The E-Word – On the Public Acceptance of Experiments

Mira Fischer (WZB Berlin)
Elisabeth Grewenig (ifo Institute)
Philipp Lergetporer (ifo Institute)
Katharina Werner (ifo Institute)

Discussion Paper No. 219

December 16, 2019

Collaborative Research Center Transregio 190 | www.rationality-and-competition.de
Ludwig-Maximilians-Universität München | Humboldt-Universität zu Berlin
Spokesperson: Prof. Dr. Klaus M. Schmidt, University of Munich, 80539 Munich, Germany
+49 (89) 2180 3405 | info@rationality-and-competition.de

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

Mira Fischer, Elisabeth Grewenig, Philipp Lergetporer, and Katharina Werner †

Abstract

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.

Keywords: experiment aversion, policy experimentation, education

JEL classification: I28, H40, C93

December 15, 2019

.

* 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 through CRC TRR 190 (project number 280092119) is gratefully acknowledged.

† Fischer: WZB Berlin; mira.fischer@wzb.eu; Grewenig: ifo Institute at the University of Munich; grewenig@ifo.de; Lergetporer: ifo Institute at the University of Munich; CESifo; lergetporer@ifo.de. Werner: ifo Institute at the University of Munich; werner.k@ifo.de.

1. Introduction

Randomized experiments are often referred to as the "gold standard" of scientific evidence (e.g., Abadie and Cattaneo 2018) and their implementation in naturally occurring contexts has been greatly increasing over the last two decades (Baldassarri and Abascal 2017). They have evolved from mostly small-scale proof-of-concept studies (Lupia 2002, Grose 2014), to experimentation as program evaluation in the field, to experimentation as an approach to governing, including but not limited to public policy (Huitema 2018). The experimental turn in policy-making is evidenced by the increasing commonness of government advisors with a background in experimental social science[1] and the OECD's promotion of policy experiments (OECD 2019). Education is a particularly important field in which randomized controlled trials are proliferating and helping to improve policy (Sadoff 2014, Connolly 2018).

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, Angrist and Lavy 2009). A case in point is the paper by Angrist and Lavy (2002), which reports that an experiment offering cash incentives for students was suspended after "extensive and mostly critical media coverage" (p. 11).[2] 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 field experiments.

In this paper, we investigate the public's preferences for reform evaluation and test the hypothesis that explicitly describing an evaluation as an "experiment" triggers public backlash. Such negative reactions might be due to several reasons. The word "experiment" 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)[3], it might also trigger concerns about policy uncertainty, as exemplified by the successful 1957 federal-election campaign slogan "*Keine Experimente*!" ("No

[1] These advisors often form so called "behavioral insights" teams that tend to focus on marketing-inspired experiments around social influence and less on program or policy evaluation. Behavioral insights teams started proliferating ten years ago, and especially since 2015, in, for example, Australia, Canada, Denmark, the Netherlands, the U.K., and the U.S. (World Bank Group 2018).

[2] More recent examples for experiments facing strong public condemnation after their implementation include the Facebook newsfeed experiment (Kramer et al. 2014, Goel 2014) or the matching score experiment of the dating platform OKCupid (Hern 2014, Hawver 2014), although the sites only implemented what is wide-spread practice on the internet (Christian 2012).

[3] 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). Another example for harmful experiments are the medical and radiation studies conducted in the 1940s to 1960s in the United States (Conahan 1994).

Experiments!") of Germany's chancellor Konrad Adenauer.[4] Finally, the word "experiment" might prime reportedly unpopular features of field experiments – such 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 shows that experimental economists often avoid the word "experiment" when communicating their research because they fear that using the word may yield a 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 some governments' reluctance to use them.[5]

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 evaluation process is described as "with experiments". If the German public takes an instant dislike against the word "experiment", then support for evaluating educational reforms should be substantially smaller in the treatment group where "experiments" are mentioned.

We have two main findings. First, a clear majority of the German public supports the idea of evaluating education reforms before rolling them out at a large scale: 75 percent are in favour of the proposal and only 14 percent oppose it (the remaining 11 percent are indifferent). This widespread support does not vary significantly across sociodemographic subgroups, with the exception of females (more patient respondents) who are significantly less (more) supportive of reform evaluations.

