RUHR
ECONOMIC PAPERS

Katja Marie Fels

Who Nudges Whom?
Field Experiments with Public Partners

RUB RWI #906

Imprint

Ruhr Economic Papers

Published by

RWI – Leibniz-Institut für Wirtschaftsforschung
Hohenzollernstr. 1-3, 45128 Essen, Germany

Ruhr-Universität Bochum (RUB), Department of Economics
Universitätsstr. 150, 44801 Bochum, Germany

Technische Universität Dortmund, Department of Economic and Social Sciences
Vogelpothsweg 87, 44227 Dortmund, Germany

Universität Duisburg-Essen, Department of Economics
Universitätsstr. 12, 45117 Essen, Germany

Editors

Prof. Dr. Thomas K. Bauer
RUB, Department of Economics, Empirical Economics
Phone: +49 (0) 234/3 22 83 41, e-mail: thomas.bauer@rub.de

Prof. Dr. Wolfgang Leininger
Technische Universität Dortmund, Department of Economic and Social Sciences
Economics – Microeconomics
Phone: +49 (0) 231/7 55-3297, e-mail: W.Leininger@tu-dortmund.de

Prof. Dr. Volker Clausen
University of Duisburg-Essen, Department of Economics
International Economics
Phone: +49 (0) 201/1 83-3655, e-mail: vclausen@vwl.uni-due.de

Prof. Dr. Ronald Bachmann, Prof. Dr. Manuel Frondel, Prof. Dr. Torsten Schmidt, Prof. Dr. Ansgar Wübker
RWI, Phone: +49 (0) 201/81 49-213, e-mail: presse@rwi-essen.de

Editorial Office

Sabine Weiler
RWI, Phone: +49 (0) 201/81 49-213, e-mail: sabine.weiler@rwi-essen.de

Ruhr Economic Papers #906

Responsible Editor: Manuel Frondel

ISSN 1864-4872 (online) – ISBN 978-3-96973-049-2

Ruhr Economic Papers #906

Katja Marie Fels

Who Nudges Whom? Field Experiments with Public Partners

Bibliografische Informationen der Deutschen Nationalbibliothek

The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available on the Internet at **http://dnb.dnb.de**

RWI is funded by the Federal Government and the federal state of North Rhine-Westphalia.

http://dx.doi.org/10.4419/96973049
ISSN 1864-4872 (online)
ISBN 978-3-96973-049-2

Katja Marie Fels[1]

Who Nudges Whom? Field Experiments with Public Partners

Abstract

Field experiments which test the application of behavioural insights to policy design have become popular to inform policy decisions. This study is the first to empirically examine who and what drives these experiments with public partners. Through a mixed-methods approach, based on a novel dataset of insights from academic researchers, behavioural insight team members, and public servants, I derive three main results: Firstly, public servants have a considerable influence on study setup and sample design. Secondly, behavioural insight team members report concerns regarding scientific rigor and limitations imposed by risk-aversion of their public partners significantly more often than academic researchers. Thirdly, transparency and quality control in collaborative research are low with respect to pre-analysis plans, the publication of results, and medium or long term effects. To remedy the current weaknesses, the study sketches out several promising ways forward, such as setting up a matchmaking platform for researchers and public bodies to facilitate cooperation, and using time-embargoed pre-analysis plans.

JEL-Code: C93, D04, D90, H11

Keywords: Behavioural public policy; field experiments; Behavioural Insights Team (BIT); research transparency; expert interviews

April 2021

1 RWI and RUB. - My gratitude goes to the interviewees of this study: Paul Adams, Dan Ariely, Christian Gillitzer, Lindsey Maser, Ruth Persian, Dina Pomeranz, Thomas Tangen, Alex Sutherland, and Wilte Zijlstra for sharing their insights with me, as well as to all attendees of the "Behavioural Exchange 2019"-conference who participated in the anonymous survey. I thank Nils aus dem Moore, Gunther Bensch, Liam Delaney, Jonathan Meer, Christoph M. Schmidt, Frederic P. Schuller, Annekathrin Schoofs, Mathias Sinning and Stephan Sommer for comments on study design or an earlier version of this paper. Thanks to Alex Bartel for his reliable research assistance. This work has been partly supported by a special grant from the German Federal Ministry for Economic Affairs and Energy and the Ministry of Innovation, Science, and Research of the State of North Rhine-Westphalia.– *All correspondence to: Katja Fels, RWI, Hohenzollernstr. 1-3, 45128 Essen, Germany, e-mail: katja.fels@rwi-essen.de*

1 Introduction

Collaborative research projects[1] involving policy makers and either academic or practical behavioural researchers are increasingly attracting attention. Just ten years ago the British government was the first to establish its own government unit to practically improve policy design based on behavioural insights (Sanders et al. 2018). Today, more than 200 institutions worldwide apply behavioural insights to public policy and test their application in the field (OECD, 2020). On the academic side, the publication of "Nudge" by Thaler & Sunstein (2008) ignited unprecedented interest of university researchers in partnering with public bodies in order to conduct behavioural field experiments.

The instrument that has received most attention in behavioural public policy are nudges. These interventions alter the decision environment of individuals "without forbidding any options or significantly changing their economic consequences" (Thaler & Sunstein, 2008). Hence nudge interventions are perceived as less intrusive than other policy tools, such as setting up bans, regulations, or monetary incentives. Nudges are widely applied in policy design (Madrian, 2014; DellaVigna, 2009) with applications ranging from automatic enrollment to pension accounts (Madrian & Shea, 2001) over inducing energy conservation of private households (Allcott, 2011) to improving tax compliance (Hallsworth et al., 2017). They provide an attractive tool for policy makers to address prevalent problems because implementation costs are low, the necessary changes to encourage citizens to make particular choices appear small, and nudge interventions can be tested via randomized controlled trials (RCTs) before roll-out (Einfeld, 2019).[2]

This study presents the first empirical investigation into the current state of collaborative research. It has the objective to empirically describe strengths as well as pitfalls of collaborative experiments and aims to inform academia and public policy institutions alike. For the analysis, I use a novel dataset of anonymously collected insights from 70 public servants, behavioural insight team members, and academic researchers with experience in collaborative research. Additionally, I conducted qualitative interviews with nine selected experts to allow for a more comprehensive interpretation of the patterns observed in the quantitative data and to develop some practical recommendations of how to remedy the observed weaknesses.

Based on the mixed-methods approach employed in this paper, I derive three main results. Firstly, public servants have an enormous influence on the design of collaborative field experiments, specifically on developing the research question and selecting the sample. At the same time, the data of this study indicate that public servants have clearly different priorities than researchers. This has implications for the scope and focus of the research endeavour.

Secondly, the study finds that behavioural insight team members seem to be working under a particular pressure to accommodate the needs of their public partner. A majority of behavioural

[1] Throughout the paper, the term "collaborative research" refers to the collaboration between a public body on the one hand and either academic researchers or behavioural insight teams on the other.

[2] For a critical discussion about what RCTs can and cannot do in evidence-based policy making, see Deaton & Cartwright (2018), Pritchett & Sandefur (2014), Harrison (2014), Cartwright & Hardie (2012), and Cartwright (2010).

insight team members explicitly state that they had to move away from an ideal scientific approach to accommodate the requirements of their cooperation partner while this does not seem to be a problem for academic researchers. Main reason for this seems to be that academic researchers have more freedom to call off an experiment if their requirements are not met whereas behavioural insight team members are bound by contracts.

Thirdly, the study documents that transparency and quality control in collaborative research — as manifested in pre-analysis plans, the publication of results, and the measurement of medium or long term effects — tends to be low.

Based on the findings presented in the paper, the study makes several suggestions for improvements, among them establishing a new behavioural insights working paper series and a matchmaking platform for interested researchers and public bodies. With respect to processes within the public bodies, internal guidelines specifying the authorized ways of collaborative experimentation and cooperation with other public institutions could facilitate cooperative research. Moreover, a co-funding by foundations or non-profit organizations would support public bodies low on resources and help to emancipate behavioural insight teams from a too strong agenda of their public partner.

The present study contributes to the small body of literature that examines the work of behavioural insight teams and collaborative experiments with public partners from a meta perspective. While a large number of papers investigate the effectiveness of different nudges (see the systematic reviews by Hallsworth (2014); Andor & Fels (2018), and the cost-benefit-analysis by Benartzi et al. (2017)), not many studies take such a meta-level approach. As an exception, Sanders et al. (2018) discuss complications, challenges and opportunities for the work of the British Behavioural Insights Team (BIT). They also touch upon some general questions related to cooperations with public partners. In his response comment, Delaney (2018) raises the general question of how professional standards can be ensured in the quickly growing field of behavioural scientists, given that many of them nowadays are practitioners mainly working in a consultancy capacity. Such deliberations help assess the current state of affairs in collaborative research. Yet up to now, no systematic empirical description is available.

By its choice of methodology, the present study also contributes to the emerging literature on analysing data from the insights of experts. An expert is someone who is either i) responsible for the development, implementation or control of a way of solving problems, or ii) has exclusive access to data regarding groups of decision-makers or the process of decision-making (Mayer, 2006). While using experts' opinions as a data source has been popular in sociology and political science for a long time, the number of economic papers that take this approach has only recently seen a notable rise. DellaVigna & Pope (2018) examine the forecasting ability of experts with a survey design that is combined with a real-effort task. Fecher et al. (2016) survey researchers who work with panel data and report which barriers are keeping them from replicating their results. Pomeranz & Vila-Belda (2019) sent an online questionnaire to researchers who have collaborated with tax authorities before and derive recommendations from their experiences. Vivalt & Coville (2017) examine how policymakers, practitioners, and researchers update their beliefs in response to study evidence by surveying the participants of several World Bank and Inter-American-Development Bank workshops.

Like them, the author of this study made use of a unique opportunity to conduct a survey among conference participants at the "Behavioural Exchange 2019 (BX2019)", one of the largest conferences on behavioural insights worldwide. In addition to that, in-depth-interviews with selected experts help to enhance the interpretation and learnings from the quantitative data.

