on assumptions regarding list balance, respondent comprehension, trust, design effects, and truthful reporting. Formal anonymity does not necessarily translate into perceived anonymity, and respondents may seek to protect themselves, preserve identities, avoid stigma, signal cooperation, or resist the task (León & Aizpurua, 2026; Sajons et al., 2026). Consequently, researchers must distinguish between measurement failure and evidence of absence: An invalid or insensitive instrument renders null findings difficult to interpret rather than substantively uninformative. More generally, direct questions, indirect questioning techniques, behavioral traces, qualitative follow-ups, expert assessments, and respondent probes should be viewed as complementary diagnostic tools rather than competing alternatives.

Taken together, these contributions do not simply report null or unexpected findings but explore where in the causal process experiments departed from researchers’ expectations and why. Collectively, they illustrate that treatment validity depends not only on random assignment but also on whether treatments and controls are plausible and meaningful, perceived as intended, interpreted consistently across respondents, and capable of manipulating the underlying construct.

These findings also point to ways of preventing avoidable failures and, when unexpected results do occur, making them more readily diagnosable. Careful treatment validation and cognitive and technical pretesting can reveal problems with treatments, controls, and measurement before an experiment is fielded. Collecting information on treatment exposure, compliance, perceptions, and implementation can then help identify where the experimental process departed from expectations. Manipulation checks, perception measures, paradata, interviewer observations, and follow-up studies should therefore be considered part of the experimental design rather than optional additions. Such diagnostic evidence is essential for distinguishing treatment or implementation failure from a true null effect and measurement failure from evidence of absence.

More generally, this special issue argues for a shift in how unsuccessful survey experiments are viewed. Unex-

pected findings should not be dismissed; instead they can help identify hidden assumptions, improve future experimental designs, and clarify the mechanisms that connect random assignment to causal inference. By documenting, scrutinizing, and reporting the processes linking treatment assignment to final estimates, researchers can design more informative survey experiments and strengthen the credibility of experimental survey research. Publishing carefully diagnosed “failures” is therefore an important part of building cumulative methodological knowledge alongside the evidence from “successful” experiments.

References

- Abadie, A. (2020). Statistical nonsignificance in empirical economics. *American Economic Review: Insights*, 2(2), 193–208. <https://doi.org/10.1257/aeri.20190252>.
- Chopra, F., Haaland, I., Roth, C., & Stegmann, A. (2024). The null result penalty. *The Economic Journal*, 134(657), 193–219. <https://doi.org/10.1093/ej/uead060>.
- Hensgen, S., & Zimmermann, F. (2026). Oops, they didn’t answer again! Dealing with systematic item nonresponse in a factorial survey experiment. *Survey Research Methods*. <https://doi.org/10.18148/srm/2026.v20i3.8545>.
- Hipp, L., & Leumann, S. (2026). Male Librarians Cause More Confusion than Female Engineers: Evidence on the Difficulties in Correctly Assessing the Gender Composition of Occupations from a Survey Experiment *Survey Research Methods*, 20(3). <https://doi.org/10.18148/srm/2026.v20i3.8579>
- Hofstetter, S., Wilke, A., & Yu, T. (2026). Performing below expectations: How to conduct survey experiments on identity threat. *Survey Research Methods* 20(3). <https://doi.org/10.18148/srm/2026.v20i3.8551>.
- Jungkunz, S., & Mayer, S. (2026). When happiness induction techniques make respondents unhappy: Failing to induce happiness in Ukraine during the Russo-Ukrainian war. *Survey Research Methods* 20(3). <https://doi.org/10.18148/srm/2026.v20i3.8503>.
- Kane, J.V. (2024). More than meets the ITT: A guide for anticipating and investigating nonsignificant results in survey experiments. *Journal of Experimental Political Science*. <https://doi.org/10.1017/XPS.2024.1>.
- Kunz, T., Gummer, T., & Neuert, C. (2026). Dealing with inconclusive results in survey research: A case study on assessing the comprehensibility of plain-language questions. *Survey Research Methods* 20(3). <https://doi.org/10.18148/srm/2026.v20i3.8543>.
- León, C.M., & Aizpurua, E. (2026). Hidden practices remain hidden? Lessons from a failed attempt to measure purchasing sexual services using the item count technique. *Survey Research Methods* 20(3). <https://doi.org/10.18148/srm/2026.v20i3.8532>.
- Litchfield, J., & Douarin, E. (2026). Unveiling the truth: Lessons from a (failed) list experiment on experiences of wartime sexual violence in Kosovo. *Survey Research Methods*, 20(3). <https://doi.org/10.18148/srm/2026.v20i3.8546>
- Lundberg, I., Johnson, R., & Stewart, B.M. (2021). What is your estimand? Defining the target quantity connects statistical evidence to theory. *American Sociological Review*, 86(3), 532–565. <https://doi.org/10.1177/00031224211004187>.
- Maslovskaya, O., Chang, G., Howe, S., Aizpurua, E., & Perelli-Harris, B. (2026). Effectiveness of using quick response (QR) codes in survey invitation letters: Experiment conducted in generations and gender survey in the UK. *Survey Research Methods*, 20(3). <https://doi.org/10.18148/srm/2026.v20i3.8522>.
- Mize, T.D., & Manago, B. (2022). The past, present, and future of experimental methods in the social sciences. *Social Science Research*, 108, 102799. <https://doi.org/10.1016/j.ssresearch.2022.102799>.
- Rauf, T., Voelkel, J.G., Druckman, J., & Freese, J. (2025). An audit of social science survey experiments. *Public Opinion Quarterly*, 89(4), 1063–1086.
- Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, 86(3), 638–641. <https://doi.org/10.1037/0033-2909.86.3.638>.
- Sajons, C., Müller, B., & Samoiluk, M. (2026). When statistical truth serum is not enough—lessons from a failed list experiment to measure corruption in Ukraine. *Survey Research Methods*, 20(3). <https://doi.org/10.18148/srm/2026.v20i3.8547>.
- Schachter, A., & Weisshaar, K. (2025). Survey experiments in sociology. *Annual Review of Sociology*, 51(1), 149–169. <https://doi.org/10.1146/annurev-soc-083024-070113>.
- Schwindt-Bayer, L., Elia, E., & Vera, S. (2026). Corruption, partisanship, and voting: Lessons from a survey experiment during the Brazilian presidential election. *Survey Research Methods*, 20(3). <https://doi.org/10.18148/srm/2026.v20i3.8548>.
- Theodoridis, A., & Goggin, S. (2026). Losing control (of the party): Conjectural bias in survey experiments. *Survey Research Methods* 20(3). <https://doi.org/10.18148/srm/2026.v20i3.8550>.
- Treischl, E., & Wolbring, T. (2022). The past, present and future of factorial survey experiments: A review for

- the social sciences. *Methods, Data, Analyses*, 16(2), 141–170.
- Waldendorf, A., Gërxhani, K., & Van De Rijt, A. (2026). The equalising potential of language: Can gender-inclusive language change behaviour? *Survey Research Methods*, 20(3). <https://doi.org/10.18148/srm/2026.v20i3.8554>.
- Zhang, N., & Baldassarri, D. (2026). Non-compliance in survey experiments with placebo treatments. *Survey Research Methods*, 20(3). <https://doi.org/10.18148/srm/2026.v20i3.8511>.