Second, using the word "experiment" to describe reform evaluations has a precisely estimated zero causal effect on overall public support for education policy evaluation. Treatment effects are very small (1 to 2 percentage points, depending on the specification) and statistically insignificant. Given our relatively large sample size, our ex-post minimum

[4] 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.

[5] While feared public backlash is a likely reason for why governments are sometimes reluctant to adopt field experiments for policy evaluation, a complementary reason highlighted in the political-economy literature is that politicians often do not have an interest in finding out whether their policies have the intended effects (Campbell 1969).

detectable effect size (MDE) is 5 to 6 percentage points, which leaves us well-powered to rule out any effects for the overall German population.

Explorative subgroup analyses do not reveal heterogeneities in the treatment effect for different sociodemographic subgroups (defined along the lines of e.g. education, income, or parental status). However, they suggest heterogeneities along political party affiliation as left-wing supporters are significantly less likely to support reform evaluation when the word "experiment" is used to describe the evaluation process, whereas respondents with other political orientations are unaffected by the treatment.

We do not find treatment effect heterogeneities by response time or survey mode, which suggests that effects are unlikely due to inattention. In sum, our results reveal broad support for the idea to evaluate education reforms before enacting them. This support is generally unaffected by framing this practice as "experiment", however different political parties possibly face different political constraints when it comes to communicating policy experimentation. Contrary to anecdotal evidence, our results suggest that the German population is not in general averse to the use of the word "experiment" when referring to the scientific evaluation of education policies. While our results speak to the language conventions around scientific policy evaluation, no conclusions can be drawn from them about public attitudes towards the use of different methods, in particular randomization, in policy evaluation.

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"[6] and several other developed countries (World Bank Group 2018), German politicians have been rather reluctant to embrace experimental policy evaluation. In 2015, a three-person 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[7] mainly lists projects that

[6] http://julkaisut.valtioneuvosto.fi/handle/10024/161308 [accessed 13 December 2019]

[7] https://www.bundesregierung.de/breg-de/themen/wirksam-regieren/ [accessed 13 December 2019]

are survey studies and survey experiments. Merely two of the listed projects (campaigns for improved hygiene in hospitals and measles vaccinations) are policy evaluations and aim at impacting objective outcomes.

Our study contributes to the emerging literature on experiment aversion and may inform the theoretical literature on the political economy of policy experimentation. While theoretical contributions have considered politicians' incentives for measuring the impact of their policies, studies on experiment aversion have investigated people's attitudes towards experimentation.

Callander and Hummel (2014) study a setting in which there is uncertainty about the outcomes of different policies. Here, small-scale policy experiments may help policy-makers to learn where other desirable policies may lie as well as which policies should be avoided. Milner, Ollivier, and Simon (2014) consider a setting in which political parties differ both in preference parameters and in empirical beliefs about the consequences of different policies. Their setting gives rise to an incentive for the incumbent party to experiment because experimentation causes both parties to update their beliefs. This in turn influences both the incumbent and the opponent party's future policy choices. Mukand and Rodrik (2005) consider a setting in which countries differ by their local state of the world, e.g. in terms of historical trajectories, institutional settings, social norms and geographical givens. Experimentation helps a country to find the combination of policies that produces the best results for its setting but is also costly. Imitating another country avoids the costs of experimentation but may lead a country to adopt a policy that is not appropriate for its context if it is not similar enough to the country it tries to imitate.

Meyer et al. (2019) conduct a series of online between-subject vignette experiments with non-representative samples across different policy domains and find evidence that people rate a randomized experiment comparing two unobjectionable policies or treatments, neither of which was known to be superior, as less appropriate than simply implementing either option for everyone.[8] They investigate several explanations for the effect and find that people tend to believe that consent is required to impose a policy on half of a population but not on the entire population, tend to have an aversion to controlled but not to uncontrolled

[8] The study's conclusion is questioned by Mislavsky, Dietvorst, and Simonsohn (2019b) who argue that the effect arises due to its between-subjects design. They argue that in its setting lower support for the experiment may be fully accounted for by the fact that all people who find one of the options objectionable and the other acceptable will reject the experiment, while in the two treatments in which people are each presented with only one option, they will accept or reject this option independently of what they might think about the other option. This, Dietvorst et al. argue, amounts to comparing mean to minimum (instead of mean to mean) acceptance across treatments.