Finally, the paper provides a contribution to the literature on transparency in economic research (Miguel et al., 2014; Christensen & Miguel, 2018). In light of the so called *replication crisis*, which first emerged in medical trials (Ioannidis, 2005) and increasingly affected other disciplines such as psychology and economics, a prevalent concern with respect to policy advice is that misleading bodies of evidence are produced (Christensen & Miguel, 2018). Collaborative experiments are mainly designed to inform policy decisions, they hence should be under special scrutiny. As Karlan & Appel (2018) put it: "A bad RCT can be worse than doing no study at all: it teaches us little, uses up resources that could be spent on providing more services (even if of uncertain value) (...), and if believed may even steer us in the wrong direction." By documenting that common tools of research transparency, such as pre-registry and publication, are not well used in collaborative research, this paper shows where collaborative research can be improved.

The paper is structured as follows: Section 2 outlines the methodology of the quantitative and qualitative data collection. Section 3 reports results from the quantitative survey on the motivation for conducting behavioural experiments with public partners and on the choice of interventions and policy fields. Section 4 takes a closer look at the role of public servants as gatekeepers in collaborative research. Section 5 reports data on the experience of behavioural insight team members. Section 6 focuses on transparency and quality control in collaborative experiments. Section 7 suggests remedies for some of the perceived shortfalls. The final section concludes.

2 Methodology

The present study applies a mixed-methods approach, combining quantitative and qualitative elements, in order to ensure breadth and depth of understanding and for corroboration (Johnson et al., 2007). The applied quantitative and qualitative methodologies are described in turn.

2.1 Quantitative survey

For collecting quantitative data, I designed an anonymous 5-minutes-questionnaire, which was answered by in total 70 participants of the "Behavioural Exchange 2019 (BX 2019)"-conference. The BX2019 is one of the largest gatherings of behavioural insights-experts with around 1,200 attendees from all around the world. As first step, I personally handed out the questionnaire on September 5-6, 2019, to attendees during breaks in the conference program.[3] The personal contact led to a good response rate: Of 152 handed out questionnaires, 43 (28.3%) respondents returned a completed survey form.

[3]Participants had three options of returning the completed form: either directly in person to me, via email, or in one of the boxes placed at the entrance hall of the conference venue.

Most likely the individuals taking part in the survey are particularly affected by the topic, while those who did not participate might not have seen much relevance of the questions for themselves. Yet since the aim of the survey is not to provide a representative picture of opinions on collaborative research in general but rather focuses on personal experiences individuals made during their collaborative research, this self selection does not seem to limit the data's scope to answer the research question.

As second step, I created an online version of the questionnaire. All BX2019-attendees who indicated their email address in the conference app and who were identified to belong to one of my the three target groups – academic researchers, behavioural insight team members, public servants – were contacted via email. The personalized email was sent out on September 16, 2019. It asked to reply with "YES" when the attendee was willing to participate in the survey, otherwise she would not be contacted again.[4] The response rate was much lower than during the personal contact. Of 260 individuals contacted via email, 45 replied with "YES" and received the link to the online survey. Only 27 of them used the link and answered the questions (10.4%).

The questionnaire comprised twelve questions (see section A, Appendix). I pilot-tested it with several researchers who had run field experiments with a public partner and incorporated their feedback. The online survey represented a slightly modified version of the paper-and-pen version.[5] It was published on www.onlineumfragen.com, a Swiss survey platform.

2.2 Qualitative interviews

In order to complement the quantitative survey, this study additionally includes insights from semi-structured interviews with selected experts. Such an approach is recommended when the study's aim is to gain sophisticated insights into aspects of social reality (Hoffmeyer-Zlotnik, 1992) and wants to capture nuances that cannot be detected by a quantitative approach (Glennerster et al., 2018).

All interviewees were chosen in their position as experts. I contacted them via email. With two exceptions, all of them either agreed to conduct the interview or referred me to a colleague who would speak to me instead. Table 1 provides an overview of the final sample of interview partners. The first interview took place on 6 September 2019, at the BX2019 conference in London. The other interviews were conducted via video call between 25 June and 14 July 2020. The duration of the interviews varied between 31 and 52 minutes. Each interviewee agreed to recording the video call to facilitate documentation. In addition to that, the direct quotes used in this study have been sent to the interviewees for authorization. All interviewees consented to be named with full name and position.

[4]This study fully complies with German data protection law, hence a follow-up reminder to non-respondents was not feasible.

[5]Main modifications of the online version were that I split some of the more complex questions into two and made use of the possibility to set junctions. This made respondents receive a slightly different set of questions depending on whether they had indicated to be a researcher, behavioural insight team member or public servant.

For the interviews, I used a semi-structured approach: a set of pre-determined questions (see section B) was posed in a flexible order. During semi-structured interviews, it is also possible to ask ad-hoc questions to let the interviewee further elaborate on aspects that come up during the interview (Ebbecke, 2008). Such an approach helps to focus the interviewees' answers to the point of interest while avoiding to exclude relevant side-information, hence capturing the full range of the topic (Bock, 1992).

Table 1: List of interview partners

Name	**Position**	**Category**	**Experience**
Dan Ariely, PhD	James B. Duke Professor of Psychology and Behavioral Economics *Duke University, United States*	Academic Researcher	Participated in more than 30 experiments. In 1996, Ariely founded the Center for Advanced Hindsight; he and his team are offering research into behavioural sciences to organizations and (public sector) partners.
Christian Gillitzer, PhD	Lecturer *The University of Sydney, Australia*	Academic Researcher	Participated in 1 experiment. Gillitzer was part of a publicly funded research cooperation with the Australian Taxation Office. The collaborative study was recently published in JEBO: Gillitzer & Sinning (2020).
Dina Pomeranz, PhD	Assistant Professor of Applied Economics *University of Zurich, Switzerland*	Academic Researcher	Participated in 7 experiments. As Phd student Pomeranz conducted her first collaborative experiment with the Chilenean Tax Authority. The study was published in the AER: Pomeranz (2015). Her current work mainly focuses on taxation and public procurement.

Dr. Alex Sutherland	Chief Scientist/ Director of Research and Evaluation *The Behavioural Insights Team (BIT), United Kingdom*	Behavioural insight team member	Participated in 15 experiments. Sutherland came to BIT in 2019 after he had worked 5 years at RAND Europe. As director of research and evaluation, he is responsible for ensuring the overall standards and quality of BIT's research.
Ruth Persian	Senior Advisor *The Behavioural Insights Team (BIT), United Kingdom*	Behavioural insight team member	Participated in 13 experiments and quasi-experimental evaluations. After first experiences in field experimentation at the World Bank, Persian joined BIT in 2016. Her current focus is on applying behavioural insights and rigorous evaluation to public policy and programmes in low and middle income countries.
Paul Adams	Manager, Behavioural Economics and Design Unit [until 2019] *Financial Conduct Authority (FCA), United Kingdom*	Behavioural insight team member	Participated in 18 field experiments. First contact with behavioural insights when he joined the FCA in 2012. In March 2019, Adams changed to the Consumer Behaviour team of the Authority for the Financial Markets, Netherlands. All quotes refer to his work at the FCA.
Wilte Zijlstra, PhD	Consumer Behavior Expert *Authority for the Financial Markets (AFM), The Netherlands*	Behavioural insight team member	Participated in 5 experiments plus several online choice experiments. Having a professional background in Evolutionary Biology, field experimentation has long been a standard scientific method for Zijlstra. He joined the AFM in 2006. When the internal behavioural insight unit was founded in 2016, Zijlstra became part of the team.

Lindsey Maser	Communications and Behavioural Science Advisor *City of Portland Bureau of Planning and Sustainability, United States*	Public Servant	Participated in 16 experiments. As part of the grant-funded "What Works City"-initiative, the City of Portland entered a partnership with BIT. They have also partnered with Ideas42 and the Center for Advanced Hindsight. Even though her official position focuses on sustainability issues, Maser acts a liaison and coordinator for all behavioural experiments that take place within of the City of Portland.
Thomas Tangen	Senior Communications Advisor *The Norwegian Tax Administration (NTA), Norway*	Public Servant	Participated in 1 experiment. As Communications Advisor, Tangen works in the Directorate, the administrative part of the tax administration. In 2013, he was part of a collaborative field study with academic researchers (published in Management Science: Bott et al. 2019).

2.3 Descriptives

This section presents descriptive results from of the anonymous survey. Table 2 reports respondents by their affiliation. Out of 70 survey respondents, 60 (86%) provided information about this.[6] Some 32% of them are academic researchers, 37% behavioural insight team members, 18% public servants and 13% employees from a private company.[7]

[6]Two respondents classified themselves as academic researchers as well as behavioural insight team members. In Table 2, they are listed as researchers. In the analysis part, they are included in both subsamples.

[7]Only very few respondents chose the option "none of the above" and specified their position. From their free text answers it became clear that they chose this option because they had not collaborated with a public body on experiments before. Releasing this condition, I could classify them into one of the four categories listed in Table 2.

Table 2: Descriptive Statistics

	researcher	**behavioural insight team**	**public servant**	**private employee**	**total**
frequency	19 (31.7%)	22 (36.7%)	11 (18.3%)	8 (13.3%)	60 (100%)
no own experiments	9	6	5	2	23
1-2 own experiments	6	4	2	3	16
3-4 own experiments	3	3	2	3	12
5-6 own experiments	0	2	0	0	3
more than 6 exp.	1	6	2	0	10

Note: Only respondents that indicated their affiliation are included in this table. The full sample comprises 70 respondents.

