Fischer, Mira; Grewenig, Elisabeth; Lergetporer, Philipp; Werner, Katharina; Zeidler, Helen

Working Paper

The E-Word – On the Public Acceptance of Experiments

IZA Discussion Papers, No. 16511

Provided in Cooperation with:
IZA – Institute of Labor Economics

Suggested Citation: Fischer, Mira; Grewenig, Elisabeth; Lergetporer, Philipp; Werner, Katharina; Zeidler, Helen (2023) : The E-Word – On the Public Acceptance of Experiments, IZA Discussion Papers, No. 16511, Institute of Labor Economics (IZA), Bonn

This Version is available at:
https://hdl.handle.net/10419/282638

I Z A Institute
of Labor Economics
Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 16511

The E-Word – On the Public Acceptance of Experiments

Mira Fischer
Elisabeth Grewenig
Philipp Lergetporer
Katharina Werner
Helen Zeidler

OCTOBER 2023

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 16511

The E-Word – On the Public Acceptance of Experiments

Mira Fischer
WZB Berlin and IZA

Elisabeth Grewenig
Kreditanstalt für Wiederaufbau

Philipp Lergetporer
Technical University of Munich, CESifo and IZA

Katharina Werner
ifo Institute at the University of Munich

Helen Zeidler
Technical University of Munich

OCTOBER 2023

ISSN: 2365-9793

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

The E-Word – On the Public Acceptance of Experiments*

Randomized experiments are often viewed as the "gold standard" of scientific evidence, but people's scepticism towards experiments has compromised their viability in the past. We study preferences for experimental policy evaluations in a representative survey in Germany (N>1,900). We find that a majority of 75% supports the idea of small-scale evaluations of policies before enacting them at a large scale. Experimentally varying whether the evaluations are explicitly described as "experiments" has a precisely estimated overall zero effect on public support. Our results indicate political leeway for experimental policy evaluation, a practice that is still uncommon in Germany.

JEL Classification: I28, H40, C93
Keywords: experiment aversion, policy experimentation, education

Corresponding author:
Philipp Lergetporer
Technical University of Munich
TUM School of Management
Professorship of Economics
Bildungscampus 9
74706 Heilbronn
Germany
E-mail: philipp.lergetporer@tum.de

* We are most grateful to Ludger Woessmann for his support and advice, and to Franziska Kugler for her help in preparing the survey. Financial support by the Leibniz Competition (SAW-2014-ifo-2) and the German Science Foundation (CRC TRR 190) is gratefully acknowledged.

1. Introduction

Randomized experiments are often referred to as the "gold standard" of scientific evidence, and their use in natural contexts has markedly increased over the past two decades (Baldassarri and Abascal, 2017). Field experiments have transformed from being primarily small-scale proof-of-concept studies (Grose, 2014) into a broader tool for program evaluation in various fields, and further, into a comprehensive approach for governance and policymaking (Huitema, 2018). The experimental turn in policymaking is highlighted by the growing presence of government advisors with expertise in experimental social science and by the OECD's advocacy for policy experiments (OECD, 2019). Education is a particularly important field in which randomized controlled trials are proliferating and helping to improve policy (Sadoff, 2014).

At the same time, backlash from political decision-makers, bureaucrats, study participants, the public at large, and other stakeholders can compromise the viability of randomized field experiments (e.g., Heckman and Smith, 1995; Krueger, 1999; Heffetz and List, 2021). A case in point is Angrist and Lavy (2002), in which the authors report that an experiment offering cash incentives for students was suspended after "extensive and mostly critical media coverage" (p. 11). Other examples for experiments facing strong public condemnation after their implementation include the Facebook newsfeed experiment (Goel, 2014) or the matching score experiment of the dating platform OKCupid (Hern, 2014). While the public's acceptance of experiments is crucial for their feasibility, little systematic evidence exists on the extent and determinants of people's support for experimental policy evaluation.