experiments and tend to fall prey to the illusion of knowledge, independently of their level of education or science literacy. Mislavsky, Dietvorst, and Simonsohn (2019a) conducted several online within-subject and pen-and-pencil vignette experiments with non-representative samples in which participants evaluated the acceptability of either corporate policy changes or of experiments testing them. When all policy changes were deemed positive (i.e better than the status quo), subjects found the experiment acceptable even when it involved deception, unequal outcomes, and lack of consent. When one of the two policy changes was negative, the experiment was rated no less acceptable than the negative policy change. In contrast to the results by Meyer et al. (2019), experiments were rated no less acceptable than the simple average acceptability of the two policy changes involved in it if both policy changes were positive. However, experiments were rated less acceptable than the simple average acceptability of the two policy changes if one of them was negative. The authors conclude that they find no evidence for experiment aversion but merely for people's tendency to overweigh negative attributes.

Finally, this study relates to the growing economics literature that uses survey experiments to study determinants of the public's policy preferences (e.g., Cruces et al. 2013, Kuziemko et al. 2015, Haaland and Roth 2017, Alesina et al. 2018a, Lergetporer et al. 2018, Roth et al. 2018). We extend this literature by applying the methodology of survey experiments to study public preferences for experimentation.

Our study is the first to investigate experiment aversion with a clean and simple manipulation in a large and representative sample of the German population. This gives its findings strong external and internal validity and statistical power, and, consequently, allows us to derive generalizable conclusions about the effect of communication on the feasibility of scientific policy evaluation.

The remainder of the paper is structured as follows. Section 2 describes the opinion survey and the experimental design. Section 3 presents our main results and analyzes effect heterogeneities. Section 4 concludes.

2. Data, Experimental Setup, and Empirical Model

2.1 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 4,081 respondents, and our experiment was conducted among a randomly chosen

subsample of 2,225 respondents.[9] Overall, the survey contained 34 questions on different topics of education policy and also collected information on respondents' sociodemographic characteristics (see Table 1). 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.[10] Sampling and polling was carried out by Kantar Public, a renowned survey company, in April and May 2017.

While rare in experimental analyses, 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.

2.2 The Survey Experiment

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?" In contrast, the treatment-group question read as follows: "Do you support or oppose that the effects of reforms in the education system, just like new medicine, should initially be tested with 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.

[9] 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 evaluations.

[10] Treatment status does not predict item non-response on the outcome variables (results available upon request).

Respondents were asked to select one of the following five answer categories: strongly support, somewhat support, neither support nor oppose, somewhat oppose, strongly oppose.[11] Similar to other recent survey-based economics papers, our outcome of interest is a self-reported preference (e.g., Kuziemko et al. 2015, Falk et al. 2018). Reassuringly, recent evidence shows that survey-based measures correspond closely to actual political behaviour such as signing petitions or donating to charity (e.g., Haaland and Roth 2017, Alesina et al. 2018b).

We test whether observable characteristics of our respondents can predict assignment into experimental groups in Table 1. Column 1 reports the covariate means and standard deviations in brackets (for non-binary covariates) for the control-group version of the question. 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 0.411 (bottom part of Table 1). Thus, our randomization worked as intended.

As insignificant coefficient estimates can either reflect true null effects or just a lack of statistical power, we report minimum detectable effect sizes (MDEs) in the results section. To compute MDEs with 80% power and $\alpha=0.05$, we follow Haushofer and Shapiro (2016) and multiply standard errors by 2.8.

2.3 Empirical Model

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

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

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 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

[11] Appendix Figure A1 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).

"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 Table 1. 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.

3. Results

This section first presents our main results on support for reform evaluation and the causal effect of using the word "experiment" to describe the evaluation, and then presents a heterogeneity analysis across sociodemographic subgroups.

3.1 Main Results

Table 2 depicts the results from regressions based on equation (1). Odd-numbered columns present estimates without controls, even-numbered columns include our set of sociodemographic control variables.[12] A broad majority of respondents in the control-group version of the question (i.e., without mentioning the word "experiment") of 75 percent supports the evaluation of education reform (see control mean). Only a small minority of 14 percent opposes it. The remainder neither supports nor opposes it. In Appendix Table A1 we regress support for reform evaluation on respondents' sociodemographic characteristics. Female respondents are less likely to support education reform evaluations, and more patient respondents are more likely to support them. The observation that all other coefficients in the

[12] We use linear probability models throughout the paper. (Ordered) probit models lead to the same qualitative results (available upon request). The controls are all variables listed in Table 1 with the exception of political preference; see notes to Table 2 for details.

table are insignificant indicates that high support for reform evaluation is a general phenomenon across most subgroups of the German population.