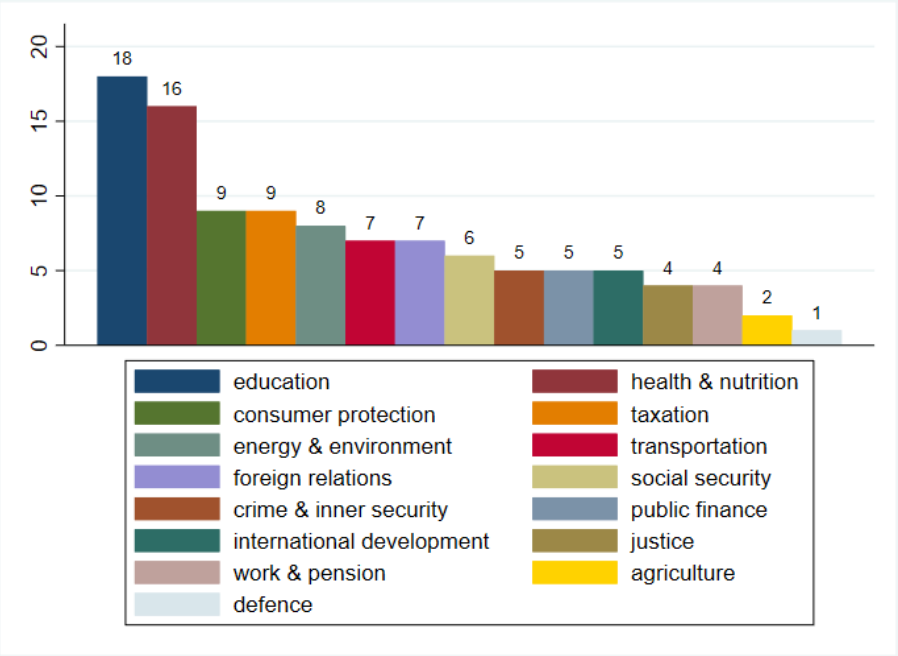
Table 2 also provides an overview of the number of experiments the respondents have conducted. Behavioural insight team members are much more experienced in collaborative experimentation than academic researchers and public servants. This reflects that their job mainly consists of designing and implementing experiments. While among academic researchers and public servants roughly half of the sample has some own experimentation experience, the great majority of behavioural insight team members (71%) reports to have conducted at least 1-2 trials. 29% of them even state to have run more than six own experiments.

Most of the respondents conducted experiments in Anglo-Saxon countries like the UK, the US, Australia, Canada, or Ireland, with an overwhelming majority being conducted in the UK. This is likely due to the fact that the BX2019-conference was organized by the British Behavioural Insights Team (BIT) and many of the behavioural insight team members in my sample stem from the BIT. The second group of countries, in which experiments were conducted, include European countries like the Netherlands, Denmark, Lithuania, Finland, Belgium, Germany, and France. Respondents also reported a few collaborative experiments in Afghanistan, Guatemala, Philippines, Kenya, Sierra Leone, and Georgia.

In terms of policy fields, two fields clearly stand out to have been most often the policy area of an own experiment: education (18) and health and nutrition (16) (see Figure 1). In the second tier, taxation (9) and consumer protection (9) attracted a lot of attention. The least experiments were conducted in agriculture (2) and defence (1).

Putting focus on the interventions that have been tested, a top 5 emerges (see Figure 2): social norms (20), simplification (16), increase in ease and convenience (16), letter design (15), and – with a little distance – reminders (12). All the interventions from this top 5 can be considered as minimally intrusive. They would not meet much resistance when discussed with policy partners, potentially in contrast to nudges like eliciting implementation intentions, disclosure, or changing the default

Figure 1: Policy fields of own experiments



Notes: The sum is equal to the number of nominations of the respective policy field not the number of experiments. If a respondent conducted more than one experiment in a certain field, it will only be counted once. Four respondents additionally chose "other".

rule, which attracted much less attention by survey respondents. This pattern fits to what has been documented in the literature elsewhere. For example, in more than 100 trials conducted by the two main behavioural insight teams in the US, a change of default settings was only tested twice (in one trial) (DellaVigna & Linos, 2020). It seems that many behavioural economists followed the advise of Sanders et al. (2018) to pursue low hanging fruit first.

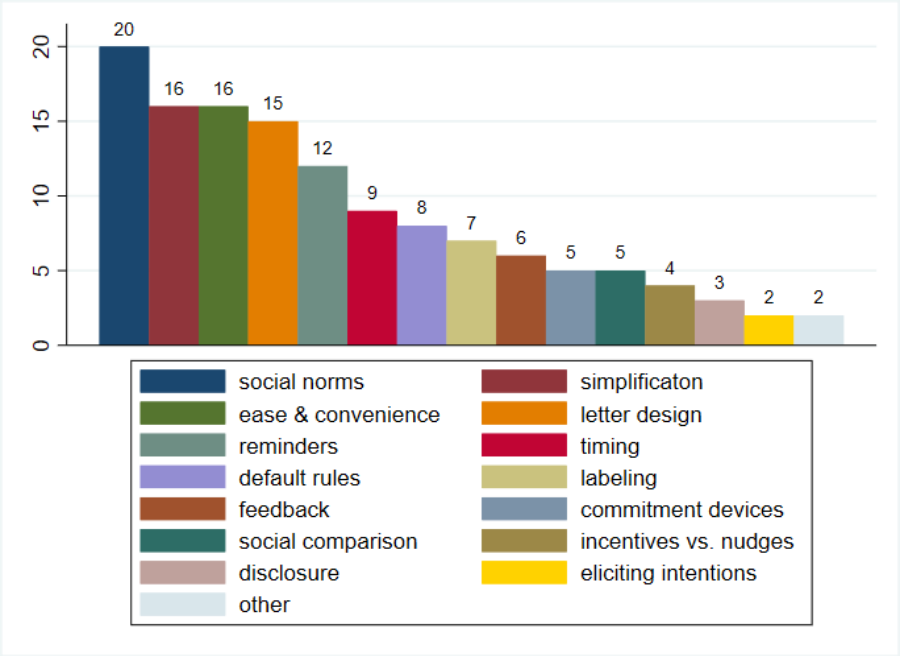
3 Behavioural insight experiments with public partners

In this section, the study presents first results from the anonymous survey on what motivates behavioural experiments with public partners and which policy fields and interventions are considered most relevant for this type of research.

3.1 Motivation

When asked what the greatest advantage of collaborative research with a public body is, the majority of respondents (64%) chose "increased political and practical relevance of research" among the four answer options (see Table 3). In addition to that, 7% enlisted reasons in the open text field option that could be summarized into a similar category: creating direct benefits of research for policy

Figure 2: Interventions tested in own experiments



Notes: The sum is equal to the number of nominations of the respective intervention not the number of experiments. If a respondent tested a particular intervention in more than one experiment, it will only be counted once.

institutions and the wider public.[8] In contrast to that, new starting points for research (16%) and access to new types of data (10%), which could be important reasons for academic researchers to start a collaboration, were not nominated as often.

The importance of making research useful for society was also emphasized during the qualitative interviews. As Persian (2020), a senior analyst at BIT, put it: "In comparison to conducting fieldwork or RCTs exclusively with academic partners, the advantage is that ideally it's relevant. So there is a public body that is interested in the research question." That was also the main motivation for researcher Christian Gillitzer (2020) to partner with the Australian Taxation Office (ATO): "A good thing about working with the ATO is you find out about questions that are of policy relevance. So not all the things are academically interesting, but they might show policy relevance. And there are a lot of dollars attached to the questions they want answers to."

On the administrative side, Tangen (2020), a senior communications advisor at the Norwegian Tax Administration, reports: "The results from the experiments – how we write letters, how we write press releases, how we word statements to the media – all things have been used a lot afterwards. We have learned a lot from it." This leads to a very distinct difference of collaborative research to other research: while many academics aim to influence public policy by publishing their study results,

[8]The specific answers were: "better outcome for citizens and consumers", "changing outcomes for the public", "helping public partners to become more effective", "better understanding of dos and don'ts in practice", and "direct impact and benefits of research for wider public".

researchers in collaborative research experience it the other way round: "impact comes first" (Sanders et al., 2018, p.156). "That's a key thing that I always stress to people if they want to start working with an institutional partner: to make sure you listen to what they care about. I look at it like a Venn diagram of the things that are academically relevant and publishable and the things that are relevant for the policy partner", Dina Pomeranz (2020) from Zurich University says.

Table 3: Greatest advantage of collaborative research

	frequency	**percent**
increased political and practical relevance of research	44	63.8
new starting points for research	11	15.9
access to new types of data	7	10.1
direct benefits of research for the wider public	5	7.3
other	2	2.9

3.2 Interventions

If the respondents were offered to test any intervention they like, the nudge chosen most often is with overwhelming majority default rules (41% of maximal points), one of the most effective but also most controversial nudges (see Jachimowicz et al. (2019), Reisch & Sunstein (2016), Sunstein et al. (2018)). The second most relevant intervention, according to the respondents, is simplification (31%), closely followed by increase in ease and convenience (30%), and social norms (30%) (see Figure 3).

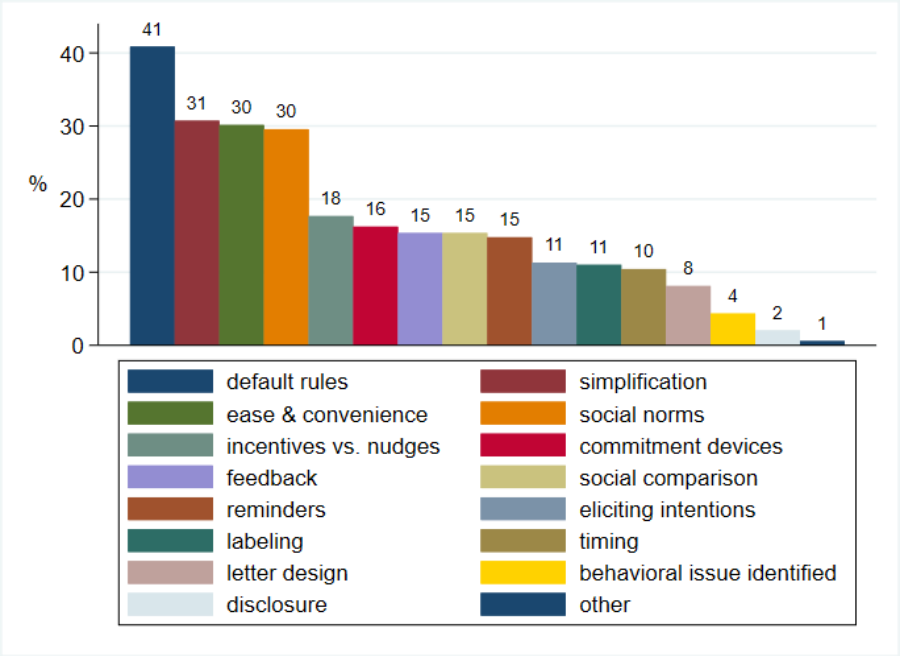
Interestingly, the top position of default rules stands in contrast to the interventions the respondents actually tested themselves. All tested nudges were much less intrusive or controversial. The top 4 interventions have another common feature: they are designed to remove friction costs. That respondents ranked this feature highly is perfectly in line with a statement by Nobel laureate Daniel Kahneman who called it "the best idea I ever heard in psychology" (Dubner, 2017).[9] Three of the interventions ranked by respondents in the top 4 - changing the default, simplification, and increase in ease and convenience - offer answers to this question.