We investigate the public's preferences for reform evaluation, and test the hypothesis that explicitly describing an evaluation as an "experiment" triggers public backlash. Negative reactions to the word "experiment" might be due to several reasons. It may make citizens think of past unethical or even criminal studies that have been referred to as "experiments" (e.g., the crimes against humanity committed by Nazi doctors during World War II)[1], it might also trigger concerns about policy uncertainty, as exemplified by the successful 1957 federal-election campaign slogan "*Keine Experimente*!" ("No Experiments!") of Germany's chancellor Konrad Adenauer.[2] Finally, the word "experiment" might prime reportedly unpopular features of field experiments – such as the use of randomization, denial of treatment to control-group members (e.g., Heckman and Smith, 1995), or lack of informed consent (e.g., List, 2008). Anecdotal evidence suggests that experimental economists

[1] These horrific crimes led to the creation of the Nuremberg Code of 1947, a code of research ethics for medical experimentation with human subjects (see also List, 2008).

[2] The slogan was used by the Christian Democratic Union (CDU) and referred to the risk that the Social Democratic Party (SPD) would leave the NATO in case of electoral victory.

often avoid the word "experiment" when communicating their research because they fear that using the word may yield backlash. If merely avoiding the word "experiment" can foster the political feasibility of field experiments, altered communication strategies could make conducting field experiments much easier and mitigate reluctance to use them.

We conduct a survey experiment among a representative sample of the German voting-age population (N>1,900) in which respondents are randomly assigned to one of two versions of a question on preferences for reform evaluation. Focusing on education policy, the baseline version of the question describes the evaluation process without explicitly mentioning the word "experiment". In the treatment group, we used the exact same wording as in the control group with the sole exception that the description of the evaluation process additionally includes the phrase "with experiments".

We find that a clear majority in the control group supports the idea of evaluating education reforms before rolling them out at a large scale: 75% are in favour of the proposal and only 14% oppose it (the remaining 11% are indifferent). Using the word "experiment" to describe reform evaluations has a precisely estimated zero causal effect on overall public support for education policy evaluation. In additional analyses, we show that treatment effects are homogeneous across sociodemographic subgroups, and that results are unlikely to be driven by respondents' inattention.

Our study contributes to the emerging experimental literature on preferences for (policy) experimentation among the public (e.g., Meyer et al., 2019; Mislavsky et al., 2019) and policymakers (e.g., Dur et al., 2023; Vivalt et al. 2023). We complement this strand of research by providing first evidence on how the use of the word "experiment" affects public preferences for experimental reform evaluation.

2. Experimental Setup

The experiment was embedded in the 2017 wave of the *ifo Education Survey*, an annual opinion survey on education policy in Germany, a country where experimental policy evaluation is rarely conducted (see Appendix B).[3] Our goal is to investigate whether using the word "experiment" to describe the evaluation of educational reforms affects public support for reform evaluation. Therefore, we randomly assigned respondents to one of two versions of a question that elicits public preferences for education reform evaluation. The control-group version of the question was worded as follows: "*Do you support or oppose that the effects of reforms in the education system, just like new medicine, should initially be tested on a small scale before they are implemented nationwide?*"

[3] The data from the ifo Education Survey are available for scientific use (Freundl et al., 2022). See Appendix B for details.

In contrast, the treatment-group question reads as follows: "*Do you support or oppose that the effects of reforms in the education system, just like new medicine, should initially be tested by experiments on a small scale before they are implemented nationwide?*" Note that the question wording is identical across experimental groups, with the sole exception being that in the treatment group the words "with experiments" were added. Respondents were asked to select one of the following five answer categories: strongly support, somewhat support, neither support nor oppose, somewhat oppose, strongly oppose (see Appendix Figure A1 for screenshots). Using a fair coin, we randomly assigned 949 (1,016) respondents to the treatment group (control group). Appendix Table A1 presents sample characteristics and shows that randomization worked as intended. We estimate treatment effects using simple OLS models (see Appendix B for the empirical model).