The small and statistically insignificant coefficients on the treatment indicator show that using the phrase "experiment" to describe reform evaluations does not affect average support for - or average opposition against - the evaluation of educational reforms. Note that the estimated effects are very small and 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. Note that our MDEs would allow us to detect any majority shift in support for or opposition against the evaluation of educational reforms induced by the word "experiment". Our heterogeneity analysis by response time below indicates that this precisely estimated zero effect is not due to respondents' inattention. In sum, the vast majority of Germans supports the evaluation of educational reforms, even if this evaluation is clearly labelled with the "E-word".

Appendix Table A2 further investigates treatment effects on each of the five answer categories separately. For each answer categories, we find small and insignificant effects of using the word "experiment", which shows that aggregating answer categories as in Table 2 does not obfuscate treatment effects in individual answer categories.

3.2 Effect Heterogeneities Across Socioeconomic Subgroups

Next, we investigate the extent to which treatment effects differ across sociodemographic subgroups using the regression framework of equation (2). For each sociodemographic subgroup, Table 3 depicts treatment effects for its members.

The fact that most coefficients reported in Table 3 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.[13] 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

[13] 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.

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 6 major German parties[14] into c*onservatives* (partisans of the CDU/CSU, and the AFD), *socialists* (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 *socialists* significantly less likely (by 11 percentage points) to support reform evaluation. Support by conservatives, progressives, and non-partisans 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 Table 3 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.

4. Conclusion

We conducted a randomized survey experiment within a representative sample of more than 1,900 respondents in Germany to investigate public support for (experimental) reform evaluation. To study the extent to which using the word "experiment" yields public backlash, we randomly assigned respondents to one of two versions of our question of interest: The control-group version elicited preferences for reform evaluation without using the word "experiment", whereas the treatment-group version explicitly mentioned "with experiments" in the question. We find that public support for policy reform evaluation is generally high (75 percent in favour) and is overall unaffected by using the word "experiment". The treatment effects are very small (1-2 percentage points, depending on the specification), insignificant, and precisely estimated. We consider reporting these zero effects important since it is the first causal evidence on whether experimenters' common practice to avoid the word "experiment" affects the public's preferences, not least in light of widespread publication biases against null results (e.g., Franco et al. 2014). Further analyses reveal that left-leaning respondents are sensitive to the treatment variation. Their support for reform evaluation is significantly

[14] 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?"

reduced when the words “through experiments” are added. Contrary to anecdotal evidence, we do not find a general aversion against the word “experiment” but our results also suggest that different political parties may face different constraints.

Our results should, however, not be read as evidence in favour of or against public support for the randomized evaluation of policies. Policy experimentation may apply various methods, with randomized controlled trials (RCTs) being just one of them. Recent studies testing for aversion against randomization in non-representative samples have reached conflicting conclusions (Meyer et al. 2019, Mislavski, Dietvorst, and Simonsohn 2019a) and further evidence is needed. The word “experiment” is often used to describe trial-and-error strategies (e.g., Batory et al. 2018), studies in which subjects are not randomly assigned to treatment and control groups (List and Metcalfe 2014) or even qualitative case studies (Blanchenay and Burns 2016). While our paper is agnostic about the exact experimental method that citizens think of when being confronted with the word “experiment”, we do think that this is an interesting avenue for future research.

References

Abadie, Alberto, Matias D. Cattaneo (2018). Econometric Methods for Program Evaluation. *Annual Review of Economics* 10: 465-503.

Alesina, Alberto, Stefanie Stantcheva, Edoardo Teso (2018a). Intergenerational Mobility and Support for Redistribution. *American Economic Review* 108 (2): 521-554.

Alesina, Alberto, Armando Miano, Stefanie Stantcheva (2018b). Immigration and Redistribution. *NBER Working Paper* 24733. Cambridge, MA: National Bureau of Economic Research.

Angrist, Joshua D, Victor Lavy (2002). The Effect of High School Matriculation Awards: Evidence From Randomized Trials. *NBER Working Paper* 9389. Cambridge, MA: National Bureau of Economic Research.