Social norms, too, could be seen as removing barriers: to provide information about injunctive norms[10] reduces uncertainty about the desired behavior (Cialdini et al., 1990). Another reason why study respondents have ranked social norms so highly might be that this interventions is very popular in the literature and has shown to be effective in many different settings, for example tax compliance

[9] As early as in 1947, social psychologist Kurt Lewin pointed out in his model of planned change that people's behavior is driven by two external forces: a driving force and a restraining force. When both forces are in equilibrium, this results in the individual's behavior. While it is intuitive and popular to think about what incentives could be created to "drive" behavior in a certain direction, the other approach might be equally if not more powerful: reducing the restraining forces by asking "What can I do to make the desired behavior easier?" (Lewin et al., 1947).

[10] The literature distinguishes between two different types of social norms: *descriptive* norms referring to what is commonly done and *injunctive* norms referring to what is commonly approved (Cialdini et al., 1990).

Figure 3: Interventions assessed as most relevant to be tested



Note: The figure depicts the percentage of maximal points each variable received in the ranking. For each time being nominated on rank 1, an intervention received five points, for rank 2 it received four points, and so on. The maximal sum of points from 69 valid responses was 345. Depicted numbers are rounded.

(Bobek et al., 2013), health behavior (Lewis & Neighbors, 2006), charitable giving (Frey & Meier, 2004), and energy conservation, also in the long run (Ferraro et al., 2011).

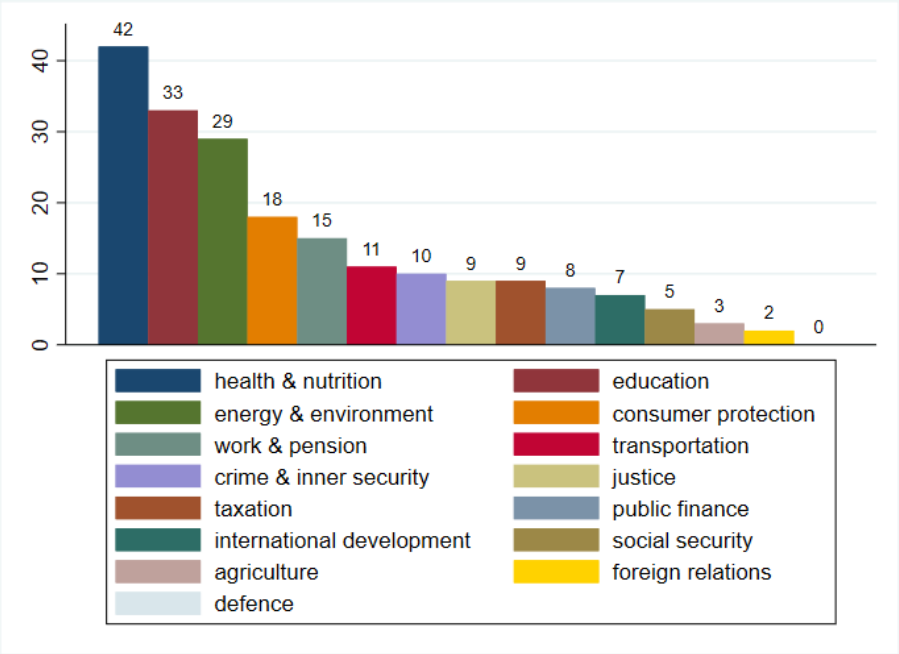
3.3 Policy fields

Among a wide range of public policy fields, a clear top 3 is considered by study respondents as most relevant for conducting collaborative experiments: health and nutrition (42 nominations), education (33), and energy and environment (29) (see Figure 4). All three fields are very closely connected to the daily life of citizens. In contrast to that, defence (0), foreign relations (2), and agriculture (3), which affect daily life rather remotely, receive the least nominations for the top 3.

When compared with the number of policy trials that have been actually conducted in different fields, some interesting lessons can be learned. For the comparison, I use the AEA RCT registry, the most popular database in economics to pre-register RCTs. On 15 May 2020, 1,159 trials were enlisted to have been completed. Looking at the top ranks of completed trials' policy fields, they mirror pretty closely what respondents of this study deem most relevant (see Figure 5). Education and health gather the top 2 positions (25.4% respective 23.1% of completed trials). Labor ranks third (20% of completed trials), a policy field that corresponds to "work and pension" in the questionnaire and gathers the 5th most nominations for the top 3 by study respondents.

Yet, also some remarkable differences occur. Firstly, while energy and environment is deemed very important by study respondents (3rd most nominations for the top 3), it only achieves rank 8 (5.8% of completed trials) in the AEA registry. This difference might be due to the fact that study respondents were asked which policy fields they find most relevant for cooperation studies with public partners while in the AEA registry, of course, not only trials conducted with a public partner are enlisted. However, in the field of energy conservation the most natural cooperation partners are private energy providers. Hence allowing greater freedom with respect to potential cooperation partners (as in the AEA registry), should rather drive results upwards than downwards.

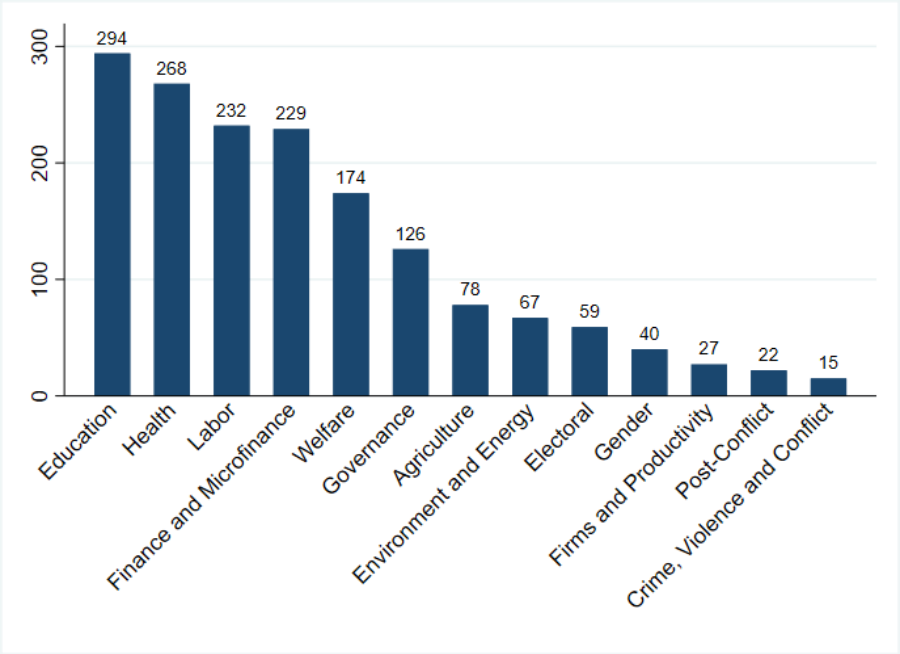
Figure 4: Policy fields – sum of top3-nominations



The opposite is true for policy fields like crime and inner security which, secondly, show a clear gap between relevance assessment and realization. While study respondents nominated crime and inner security ten times for the top 3, the corresponding policy field "crime, violence and conflict" makes up only 1.3% of completed trials; it holds the final position in the AEA RCT registry. In this case, a potential upward bias of study respondents' answers seems to be likely because crime and inner security is a core policy field of the public hand, and they were asked about cooperational research with a public partner.

Thirdly, among study respondents the policy field agriculture is not deemed very relevant (only 3 nominations for the top 3), while it makes up for 6.7 percent of actually completed trials (more than energy and environment). This might be due to the fact that the most popular region for trials enlisted in the AEA RCT registry is Africa where agriculture plays a crucial role for income generation. Study respondents, on the other hand, mainly focus their research on Anglo-Saxon countries (see section 2.3) and hence might take a different perspective.

Figure 5: Policy fields – number of completed trials



Note: Total number of completed trials in the AEA RCT registry on 15 May, 2020: 1,159. Each trial may belong to multiple policy fields.

Unfortunately, policy fields like consumer protection, transportation, justice, and taxation, which are deemed highly relevant by study respondents, do not have an equivalent category in the AEA RCT registry. A possible explanation is that these fields do not make up many of the studies enlisted in the registry, since otherwise the categories would have been added, but this interpretation is up for further investigation.

In conclusion, education and health, the two policy fields assessed as most relevant by study respondents, also have been most researched. In contrast, energy and environment seems to be an under-researched policy field, and - with some caution due to potential bias - crime and inner security as well.

4 Public servants: influential gatekeepers

The data of this study reveals that public servants yield a great influence on collaborative research. This section takes a closer look at their role.

4.1 Influence on study design

Asked how the research question in their collaborative research was derived, only a small minority (13%) of respondents indicated that a knowledge gap identified by the researcher was the starting point (see Table 4). In contrast, 40% indicated that a knowledge gap identified by the public body has played this role, and another 48% referred to consultations between the researcher and the public body. This matches results from an earlier survey by Pomeranz & Vila-Belda (2019) which focussed

on collaborations with tax authorities. They find that in 40% of cases the research idea emerged from jointly exploring topics of common interest between researchers and their public partner.

Moreover, all interviewees confirmed that the public body has a decisive influence on designing the research question. As Ruth Persian (2020) clearly puts it for studies conducted by the BIT: "The research question in a way is set by the public sector partner." Academic researchers like Christian Gillitzer (2020) had similar experiences in collaborative research: "This was more of a partnership where the ideas and proposals were directed by the ATO: they had a business need and something they wanted evaluated. Our scope and role was to refine and design the intervention such that it could be tested scientifically." Yet there seems to be some scope to increase researcher's influence over time. Paul Adams (2020), a former manager of the behavioural insight unit at the Financial Conduct Authority (FCA) in the UK, recounts: "In some of the earlier trials, the policy interventions were mostly designed by the policy makers. Over time, when results were not as positive as expected, we started to develop a bigger role earlier in the process, and used some other techniques to help design and develop more effective interventions."