3. Results

Table 1 depicts our main results. Odd-numbered columns present estimates without controls, even-numbered columns include our set of sociodemographic control variables. A 75% majority of respondents in the control group supports the evaluation of education reform (see control mean), only 14% oppose it. The remainder is neutral. This widespread support echoes the majority backing for policy experimentation in other areas, for example, among Dutch voters as documented by Dur et al. (2023).

The small and statistically insignificant coefficients on the treatment indicator in Table 1 show that using the phrase "experiment" to describe reform evaluations does not affect average support for the evaluation of educational reforms. Note that the estimated effects are very small and that we are powered to detect treatment effects of 6 percentage points in columns 1 and 2, and 5 percentage points in columns 3 and 4.[4] In Appendix Table A2 we show that the treatment has small and insignificant effects on each of the five answer categories. Appendix Table A3 shows how support for reform evaluation with experiments varies with respondents' characteristics. In sum, the vast majority of Germans supports the evaluation of educational reforms, even if this evaluation is clearly labelled with the "E-word".

Treatment effects do not differ across sociodemographic subgroups defined by, for example, educational background or employment in education (Appendix Table A4). The only exception is political orientation: partisans of left parties are less likely to support reform evaluation if they are described as "experiment". Importantly, we find no heterogeneity by response time, indicating that the overall zero effect is not due to respondents' inattention.

[4] To compute minimal detectable effect sizes with 80% power and α=0.05, we follow Haushofer and Shapiro (2016) and multiply standard errors by 2.8.

4. Conclusion

In our representative survey experiment, we find majority support for scientific reform evaluation, irrespective of whether it is termed as an “experiment”. While there is widespread publication bias against null results (Chopra et al., 2023), we consider reporting this zero effect important as it provides the first causal evidence on whether using the “E-word” causes public backlash.

Our results shed light on the effect of the terminology used to describe scientific policy evaluation, but they do not allow for conclusions about how different elements of policy evaluations, like randomization, influence public attitudes. Studies examining aversion to randomization yield conflicting results (Meyer et al., 2019; Mislavski et al., 2019), which highlights the need for further studies. Additionally, replicating our experiment in policy areas beyond education would be valuable for future research to determine the generality of our findings.

References

Angrist, Joshua D, Victor Lavy (2002). The Effect of High School Matriculation Awards: Evidence From Randomized Trials. NBER Working Paper 9389.

Baldassarri, Delia, Maria Abascal (2017). Field Experiments across the Social Sciences. *Annual Review of Sociology* 43: 41-73.

Chopra, Felix, Ingar Haaland, Christopher Roth, Andreas Stegmann (2023). The Null Result Penalty. *Economic Journal*, forthcoming.

Dur, Robert, Arjan Non, Paul Prottung, Benedetta Ricci (2023). Who's Afraid of Policy Experiments? *Tinbergen Institute Discussion Papers* 23-027/V.

Freundl, Vera, Elisabeth Grewenig, Franziska Kugler, Philipp Lergetporer, Ruth Schüler, Katharina Wedel, Katharina Werner, Olivia Wirth, Ludger Woessmann (2022). The ifo Education Survey 2014-2021: A new Dataset on Public Preferences for Education Policy in Germany. *Journal of Economics and Statistics*, ahead of print.

Goel, Vindu (2014). Facebook Tinkers with Users' Emotions in News Feed Experiment, Stirring Outcry, New York Times 29/07/2014.

Grose, Christian R. (2014). Field Experimental Work on Political Institutions. *Annual Review of Political Science* 17: 355-370.

Haushofer, Johannes, Jeremy Shapiro (2016). The Short-term Impact of Unconditional Cash Transfers to the Poor: Experimental Evidence from Kenya. *Quarterly Journal of Economics* 131: 1973–2042.