Angrist, Joshua D., Victor Lavy (2009). The Effects of High Stakes High School Achievement Awards: Evidence From a Randomized Trial. *American Economic Review* 99 (4): 1384-1414.

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

Batory, Agnes, Andrew Cartwright, Diane Stone (Eds.) (2018). *Policy Experiments, Failures and Innovations: Beyond Accession in Central and Eastern Europe*. Cheltenham: Edward Elgar Publishing.

Blanchenay, Patrick, Tracey Burns (2016). Chapter 8 - Policy Experiments in Complex Education Systems, in: Tracey Burns and Florian Köster (Eds.) *Educational Research and Innovation - Governing Education in a Complex World*. Paris: OECD Publishing: 170.

Callander, Steven, Patrick Hummel (2014). Preemptive Policy Experimentation. *Econometrica* 82 (4): 1509-1528.

Campbell, Donald T. (1969). Reforms as Axperiments. *American Psychologist*, 24: 409-429.

Christian, Brian (2012). The A/B Test: Inside the Technology That's Changing the Rules of Business, *Wired* 04/25/2012, https://www.wired.com/2012/04/ff-abtesting/ [accessed 13 December 2019].

Conahan, Frank C. (1994). Human Experimentation: An Overview of Cold War Experimentation Programs. Washington D. C.: Unites States General Accounting Office, http://archive.gao.gov/t2pbat2/152601.pdf [accessed 13 December 2019].

Connolly, Paul, Ciara Keenan, Karolina Urbanska (2018). The Trials of Evidence-Based Practice in Education: A Systematic Review of Randomised Controlled Trials in Education Research 1980–2016. *Educational Research* 60 (3): 276-291.

Cruces, Guillermo, Ricardo Perez-Truglia, Martin Tetaz (2013). Biased Perceptions of Income Distribution and Preferences for Redistribution: Evidence From a Survey Experiment. *Journal of Public Economics* 98: 100-112.

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 13 December 2019].

Falk, Armin, Anke Becker, Thomas Dohmen, Benjamin Enke, David Huffman, Uwe Sunde (2018). Golbal Evidence on Economic Preferences. *Quarterly Journal of Economics* 133 (4): 1645-1692.

Franco, Annie, Neil Malhotra, and Gabor Simonovits (2014). Publication Bias in the Social Sciences: Unlocking the File Drawer. *Science* 345 (6203): 1502-1505.

Goel, Vindu (2014). Facebook Tinkers With Users' Emotions in News Feed Experiment, Stirring Outcry, *The New York Times* 29/07/2014, https://www.nytimes.com/2014/06/30/technology/facebook-tinkers-with-users-emotions-in-news-feed-experiment-stirring-outcry.html [accessed 13 December 2019].

Grose, Christian R. (2014). Field Experimental Work on Political Institutions. *Annual Review of Political Science* 17: 355-70.

Haaland, Ingar, Christopher Roth (2017). Labor Market Concerns and Support for Immigration. *Unpublished Working Paper*.

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

Hawver, Mark (2014). OKCupid Looking Like OKStupid as Backlash Hits Over Social Experiments, *Tech Times* 07/30/2014, http://www.techtimes.com/articles/11640/20140730/okcupid-looking-okstupid-backlash-hits-over-social-experiements.htm [accessed 13 December 2019].

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

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 13 December 2019].

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

Kramer, Adam DI, Jamie E. Guillory, Jeffrey T. Hancock. (2014). Experimental Evidence of Massive-scale Emotional Contagion Through Social Networks. *Proceedings of the National Academy of Sciences* 111 (24): 8788-8790.

Krueger, Alan B. (1999). Experimental Estimates of Education Production Functions. *Quarterly Journal of Economics* 114: 497–532.

Kuziemko, Ilyana, Michael I. Norton, Emmanuel Saez, Stefanie Stantcheva (2015). How Elastic Are Preferences for Redistribution? Evidence from Randomized Survey Experiments. *American Economic Review* 105 (4): 1478-1508.

Lergetporer, Philipp, Guido Schwerdt, Katharina Werner, Martin R. West, Ludger Woessmann (2018). How Information Affects Support for Education Spending: Evidence from Survey Experiments in Germany and the United States. *Journal of Public Economics* 167: 138-157.

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): 585–96.

Lupia, Arthur (2002). New Ideas in Experimental Political Science. *Political Analysis* 10 (4): 319-324.