With respect to selecting the sample, the data again documents a great influence of the public body. As Table 4 depicts, only one fifth of respondents (21%) state that the researchers were free to choose any sample from the target population. In more than half of the cases (53%), the researchers had to choose their experimental sample from a sub-population, which the public body selected beforehand. Some 26% of respondents even indicate that the sample was entirely chosen by the public body. In sum, public servants at least pre-selected the sample in 80% of studies. "There was a trend in all the trials that went ahead that over time the samples tended to get smaller and smaller and smaller. So there was a bit of an incentives for the researchers to go in an overbid on the sample size to the expectation that there would be a reduced size by the time that the actual intervention went into the field", Christian Gillitzer (2020) from Sydney University remembers.

Table 4: Research question and selection of the sample

	frequency	percent
How was the research question derived?		
knowledge gap identified by researchers	5	12.5
knowledge gap identified by public servants	16	40.0
consultation between research. and public servants	19	47.5
Who selected the sample?		
researchers were free to choose	8	21.0
researchers chose from a pre-selected population	20	52.6
public body chose the sample	10	26.3

Yet all behavioural insight team members and academic researchers made clear during the interviews that they would not go ahead with ill-powered studies or a sample design that does not comply with scientific standards. "Basically, for quantitative research we need a certain number of observations. If they don't have that, we cannot do the study. The reason is that we need a certain sample size to have statistical precision", Dina Pomeranz (2020) from Zurich University summarizes. A similar view is put forward by Dan Ariely (2019), founding member of the Center for Advanced Hindsight: "For me, experiments with public partners do not mean a step back in scientific rigor. Because I am also happy to say *no* to an experiment. It's helpful that we are an outside company."

For behavioural insight teams like the BIT, the focus lies on explaining the statistical needs in plain language: "It's being very clear upfront about what standards you have and why. It's good to explain very carefully and be prepared to have these conversations several times. You will have to tell different people. Explain concepts you have. Power calculations is not something that's intuitive to many people but it's finding a way to making it intuitive", Alex Sutherland (2020), chief scientist and director of research and evaluation at the BIT, explains. Two lessons also Karlan & Appel (2018, p. 36) distill from their case studies is "to watch these steps closely, especially if they are being executed by partners unfamiliar with randomization" and to know when to walk away from a failing partnership.

However, when the interviewees were asked whether they were free to walk away from an experiment with a public partner in case of quality concerns, one interviewee admitted: "Walking away from an entire project is difficult. There are contractual obligations. I mean, there are implications, not for me personally, but for the project itself." Another one added: "We could not walk away at any time, no. Once we've invested the political capital and the energy of our partner, walking away from a trial is probably not going to go down very well. That's why we were always very careful in our negotiations and be very clear that there is a contracting process, and there's always a stop-go-decision at the very end."

On the administrative side, the interviewed public servants were despite their adherent interest and experiences in behavioural experiments very aware of their organisation's limited expertise in experimental design and the current state of research. "The advantage is definitely the level of expertise we get. We don't have anyone employed at the city of Portland who is a behavioural scientist or has that level of expertise, that connection to the ongoing, rapidly evolving field of research", Lindsay Maser (2020) says. Thomas Tangen (2020) from the Norwegian Tax Administration consents: "All the very young academics we have hired are from different universities. But I still think it is important to have a close relationship with people outside the administration."

4.2 Differing priorities

The great influence of the public body on study design might have meaningful consequences if priorities of public servants and researchers in testing behavioural interventions substantially differed. In these cases, collaborative research with a public partner would lead to a systematically different research

focus than other academic work, potentially along a research agenda that is sub-optimal (Levitt & List, 2009).[11] The data of this study provides some first empirical hints into that direction.

Firstly, public servants and academic researchers clearly yield different priorities regarding the interventions which they feel are most relevant to be tested (see Figure 6). Several interventions which researchers rank highly have not been nominated by a public servant once: monetary incentives versus nudges (28% vs. 0% of maximal points), eliciting implementation intentions (14% vs. 0%), and disclosure (7% vs. 0%). In addition to that, academic researchers assess several interventions as much more relevant than public servants: feedback (29% vs. 11%), labeling (18% vs. 4%), and commitment devices (15% vs. 5%). Given the great influence of public servants on study design, this will likely lead to these interventions being under-researched in collaborative experiments. Only default rules are assessed as similarly important by both groups. Yet while public servants rank simplification higher than defaults, this evaluation is not shared by researchers: among them, simplification only receives half as much appreciation as among public servants (23% vs. 58%). A similar picture occurs with increase in ease and convenience, an intervention which is ranked highly by public servants but much less so by researchers (20% vs. 45%). An interview confirmed that research interests between these two groups differ: "We did have early on a conference where we met with many senior people from the public partner. There we came on with proposals. And they were receptive. But ultimately the things that went ahead were dictated primarily by their business needs."[12]

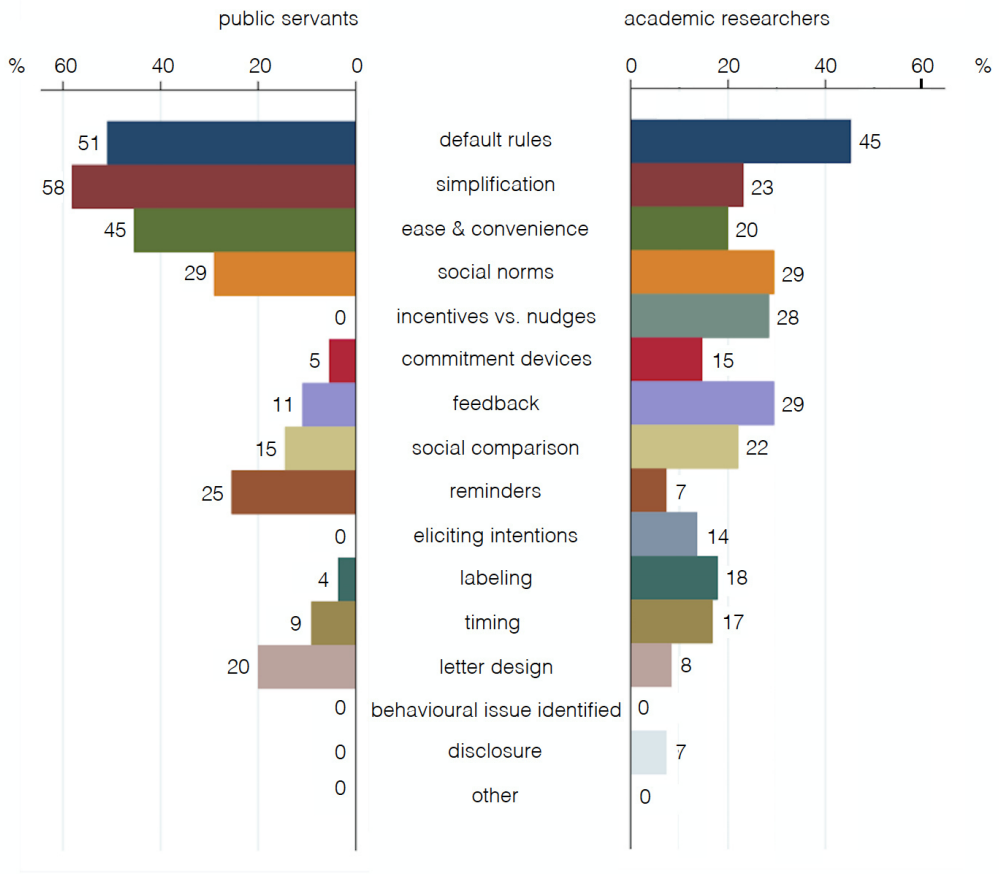
Secondly, with respect to policy fields in which collaborative experiments should be conducted, less pronounced differences can be observed. Researchers and public servants overall agree that health and nutrition, education, and energy and environment belong to the top 3 of relevant fields to test behavioural insights (see Figure 7). Yet while academic researchers put similar emphasis on health and nutrition (55% of maximal points) and education (53%), for public servants, health (60%) clearly ranks before education (42%). Academic researchers also attribute high relevance to two fields that are considered by public servants much less frequent or not at all: international development (12% vs. 0%) and public finance (18% vs. 5%). Three other fields, on the other hand, are considered relatively more important by public servants than by academic researchers: social security (5% vs. 24%), transportation (14% vs. 22%) and consumer protection (21% vs. 29%). Overall, there seems unity that fields like foreign relations, agriculture and defence are not very relevant for testing the application of behavioural insights in the field. However, in practise, respondents from the sample indicated to have conducted own experiments in these policy fields: foreign relations was nominated seven times, agriculture two times, and defence at least once (see Figure 1).

Thirdly, when asked what the greatest advantage of collaborative research is, a substantial share of public servants (36%) chose the option "access to new types of data". Among the group of researchers, only 11% picked this answer (see Table 5). This is surprising as one might have expected

[11] Additionally, there might be competing priorities on the individual level, which Karlan & Appel (2018) document for organizations from the international development context: If individuals parallel to implementing an intervention also need to achieve other organizational goals, it might be rational for them to not fully follow the research protocol.

[12] The interview partner wants this quote to be cited anonymously.