Heckman, James J., Jeffrey A. Smith (1995). Assessing the Case for Social Experiments. *Journal of Economic Perspectives* 9: 85–110.

Heffetz, Ori, John List (2021). Who's Afraid of Evidence-Based Policymaking? *Project Syndicate*.

Hern, Alex (2014). OKCupid: We Experiment on Users. Everyone Does, The Guardian 07/24/2014, https://www.theguardian.com/technology/2014/jul/29/okcupid-experiment-human-beings-dating, [accessed 27 September 2023].

Huitema, Dave, Andrew Jordan, Stefania Munaretto, Mikael Hildén (2018). Policy Experimentation: Core Concepts, Political Dynamics, Governance and Impacts. *Policy Sciences* 51(2): 143-159.

List, John A. (2008). Informed Consent in Social Science. *Science* 322(5886): 672.

List, John A., Robert Metcalfe (2014). Field experiments in the Developed World: An Introduction. *Oxford Review of Economic Policy* 30(4): 585–596.

Meyer, Michelle N., Patrick R. Heck, Geoffrey S. Holtzman, Stephen M. Anderson, William Cai, Duncan J. Watts, Christopher F. Chabris (2019). Objecting to Experiments that Compare Two Unobjectionable Policies or Treatments. *PNAS* 116(22): 10723-10728.

Mislavsky, Robert, Berkeley Dietvorst, Uri Simonsohn (2019). The Minimum Mean Paradox: A Mechanical Explanation for Apparent Experiment Aversion. *PNAS* 116(48): 23883-23884.

OECD (2019). Tools and Ethics for Applied Behavioural Insights: The BASIC Toolkit, Paris: OECD Publishing.

Sadoff, Sally (2014). The Role of Experimentation in Education Policy, *Oxford Review of Economic Policy* 30(4): 597-620.

Vivalt, Eva, Aidan Coville, Sampada KC (2023). Weighing the Evidence: Which Studies Count? *Working Paper*.

Table 1: Effects of using the word "experiment" on preferences for reform evaluation

	Support for reform evaluation		Opposition reform evaluation	
	(1)	(2)	(3)	(4)
„Experiment" treatment	-0.009	-0.015	0.019	0.021
	(0.022)	*(0.023)*	*(0.018)*	*(0.018)*
Covariates	No	Yes	No	Yes
Control mean	0.751		0.140	
Observations	1,957	1,902	1,957	1,902
R^2	0.000	0.022	0.001	0.024

Notes: OLS regressions. "Experiment" treatment: 1 = word "experiment" is included in the question text, 0 otherwise. Dependent variable: Columns 1-2: Dummy variables 1 = "strongly support" or "somewhat support" reform evaluation, 0 otherwise; columns 3-4: Dummy variables 1 = "strongly oppose" or "somewhat oppose" reform evaluation, 0 otherwise. Residual category: "neither support nor oppose." Control mean: mean of the outcome variable in the control group. Data source: ifo Education Survey 2017. Regressions weighted by survey weights. Robust standard errors in parentheses. Significance levels: *** p<0.01, ** p<0.05, * p<0.10.