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. *Proceedings of the National Academy of Sciences*, 116 (22): 10723-10728.

Millner, Antony, Hélène Ollivier, Leo Simon (2014). Policy Experimentation, Political Competition, and Heterogeneous Beliefs. *Journal of Public Economics* 120 (2014): 84-96.

Mislavsky, Robert, Berkeley Dietvorst, Uri Simonsohn (2019a). Critical Condition: People Don't Dislike a Corporate Experiment More than They Dislike Its Worst Condition. *Marketing Science*, Articles in Advance: 1-13.

Mislavsky, Robert, Berkeley Dietvorst, Uri Simonsohn (2019b). The Minimum Mean Paradox: A Mechanical Explanation *for Apparent Experiment Aversion. Proceedings of the National Academy of Sciences*, 116 (48), 23883-23884.

Mukand, Sharun W., Dani Rodrik. (2005). In Search of the Holy Grail: Policy Convergence, Experimentation, and Economic Performance. *American Economic Review* 95 (1): 374-383.

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

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

Roth, Christopher, Sonja Settele, Johannes Wolfart (2018). Public Debt and the Demand for Government Spending and Taxation. *Unpublished Working Paper*.

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

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 13 December 2019].

Figures and Tables

Table 1: 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 high 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)*
Political leaning: conservative	0.336	0.018 *(0.029)*
Political leaning: socialist	0.268	-0.027 *(0.030)*
Political leaning: progressive	0.093	0.044 *(0.044)*
Non-partisans	0.303	-0.012 *(0.029)*
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)*
Observations	1,965	
F-Test for joint significance (p-value)		0.466

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. Municipality size: categorical variable equal to 1 for "1 to 1.999 residents" to 7 for "more than 500.000 residents". Observations might differ for individual rows due to missing values. Maximum number of observations: 1,965. Data source: ifo Education Survey 2017. Regressions weighted using survey weights. Data source: ifo Education Survey 2017. Significance levels: *** p<0.01, ** p<0.05, * p<0.10.

Table 2: 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: experimental treatments in the survey experiment. 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. 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.

Table 3: Treatment effects in sociodemographic subgroups

	Dependent variable: Support for reform evaluation
Treatment effects for the following subgroups:	(1)
No or basic school degree	0.014
	(0.037)
Middle school degree	0.006
	(0.042)
University entrance degree	-0.061*
	(0.036)
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: conservative	0.058
	(0.039)
Political leaning: socialist	-0.108***
	(0.041)
Political leaning: progressive	-0.012
	(0.072)
Non-partisans	0.003
	(0.044)

Notes: Each line represent the coefficient of a separate OLS regression. Reported coefficients are for interaction terms between the treatment indicator and the respective variables indicated in the left column. 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). 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.

Appendix Table A1: Who supports reform evaluations?

	Dependent variable: Support for reform evaluation
	(1)
Age	-0.001 *(0.001)*
Female	-0.062** *(0.031)*
Born in Germany	0.073 *(0.074)*
Municipality size	0.003 *(0.008)*
Monthly household income	-0.000 *(0.000)*
Partner in household	-0.011 *(0.033)*
Has parent(s) with high degree	-0.021 *(0.035)*
Works in education sector	0.015 *(0.052)*
Parents of school-aged children	-0.026 *(0.035)*
Lives in West Germany	0.038 *(0.038)*
Middle school degree	0.007 *(0.041)*
University entrance degree	0.068 *(0.051)*
University student	0.006 *(0.057)*
Vocational track	-0.022 *(0.050)*
Academic track	-0.050 *(0.059)*
Unemployed/Retired	0.013 *(0.038)*
Risk tolerance	-0.001 *(0.006)*
Patience	0.015** *(0.006)*
Offline-survey mode	0.088 *(0.063)*
Constant	0.660*** *(0.128)*
Observations	917
R^2	0.0211

Notes: OLS regressions. Dependent variable: Dummy variable 1 = "strongly support" or "somewhat support" reform evaluation, 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.

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	0.218	0.532	0.106	0.121	0.024
Observations			1,902		
R^2	0.019	0.011	0.021	0.017	0.019

Notes: OLS regressions. *"Experiment" treatment:* experimental treatments in the survey experiment. 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.

Appendix Figure A1: Screenshots of the survey questions

Baseline version

Experiment 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").