Figure 6: Ranking of interventions – subsamples public servants and academic researchers

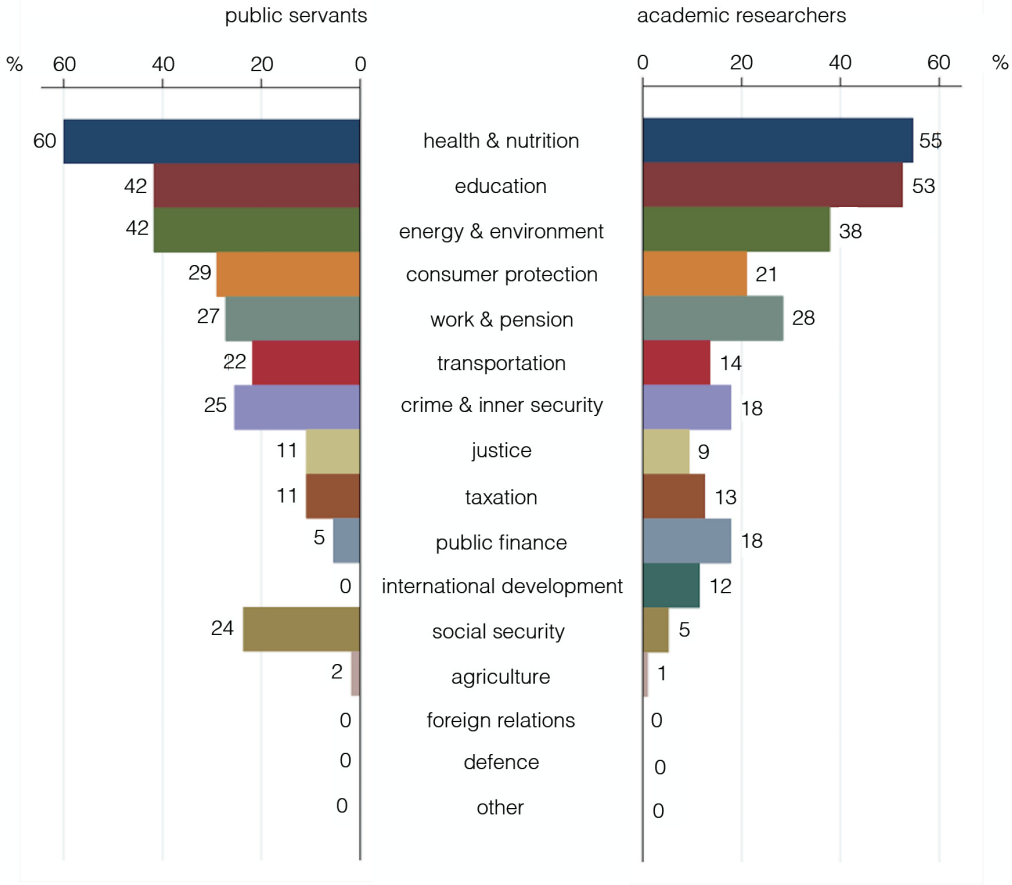


Note: Both figures depict the percentage of maximal points each variable received in the ranking. For each time being nominated on rank 1, an intervention received five points, for rank 2 it received four points, and so on. In the subsample public servants, the maximal sum of points from 11 valid responses was 55, while in the subsample academic researchers, the maximal sum of points from 19 valid responses was 95. Depicted numbers are rounded.

the pattern to occur the other way round: while academic researchers hope to get access to new types of (administrative) data through the cooperation with a public body, public servants are the ones working at the source. Yet it seems that it is the experimental approach which promises novel insights to public servants, an interpretation for which also the qualitative interviews provided anecdotal evidence (Adams (2020); Tangen (2020)). Meanwhile, academic researchers see a great advantage of collaborative experiments in new starting points for research; a motivation which intuitively is less relevant for public servants but fits the observation that in the majority of cases the public body had a strong part in deriving the research question.

Some interviewees also explicitly mention that the perspectives of academic researchers differ from those of the public body. "I have realized when collaborating with academics quite often, because interests are different, they might be more interested in testing a theory or in finding interesting

Figure 7: Ranking of policy fields – subsamples public servants and academic researchers



Note: Both figures depict the percentage of maximal points each variable received in the ranking. For each time being nominated on rank 1, a policy field received five points, for rank 2 it received four points, and so on. For the subsample of public servants, the maximal sum of points from 11 valid responses was 55, while for academic researchers, the maximal sum of points from 19 valid responses was 95. Depicted numbers are rounded.

results that can be published nicely. We got a few projects where we take more the research angle, but ultimately if I have to run the 50th social norms tax trial, it will probably not be intellectually stimulating but if I think that's actually the most effective thing, then that's ultimately what I would do", Ruth Persian (2020) describes how she at BIT incorporates her client's interests. Wilte Zijlstra (2020) from the internal behavioural insight team at the Authority for the Financial Markets in the Netherlands puts it this way: "We're not a university. Our main task is to promote fair and transparent financial markets. We do research together with academics. But research is not our only aim."

In sum, while the public body is interested in finding out *what works*, researchers moreover want to dive into the *whys* of the causal relationship (Czibor et al., 2019) in order to inform ongoing theoretical debates (Christensen & Miguel, 2018). In clinical research this distinction is used to distinguish *pragmatic trials* from *explanatory trials* with the former reportedly being most attractive for policy makers (Patsopoulos, 2011).

Table 5: Greatest advantage of collaborative research – subsamples

	% of academic researchers	% of public servants
increased relevance of research	66.7	45.5
access to new types of data	11.1	36.4
new starting points for research	16.7	9.1
Other	5.6	9.1

Notes: Values may sum up to more than 100% due to rounding.

4.3 Constraints within the public administration

When asked what reservations they met within the public body while pursuing an experimental test of interventions, the top answer, mentioned by every interviewee, was: *time constraints*. This comprises two aspects: First, experiments take time until they produce results, and second, they need a lot of time and effort invested by the people running them. Thomas Tangen (2020) from the Norwegian Tax Administration describes the tradeoff: "Often it is like 'we have a good idea, let's implement it'. So that's one of the main issues with academia, it takes so long time before you get the results." Paul Adams, formerly Financial Conduct Authority in the UK, confirms: "Within every organization there is a time pressure and people like to get things done quickly. Field experiments often take more time. That was often the main discussion point with our policy making colleagues."

Ruth Persian (2020) from the BIT made a similar observation: "I think for them it's about how much effort goes into everything. 'Why do we have to think about everything so many times? Why do we have to make sure everything is randomized perfectly, and then we have to go back to the data again because there was a mistake. . . ?' I think it's this planning, this being very very detail oriented upfront. While they would prefer just doing something, just sending something out and see." Since "testing takes more time and is not expected yet" (Maser, 2020), incentives for public servants are very low. Some would go so far as to just implement an intervention even though it could have easily been tested before. Even after the decision to test a new policy was made, the implementing staff may face competing priorities with respect to their day-to-day job and the new tasks the study brings about (Karlan & Appel, 2018).

Those public servants who bought into the idea of running a trial still *need to involve a lot of people* within their administration: their direct superiors and the colleagues who are affected by the implementation, the IT department, the legal department, the communications department, to name just a few. "Yes, it's a bit of bureaucracy. But it's also people's declarations. This is real. So you have to be very thorough. You have to do a proper job. You have to make it right", Thomas Tangen (2020) from the Norwegian Tax Administration describes. BIT-director of research, Alex Sutherland (2020), therefore sees a main challenge for external cooperation partners in making it maximally easy for the public partner: "The emphasis is really on the evaluator to reduce the burden as much as possible on those you are working with. For example, by relying on administrative data, data that is collected

anyway, rather than requiring a battery of 20 tests to be done for an outcome." Dina Pomeranz (2020) from Zurich University adds: "It is very helpful if there is at least someone within the partner organization that has as much excitement and interest in doing this study as the researchers. Then you become a team and can present this to the director of the institution, and help each other to make sure that it's academically sound but also interesting for the leaders of the organization."

Another substantial constraint for collaborative experiments seem to be *outdated IT systems within the administration*. Both interviewed public servants mentioned that external communication to citizens has to run via databases which are not suited for changing correspondence for different test groups (Tangen, 2020; Maser, 2020). This creates yet another bottleneck: "We've been working with the department staff, but they are reliant on our technology department to help them access their data and, when necessary, edit databases that auto-generate communications. And our technology department doesn't have time to help them or sometimes the databases are so old that it's difficult to try and change them. And obviously to try and run that experiment by having staff manually send out a letter would be prohibitively time consuming", Lindsey Maser (2020) from the City of Portland describes. "We have to think more like a web-administration", Thomas Tangen (2020) points towards a way into the future. "When you have an internet side, you are running experiments all the time. But on our systems, it's not like that. Our system is quite old. And security is important."

On a more principal level *ethical concerns* need to be addressed. "There is number of different concerns. First the general idea of experimenting on people but then also: Why are we preventing a certain group of the population from an intervention that we are thinking is beneficial?", Ruth Persian (2020), senior advisor at the BIT, recounts. Thomas Tangen (2020) from the Norwegian Tax Administration consents: "Experimenting with people's declaration, you have to have your ethics in place." Consequently, the public partner needs to get convinced that the experimental approach is ethical, legal, and that necessary insights cannot be derived by other econometric techniques (Gueron, 2017).

Doing that, researchers face a tendency of *risk-aversion* and "political scepticism" (Gueron, 2017, p. 7)) within the public body. "Public servants can lose a lot and do not have much to win. They can win for the country but not for themselves. They need to step out of their comfort zone a big way", Dan Ariely (2019), professor at Duke University, says. Dina Pomeranz (2020), professor at Zurich University, shows understanding for constraints resulting from that: "Responsible leaders of course have to be risk averse to some degree. If the payoff is small and it's not worth taking any risks for it, they are unlikely to agree to a collaboration. If knowing the answer to the research question is of value to them, they tend to be more willing to take the risk." Alex Sutherland (2020) from the BIT describes his experiences with risk aversion on side of the public partner like this: "It's the fear of the unknown. If people have done other kinds of evaluations like quasi-experimental designs, they may be familiar with those kinds of things. But changing something that's been done and planned, going against the status quo, that's more difficult. The process of change might feel alien to them."

A decisive role in this plays the *media*, several interviewees emphasize ((Ariely, 2019; Maser, 2020; Pomeranz, 2020; Tangen, 2020; Zijlstra, 2020)). Thomas Tangen (2020) from the Norwegian Tax

Authority personally had to deal with bad publicity after cooperating with academic researchers on a tax trial (published by Bott et al. (2019)): "It actually became a media issue afterwards. Because some lawyers argued that we treated people differently. Some taxpayers were told 'We know you have some money abroad' and others were told 'You should just declare everything', we did not say what we knew. That was a big discussion afterwards." Even though the Director General publicly defended the experiment and no legal issue followed, attitudes within the administration changed: "The notion was that we have to be very cautious, because at the end of the day we are dependent on people's trust. We still are doing experiments but we have to have that discussion every time" (Tangen, 2020).