Online Appendix A: Additional Tables and Figures

Appendix Table A1: Summary statistics and balancing table

	Mean [SD]	Treatment effects on covariates
	(1)	(2)
Age	50.1 [18.5]	0.001 *(0.001)*
Female	0.505	0.031 *(0.026)*
Born in Germany	0.940	0.092 *(0.059)*
Municipality size	4.31 [1.8]	-0.004 *(0.007)*
Monthly household income	2324.1 [1471.8]	0.000 *(0.000)*
Partner in household	0.552	0.020 *(0.027)*
Has parent(s) with univ. degree	0.292	-0.018 *(0.028)*
Works in education sector	0.079	0.065 *(0.046)*
Parents of school-aged children	0.258	0.000 *(0.029)*
Lives in West Germany	0.805	0.071** *(0.030)*
No or basic school degree	0.366	0.079*** *(0.029)*
Middle school degree	0.306	-0.058** *(0.026)*
University entrance degree	0.328	-0.028 *(0.027)*
University student	0.097	-0.054 *(0.045)*
Employed	0.517	-0.012 *(0.026)*
Unemployed/Retired	0.386	0.033 *(0.028)*
Left-leaning political preferences	0.450	-0.038 *(0.032)*
Risk tolerance	4.2 [2.5]	-0.002 *(0.005)*
Patience	6.0 [2.5]	-0.005 *(0.005)*
Offline-survey mode	0.169	0.038 *(0.043)*
F-Test for joint significance (p-value)		0.4142

Notes: Column 1: Weighted group means (standard deviations of non-binary variables in brackets). Column 2: coefficients and standard errors of regressions of the respective covariate on the treatment indicator. Each cell represents a separate regression. Data source: ifo Education Survey 2017. Regressions weighted using survey weights. Significance levels: *** p<0.01, ** p<0.05, * p<0.10.

Description of Appendix Table A1: We test whether observable characteristics of our respondents can predict assignment into experimental groups in Appendix Table A1. Column 1 reports the covariate means and standard deviations in brackets (for non-binary covariates) for the analysis sample. Column 2 reports coefficients from regressing each covariate on the treatment indicator. Overall, the table shows that there are small but significant differences ($p<0.05$) in only 3 out of 20 pairwise comparisons. In addition, regressing treatment status simultaneously on all listed covariates yields a p-value for joint significance of 0.414 (bottom part of Appendix Table A1). Thus, our randomization worked as intended.

Appendix Table A2: Effects of using the word "experiment" on preferences for reform evaluation: Five answer categories

	Strongly support	Somewhat support	Neither support nor oppose	Somewhat oppose	Strongly oppose
	(1)	(2)		(3)	(4)
„Experiment" treatment	-0.012	-0.002	-0.007	0.007	0.015*
	(0.022)	*(0.027)*	*(0.016)*	*(0.017)*	*(0.008)*
Covariates	Yes	Yes	Yes	Yes	Yes
Control mean	0.222	0.529	0.110	0.117	0.022
Observations			1,902		
R^2	0.019	0.011	0.021	0.017	0.019

Notes: OLS regressions. "Experiment" treatment: 1 = word "experiment" is included in the question text, 0 otherwise. Dependent variables: Dummy variables 1 = respondent selected respective answer category, 0 otherwise. Control mean: mean of the outcome variable in the control group. Covariates include age, municipality size, income, risk tolerance, patience, and dummies for gender, born in Germany, living with partner in household, parents' higher degree, working in the education sector, parent status, living in West Germany, highest school degree, employment status, and offline-survey mode. Data source: ifo Education Survey 2017. Regressions weighted by survey weights. Robust standard errors in parentheses. Significance levels: *** p<0.01, ** p<0.05, * p<0.10.

Description of Appendix Table A3: The table replicates Table 1 using each of the five answer categories as dependent variables. For each answer category, we find small and insignificant effects of using the word "experiment", which shows that aggregating answer categories as in Table 1 does not obfuscate treatment effects in individual answer categories.

Appendix Table A3: Who supports experimental reform evaluations?

	Dependent variable: Support for reform evaluation with experiments
	(1)
Age	-0.002 *(0.001)*
Female	-0.070** *(0.033)*
Born in Germany	0.018 *(0.071)*
Municipality size	0.003 *(0.009)*
Monthly household income	-0.000 *(0.000)*
Partner in household	0.007 *(0.039)*
Has parent(s) with univ. degree	-0.030 *(0.039)*
Works in education sector	0.047 *(0.055)*
Parents of school-aged children	-0.016 *(0.037)*
Lives in West Germany	0.058 *(0.045)*
Middle school degree	-0.023 *(0.043)*
University entrance degree	0.024 *(0.052)*
University student	-0.026 *(0.072)*
Unemployed/Retired	0.033 *(0.040)*
Risk tolerance	-0.002 *(0.007)*
Patience	0.014** *(0.007)*
Offline-survey mode	0.090 *(0.064)*
Constant	0.708*** *(0.131)*
Observations	917
R^2	0.0220

Notes: OLS regressions. Treatment group only. Dependent variable: Dummy variable 1 = "strongly support" or "somewhat support" reform evaluation with experiments, 0 otherwise. Data source: ifo Education Survey 2017. Regressions weighted by survey weights. Robust standard errors in parentheses. Significance levels: *** p<0.01, ** p<0.05, * p<0.10.

Description of Appendix Table A2: The table shows that female respondents are less likely to support education reform evaluations with experiments, and more patient respondents are more likely to support them. The observation that all other coefficients in the table are insignificant indicates that high support for reform evaluation with experiments is a general phenomenon across most subgroups of the German population.

Appendix Table A4: Treatment effects in sociodemographic subgroups

	Dependent variable: Support for reform evaluation
Treatment effects for the following subgroups:	(1)
No or basic school degree	0.006
	(0.042)
Middle school degree	-0.061*
	(0.036)
University entrance degree	0.014
	(0.037)
Income above median	-0.014
	(0.030)
Well-informed about educ. system	0.007
	(0.035)
Positive evaluation of educ. system	0.016
	(0.033)
Works in education sector	-0.038
	(0.071)
Response time above median	-0.026
	(0.035)
Offline-survey mode	-0.046
	(0.061)
Political leaning: right	0.058
	(0.039)
Political leaning: left	-0.108***
	(0.041)
Political leaning: progressive	-0.012
	(0.072)
Non-partisans	0.003
	(0.044)

Notes: Each line represents the coefficient of a separate OLS regression. Dependent variable: Dummy variable 1 = "strongly support" or "somewhat support" reform evaluation, 0 otherwise. The table displays coefficients on the interaction term between treatment and subgroup indicators from estimates based on equation (2) in Appendix B. Data source: ifo Education Survey 2017. Regressions weighted by survey weights. Robust standard errors in parentheses. Significance levels: *** p<0.01, ** p<0.05, * p<0.10.

Description of Appendix Table A4: The table investigates the extent to which treatment effects differ across sociodemographic subgroups using the regression framework of equation (2) in Online Appendix B. The fact that most coefficients reported in the table are insignificant shows that almost no subgroup of the German population reacts adversely to the word "experiment". In particular, we find no heterogeneous treatment effects by educational attainment, income, respondents' information status about the German education system, and whether they think that the

German school system performs well.[5] Furthermore, we find no significant treatment effects for two groups that would be directly affected by an (experimental) evaluation process: Parents of children below age 18 years, and respondents working in the education sector. Interestingly, we also find no effect heterogeneity by proxies of respondents' attention: Respondents with longer response times and those interviewed offline in the presence of an interviewer exhibit no differential treatment effects than their counterparts. This suggests that inattention cannot explain why we do not find an effect of using the word "experiment".

The only significant treatment effect heterogeneity that we detect is by respondents' political leaning. Grouping partisans of the six major German parties[6] into *right* (partisans of the CDU/CSU, and the AfD), *left* (partisans of the SPD and Die Linke), *progressives* (partisans of the FDP and Die Grünen), and *non-partisans*, we find that using the word "experiment" to describe the evaluation process makes *leftists* significantly less likely (by 11 percentage points) to support reform evaluation. Support by those categorized as right, progressives, and nonpartisans stays unchanged. The finding that treatment effects are homogeneous by sociodemographic background suggests that the sociodemographic composition of different partisan groups cannot account for the significant effect heterogeneities by political leaning. At the same time, the significant coefficients in the table need to be interpreted with some caution given the large number of hypotheses tested in the table, and the related risk of false-positive results.