Even with respect to solely internal communication, experiments might bring about *unwanted evidence*. "Science is kind of risky. You don't know what the answer is going to be. When you experiment, it might result in something that is contrary to what I would like it to be", Wilte Zijlstra (2020) from the Authority for the Financial Markets summarizes his colleagues' reservations. Alex Sutherland (2020) from the BIT consents: "It requires a great deal of faith from the delivery partner knowing that the end result of the evaluation could be that their intervention was not successful in improving outcomes. It's really hard if they have to sell that it hasn't really worked or – as worst outcome – that it made things worse." Yet researchers looking for collaboration can actively address this fear: "One aspect that they highlighted about why they accepted this proposal as opposed to some other ones is that it had a lot of safeguards about how we would avoid unintended outcomes. The proposal was careful to protect their institution from potential problems", Dina Pomeranz (2020) recounts her very first collaborative experiment with a public partner.

Another way out of this is "*building trust and reputation*", Dan Ariely (2019) from Duke University emphasizes. "Even with being well known as a researcher, I do so much free advice and trustbuilding. Normally, I give an introduction into behavioural economics, describe a small problem I am working on and sketch out a project that will come in 20 years. I call it lubricating the trust machine." Yet the public partner not only needs to build trust into the researcher but also into the intervention. For this, best practise examples from the public sector help, Lindsey Maser (2020) from the City of Portland emphasizes: "Since behavioural science and running RCTs is still fairly new in US government, and government tends to be very risk-averse, it's incredibly helpful to have examples of successful applications from other governments. It's easier to get approval to try something another government had success with than to try something that's never been done and might fail." Paul Adams (2020), formerly at the behavioural insights unit of the UK financial regulator, confirms: "You can then rely on external expertise to back your approach."

However, having won the trust of certain public servants does not mean the experiment will go along as planned. "One of the things that we faced is that there is very *high turnover internally*. We had to deal with different people on a frequent basis. So that made it a little bit difficult that some of the conversations which had gotten going then had to start again", Christian Gillitzer (2020) from Sydney University recounts. This could even lead to cancelling entire projects that were already agreed upon, Dina Pomeranz (2020) says: "When working with large institutions, changes in the environment or in the leadership are likely to occur. There can, for example, be an unexpected change

in the head of the authority who may not share that priority. So there is a substantial risk that in the middle of a research project somebody comes in and says: 'Actually, we're not going to finish the study'."

4.4 Discussion

In collaborative research, public servants apparently yield a great influence on the design of the experiment while they themselves face many constraints within their public body. Their influence bears risks and opportunities alike. On the upside, it ensures that the research endeavour is designed in a way that it will face substantial political interest and impact policy decisions one way or the other (Karlan & Appel, 2018) - a goal, which many researchers aim for. Moreover, since public partners are experts for their policy implementation processes, knowledge gaps identified by them bring new starting points for research.

On the downside, there is a substantial risk that collaborative studies are "carried out opportunistically" (Levitt & List (2009), p. 21) and under-research important aspects that are not considered relevant by the public partner. According to Deaton & Cartwright (2018), running an RCT does not require much prior knowledge and hence is suited to convince a distrustful audience. Yet, with respect to advancing knowledge accumulation, which is necessary to sustainably improve public policy, these kind of RCTs do not bring about much progress.

In addition to that, they give rise to *confirmation bias*, a phenomenon which is broadly discussed in hypothesis testing (see, for example, Oswald & Grosjean (2004) and McMillan & White (1993)). Confirmation bias occurs when the motivation for evaluating a program or a new policy is to produce external evidence for observations that have been already made. If this was the case, the research endeavour would be designed in a way to favour expectancy congruent information over incongruent information (Oswald & Grosjean, 2004) and not to search for the true underlying behavior. One way to avoid that is external quality control by independent researchers. This, however, can only be effective if the researchers are enabled to bring their full expertise to the table and do not have to compromise at too many stages. Several promising *ways forward* how this could be ensured are discussed in section 8.

5 Behavioural insight teams: caught between research and politics

The data from the quantitative survey reveals that behavioural insight team members experience collaborative experiments very differently from academic researchers. This section hence takes a closer look at their role.

5.1 Internal and external teams

Two different types of behavioural insight teams can be distinguished: i) internal units dedicated to applying behavioural insights within their organisation, and ii) external units like the BIT which

started as a British government institution and became a social purpose consulting company in 2014. It now has offices in London, Manchester, New York, Sydney, Singapore, and Wellington and works for public bodies all over the world.

Members of an internal behavioural insight unit have the great advantage that they know the institution from the inside, as Wilte Zijlstra (2020) from the behavioural insight team of the Dutch Authority of Financial Markets (AFM) points out: "You also have to deal with the realities of an institution. You know what's going to fly, what's not going to fly, what's feasible, what's not feasible. It's harder for an external consultant or an external researcher. Because I work at the AFM, I know better what my colleagues want and I know where the goals are." This view is confirmed by academic researcher Christian Gillitzer (2020) who worked with the behavioural insight team of the Australian Taxation Office (ATO): "They're much closer to the institution, they are ATO staff. Many of them have worked in operational parts of the organization themselves before coming to the behavioural insight team. I think they are more receptive to academic research than the rest of the organization. Their priority is demonstrating usefulness on a day to day basis to the ATO's activities."

The main challenge for internal behavioural insight units seem to be gaining reputation and support within their administration. "There's not yet much awareness at the higher levels of our city government of behavioural science and randomized control trials. Our efforts are initiated by staff rather than leadership so far. We're fortunate to have such engaged, interested employees, but in order to grow our behavioural insights and RCT efforts, we'll need someone in leadership to champion this work", Lindsey Maser (2020) from the City of Portland points out. "I think it's the role of the behavioural insight team internally to do some of that selling and convincing of the operational teams", Christian Gillitzer (2020) from the University of Sydney says. Yet even for very motivated behavioural insight units it seems to be a long way to go. "I would like it to be even more demand-driven, so that colleagues in supervision and policy would know better about us and approach us", Wilte Zijlstra (2020) from the internal behavioural insight team of the AFM says. Another interviewee adds: "It's often a case of individuals in an organisation who are willing to support a more evidence based approach."

External units like the BIT, on the other hand, are called to the table when the decision of potentially applying behavioural insights to policy design has already been made. "As a consultant, someone gives you a problem and you try to solve the problem with the tools you have. We are a research consultancy, both in terms of business model but also in the way we look at problems", Ruth Persian (2020) describes. Yet despite being an external partner, Alex Sutherland (2020) claims that there is a difference between working with the BIT and collaborating with academic researchers: "Because of our early ties to government, we still have good relationships with people within that space. This network, combined with the fact that many of our staff are previous civil servants, means that we can navigate the government landscape quickly and effectively." For the public body, also BIT's expertise and contacts seem to play a crucial role: "It's very hard to run a randomized controlled trial with a small sample size. BIT helped cities overcome this challenge by coordinating multi-city

Table 6: Risk-aversion and scientific standards – subsamples

	% of academic researchers	% of behavioural insight teams
Risk aversion in the public partner		
more frequent	12.5	13.3
less frequent	12.5	20.0
equally frequent	75.0	26.7
always high risk-aversion	0	40.0
never high risk-aversion	0	0
Moving away from ideal scientific approach		
never	10.0	13.3
in the minority of cases	40.0	20.0
50% of the time	30.0	26.7
in the majority of cases	10	33.3
always	10	33.3

Notes: Values may sum up to more than 100% due to rounding.

efforts. With this we could see trends and differences of what worked and what didn't among different populations", Lindsey Maser (2020) from the City of Portland says.

Many interviewees also point to the fact that communicating research results in a way that makes it accessible to a non-scientific audience is a core skill of a behavioural insight team. "It's knowing where to put the information you are producing. We definitely don't prioritize publishing journal articles but we do prioritize to make sure that whatever is produced is accessible to people who are not specialists, who are not researchers, and we make it as easy as possible for them to understand the implications of results that we find", Alex Sutherland (2020) says. It is therefore not surprising that three out of four interviewees working on behavioural insights within a public body held positions as communications professionals in their organisation before (Maser, Tangen, Zijlstra).

5.2 Boundaries of consultancy work

The anonymous survey of this study unveils that behavioural insight team members feel much more restrained by their public partners in collaborative experiments than academic researchers. Asked how frequently they had the impression their research opportunities were limited by a high degree of risk aversion in the public cooperation partner, 40% of behavioural insight team members indicate that this was *always* the case. Another 13% report that they had this problem more often than with other partners (see Table 6). In contrast, for academic researchers this aspect does not seem to pose a problem: only 12% indicate to have been limited by high risk aversion in the public partner more often than with other partners. 75% experience this equally often, and 12.5% even less frequent.

In addition to that, 67% of behavioural insight team members explicitly state that *in the majority of cases* or *always* they had to move away from an ideal scientific approach to accommodate the requirements of their cooperation partner (see Table 6). Most academic researchers, on the other hand, encountered such a pressure only *in the minority of cases* (40%) or *never* (10%); a mere 20% indicate that they experienced it *in the majority of cases* or *always*. One reason might be that behavioural insight team members cannot act as independently as academic researchers. The qualitative interviews point to the fact that members of behavioural insight teams are more likely to go along with the public body's constraints and implement mainly low-risk nudges. "At the start of the seven year period, we would say, that's fine, let's just do it anyway to get the experience and to try something out. But we sort of realised actually this was not a sensible way to do things", Paul Adams (2020) recounts his time in the behavioural insight team at the Financial Conduct Authority in the UK. "So our approach changed a little bit over time. I think early on we were happy to be more flexible." Another interviewee adds: "Obviously we cherish our autonomy. But I also know I might have to collaborate with this person again sometime."

On the other hand, Wilte Zijlstra (2020) from the behavioural insights team of the Dutch Authority for the Financial Markets, clearly sees himself as an advocat of his organization's interests: "When it gets too academic, I tell my colleagues at the behavioural insight team: Don't forget about the practical aspects. Think about what is relevant for supervision." In contrast, Ruth Persian (2020) from the BIT points out: "The advantage of being external is having a fresh pair of eyes. We can suggest things and get away with them, where civil servant might be a bit more hesitant because they will still be around at the end of the project." Yet she also is very aware of certain constraints: "We are a consultancy. So ultimately, if the government partner refuses or feels very uncomfortable with a certain design, then we might have to adapt our approach."