[5] We construct an information measure by using respondents' answers to several guess questions on facts about the educational system. A respondent is classified as "informed" if her beliefs are closer to the correct values than those of the median respondent. To categorize respondents' beliefs about the performance of the school system, we assume respondents have a positive evaluation of the school system if they say they would give schools in their local area one of the top two grades on a 6-point scale.

[6] The categorization is based on the following question about the respondents' long-term party attachment: "Many people in Germany lean towards a particular political party in the long term, even if they occasionally also vote for another party. With which party do you sympathize in general?"

Appendix Figure A1: Screenshots of the survey questions

"Experiment" treatment

Control group version

Notes: Screenshots of the two versions of the question. The wording of the question in the two treatments only differs by whether it includes the worlds "mit Experimenten" ("with experiments").

Description of Appendix Figure A1: The figure provides screenshots of the survey questions as they appeared on respondents' devices. To prompt people to give a considered answer and to minimize the error of central tendency, the category "neither favor nor oppose" was placed below the

other answer categories for both questions. We implemented a methodological experiment on another survey question (on granting teachers civil service protections) and found that the position of the neutral category does not change relative support and opposition towards the policy proposal (results available upon request).

Online Appendix B: Additional Information

The Opinion Survey. Our paper is based on data from the 2017 wave of the ifo Education Survey, an annual representative opinion survey on education policy in Germany. The survey comprised a total of 3,968 respondents, and our experiment was conducted among a randomly chosen subsample of 1,965 respondents.[7] The respondents not included in our analysis answered unrelated questions about education spending or the PISA test instead of answering a question on (experimental) policy evaluation. Overall, the survey contained 34 questions on different topics of education policy and also collected information on respondents' sociodemographic characteristics. Median completion time was 17 minutes, and item-non-response was very small, for instance below 0.3 percent for our main outcome question of interest. Treatment status does not predict item non-response on the outcome variables (results available upon request). Sampling and polling was carried out by Kantar Public, a renowned survey company, in April and May 2017.

Survey representativeness is an important feature of our study which enables us to derive generalizable statements for the political economy of policy evaluation. Since computerized surveys do not cover the part of the population that does not use the internet, Kantar Public collected the data in two strata. First, people who use the internet (83 percent) were drawn from an online panel and answered all questions autonomously on their devices. Second, people who reported not to use the internet (17 percent) were surveyed at their homes by trained interviewers. These respondents were provided with a tablet computer for completing the survey. This mixed-mode design assures that our findings are representative for the entire German population.

All analyses presented in this paper use survey weights that were designed to match official statistics with respect to age, gender, parental status, school degree, federal state, and municipality size. Grewenig et al. (2023) show that using survey weights achieves representativeness of the *ifo Education Survey* data for the German adult population.

Empirical Model. We estimate the causal effects of using the word "experiment" on support for education-reform evaluation with the following regression model:

$$y_i = \alpha_0 + \alpha_1 T_i + \delta' X_i + \varepsilon_i \qquad (1)$$

where y_i is respondent i's preference for educational reform evaluation, T_i indicates whether respondent i received the version of the question contains the word "experiment", X_i is a vector of

[7] Eight respondents did not provide an answer to the question on preferences for reform evaluation, reducing our analytical sample to 1,957 respondents.

control variables, and ε_i is an error term which is uncorrelated with all right-hand side variables. The parameter of interest α_1 represents the causal effect of using the words "with experiments". While further control variables are not required to identify the causal treatment effect because of random assignment, we include further controls in some specifications to increase the precision of our estimates, and to account for the slight imbalances reported in Appendix Table A1. Our main outcomes of interest are dummy variables coded 1 if a respondent (strongly or somewhat) supports or opposes reform evaluations, and 0 else, but we also analyze effects on each of the five answer categories separately to investigate preference intensity.

To analyze heterogeneous treatment effects by respondents' background characteristics, we additionally employ the following regression model:

$$y_i = \beta_0 + \beta_1 T_i + \beta_2 Subgroup_i + \beta_3 T_i Subgroup_i + \theta' X_i + \eta_i \quad (2)$$

where $Subgroup_i$ equals 1 if respondent i belongs to the respective subgroup, and 0 otherwise. In this specification, β_1 measures the treatment effect on non-members of the subgroup, and β_3 measures the additional effect on the subgroup. We use linear probability models throughout the paper. (Ordered) probit models lead to the same qualitative results (available upon request). Covariates include age, municipality size, income, risk tolerance, patience, and dummies for gender, born in Germany, living with partner in household, parents' higher degree, working in the education sector, parent status, living in West Germany, highest school degree, employment status, and offline-survey mode.

Experimental Policy Evaluation in Germany. Germany's political institutions may offer particularly good conditions for experimentation and learning to address a variety of social and economic problems because of their high degree of decentralization (Oates, 1999), including in the area of general education. However, compared to, for instance, the Finnish government that declared that it wants Finland to become "the world's best environment for innovating and experimenting by 2025" (http://julkaisut.valtioneuvosto.fi/handle/10024/161308 [accessed 27 September 2023]) and several other developed countries (World Bank Group 2018), German politicians have been rather reluctant to embrace experimental policy evaluation.

In 2015, a unit named "Wirksam Regieren" ("Governing Effectively"), subject to the Chancellery, took up work. Its declared aim is the use of "ex-ante-effectiveness analyses to gain empirical insights for the evaluation of alternative problem-solving approaches and to increase the effectiveness of policy measures" to which end it is supposed to run "pilot-projects" (Deutscher

Bundestag 2015). However, to date the unit's website mainly lists projects that are survey studies and survey experiments. Merely three of the listed projects (campaigns for improved hygiene in hospitals and measles vaccinations) are policy evaluations and aim at impacting objective outcomes. Beyond the absence of randomized evaluations, limited access to pertinent data for researchers has been criticized as a factor undermining effective policy evaluations in Germany (see, e.g., Riphahn et al., 2016; Blesse et al., 2023).

Appendix References

Blesse, Sebastian, Philipp Lergetporer, Justus Nover, Katharina Werner (2023). Transparency and Policy Competition: Experimental Evidence from German Citizens and Politicians. CESifo Working Paper 10292.

Deutscher Bundestag (2015). Schriftliche Fragen mit den in der Woche vom 4. Mai 2015 eingegangenen Antworten der Bundesregierung: Ziel der Arbeitsgruppe "wirksames Regieren" sowie Aufgaben der drei im Bundeskanzleramt eingestellten Experten und neutrale Aufklärung der Bürger", Drucksache 18/4856, Berlin: Deutscher Bundestag, https://dipbt.bundestag.de/extrakt/ba/WP18/672/67298.html [accessed 27 September 2023].

Grewenig, Elisabeth, Philipp Lergetporer, Lisa Simon, Katharina Werner, Ludger Woessmann (2018). Can Internet Surveys Represent the Entire Population? A Practitioners' Analysis. *European Journal of Political Economy* 78, 102382.

Oates, Wallace E. (1999). An Essay on Fiscal Federalism. *Journal of Economic Literature* 37 (3) (1999): 1120-1149.

Riphahn, Regina T., Ludger Woessmann (2016). Mehr Transparenz in der Bildungspolitik. *Wirtschaftsdienst* 96 (7): 474-478.

World Bank Group (2018). Behavioral Science around the World: Profiles of 10 Countries, Brief 132610, Washington D.C.: World Bank Group http://documents.worldbank.org/curated/en/710771543609067500/pdf/132610-REVISED-00-COUNTRY-PROFILES-dig.pdf [accessed 27 September 2023].