Behavioural insight team members, moreover, seem to be more willing to adjust to the time constraints set by the public partner. "One difference to academia is the timelines we are working on can often be very different. We might be trying to turn around a trial in a number of days or weeks rather than months or years", Alex Sutherland (2020) from the BIT describes.

5.3 Discussion

Behavioural insight teams take on a very special role in collaborative research. On the one hand, they are often embedded within the organisation, "know all the nuances" (Zijlstra, 2020) and can use that to apply their expertise on behavioural insights and experimental research very targeted to prevalent policy problems. On the other hand, behavioural insight team members are much more subject to the goals and constraints of the public administration than academic researchers. This is either the case because they are employees of the administration themselves, or they are paid via a consultancy contract.

In both positions - being an employee or a consultant - they experience firm boundaries around how much freedom they have to pursue their priorities and ideas if these are not fully aligned to those

of the public body. A systematically different research agenda might be the consequence. As Liam Delaney (2018) sketches out in his recent paper on the BIT: "There is a danger that behavioural insights trials will accumulate a large amount of local information on projects specifically selected for their suitability for treatment and with outcomes determined by local agency pressure." This study provides some first explorative evidence to support this apprehension.

6 Transparency and quality control in collaborative experiments

High-quality evidence is crucial in evidence-based policy advice, as discussed by Pomeranz (2017), Schmidt (2014), and Smets (2020). This section will hence take a closer look at how transparency and quality enhancing processes like pre-registry and publication, the comparison of short and long-term effects, and internal mechanisms of quality control are applied within the realm of collaborative experiments.

6.1 Pre-registry and publication

In the scientific community, publishing a pre-analysis plan for experimental research has become a highly recommended means of quality control. The main objectives of pre-specifying aspects like study design and hypotheses before running the experiment, and uploading this information to an external database are to prevent data mining and specification searching. Yet setting up a pre-analysis plan also helps researchers to finetune their analysis strategy and to increase credibility of research findings (World Bank, 2020).[13] Most popular in economic research is the platform AEA RCT registry which has been installed in 2013 and in recent year experienced a rising pace of registrations (Christensen & Miguel, 2018). As of 10 July 2020, more than 3,700 experiments in over 150 countries have been registered (AEA RCT registry, 2021).[14]

In collaborative experiments with a public partner, however, pre-registration does not seem to be a common way of quality control. Almost 40% of study respondents indicate that they have never pre-registered any of their collaborative experiments before (see Table 7). Another 15% did this less frequently than with other partners. Only 6% of respondents indicate that they always pre-register collaborative experiments. Interestingly, those respondents always pre-registering are not the academic researchers: none of them indicates to have done so. The majority of them rather states to pre-register experiments with public partners equally frequent than experiments with other partners, which apparently is not always the case. When considering the subsample of behavioural insight team members, one can see another interesting divide: Roughly half of the sample reports to pre-register

[13] For an overview of what should be included in a pre-analysis plan see World Bank (2020) and Christensen & Miguel (2018), p. 42.

[14] Other databases to pre-register are the Registry for International Development Impact Evaluations (RIDIE, http://ridie.3ieimpact.org) by the International Initiative for Impact Evaluation (3ie), the Experiments in Governance and Politics (EGAP) registry (http://egap.org/content/registration), and the Center for Open Science's Open Science Framework (OSF, http://osf.io).

their experiments more frequently than with other partners or to always pre-register experiments. The other half states that they never pre-registered collaborative experiments with public partners or to have done so less frequently than in experiments with other partners.

Table 7: Frequency of pre-registering

	% of full sample	**% of researchers**	**% of behavioural insight teams**
less frequent	15.2	11.1	5.9
equally frequent	39.4	55.6	35.3
more frequent	0	0	0
always pre-register	6.1	0	11.8
never pre-register	39.4	33.3	47.1

Notes: Respondents were asked how frequently they pre-registered their experiments with public partners compared to experiments with other partners. Values may sum up to more than 100% due to rounding.

According to the qualitative interviews, three main concerns are preventing researchers and behavioural insight team members from pre-registering: time pressure, confidentiality issues, and the aim of keeping experimental subjects and the media uniformed that a trial is underway (Adams, 2020; Persian, 2020; Sutherland, 2020; Tangen, 2020; Zijlstra, 2020). Wilte Zijlstra (2020) from the behavioural insight team at the Netherlandian Authority for Financial Conduct sees pros and cons: "Setting up a pre-analyis plan helps you with your design. But, again, it costs time. And you are operating under a deadline." For the work of the BIT, Ruth Persian (2020) explains: "We do have research protocols for every single experiment that we run, but usually we do not pre-register publicly. Often this is because of confidentiality issues with our partners. Social science registries are also quite a new development - but the importance of pre-registration is definitely something we are conscious of and are thinking about." She also thinks that more pre-analysis plans will be openly published in future: "To be honest, I think it's partly a resource problem that we don't do that by default. If we come up with a process, it's probably not that much work."

Yet it is not only lacking pre-registration that gives rise to concerns with regard to research transparency. According to the interviewees, a substantial number of trials with public partners does not even get published after completion. This is in line with findings of a recent meta-study by DellaVigna & Linos (2020). They document that as much as 90% of the trials conducted by the two largest Nudge Units in the United States have not been published to date, neither as working paper nor in any other academic publication format. According to Sanders et al. (2018), two different cases create problems with transparency: i) trials that are not published at all, producing a *public file drawer problem*, especially when null and negative results are selectively held back, and ii) trials that are published with insufficient details regarding their methodology, as the reliability of their results cannot be assessed appropriately.

Alex Sutherland (2020) describes the tradeoff for consultancies like the BIT: "We want to push for greater transparency in our work. We engage with people on this point quite frequently. Yet being a commercial research organization, the incentives are not towards publishing. The incentives are to get the job done." Ruth Persian (2020) from BIT moreover emphasizes the personal considerations: "That's all great if you actually get something out of the publication. And our publication on my CV is of course great. But it was also a lot of work that has to happen next to our day job."

Wilte Zijlstra (2020) from the behavioural insight team at the Netherlandian Authority for Financial Conduct also points to the resource tradeoff: "If you want to publish externally, you think about: how will this be interpreted and understood? So it's a cross-benefit assessment: how much extra time would it cost to get an external publication and is it worth the time? For reports internally, people know the context, you don't have to explain everything." He also sees the risk that some firms would use null or negative results for legal complaints against the regulator: "When you get court cases, they can use published reports against us. They would quote us: 'you're saying we should do this. But you're also saying it doesn't work'."

In contrast, for academic researchers like Dina Pomeranz (2020) from Zurich University publishing the results of an experiment is a non-negotiable prerequisite for any cooperation: "It is important to always set clear terms ex ante of what can get published, for example general results can be published, but no individual data. For scientific integrity it is important that the institution does not have a veto power at the end if the results are not what they hoped for." Christian Gillitzer (2020) from Sydney University, too, did not experience any constraints with respect to publication, just different priorities: "They're not primarily interested in making contributions to academic literature. Once the initial report was written with the findings, there was a conference call, and after that they briefed their senior people and then considered essentially the case closed. We contacted them with some follow up questions during the revision process to the paper, and they were receptive and helpful and wished us well. But they had gotten out of what they wanted to do and had moved on." From the partner's perspective a disengagement might be rational after having received their answers (Karlan & Appel, 2018). Yet in sum it contributes to a situation where publishing the findings of a collaborative experiments is not considered the default.

6.2 Short- and long-term effects

Another aspect of quality control is to check whether effects are sustainable over time by comparing short-term to medium or long-term effects. In collaborative research with public partners, this does not seem very common. The vast majority of study respondents (73%) indicates that the maximum time of observation to measure a (long term) effect in any of their collaborative experiments was 12 months (see Table 8). Only 9% measured an effect after more than 24 months.

This concentration on short term effects could be a phenomenon mirroring the relative young age of testing behavioural interventions with a public partner, as Dan Ariely (2019) points out: "We started with low-hanging fruits to show success. There was a focus on short-term effects. Now the

Table 8: Maximum period of observation

	frequency	**percent**
less than 4 hours	2	6.1
1-3 months	4	12.1
4-6 months	8	24.2
7-12 months	10	30.3
13-24 months	6	18.2
more than 24 months	3	9.1

Notes: Values may sum up to more than 100% due to rounding.

field will develop into longer term experiments." Yet long term studies also need a public partner to go along. While some researchers report the experience that policymakers find long run effects of great importance (Czibor et al., 2019), other document exactly the opposite (Sanders et al., 2018). This study contributes to the latter view by finding a high institutional impatience when it comes to the time span of research projects. Whether policy makers will truly provide enough administrative resources to measure long term effects, remains an open question. "I think it depends a lot on the person. The key is to find partners who share the interest in learning the answers. It also depends on the institutions. Some institutions have a tradition of research and innovation. Others have less of a culture of learning", Dina Pomeranz (2020) summarizes. It seems that for public bodies the same is true what Karlan & Appel (2018) document for organizations from the international development context: "The overarching lesson (...) is to choose carefully. Seek out partners who genuinely want to learn about their programs and products; who are ready, willing, and able to dedicate an appropriate amount of organizational capacity to research; and who are open to the possibility that not all answers will be rosy."

6.3 Internal quality control

Some institutions have installed their own processes of quality control. At the BIT, an internal research team keeps an overview of all research protocols and their proper set up (Persian, 2020). Being the Chief Scientist and Director of Research and Evaluation at the BIT, Alex Sutherland (2020) ensures the overall standards and quality: "We are trying to operate a similar sort of standards as external researchers. We have power protocols. We pre-specify. We have quality ensurance reviews of our analyses. We also often collaborate with universities who keep us accountable and shine a critical eye over our methods and tools."

Collaborating with external academics was also the way chosen by the internal behavioural insight team of the Financial Conduct Authority (FCA): "For all our research publications at the FCA, they had to be peer reviewed by an external academic to make sure that the methods were academically sound and rigorous. So all of the publications we put out had to have that external check", Paul

Adams (2020) recounts. The results would then be published in an FCA-own "Occasional Paper"-series (Financial Conduct Authority, 2020).

Public bodies, on the other hand, rely on external research consultants like the BIT for quality control: "In times, when less funding is available, we design and run the trial and BIT provides some advice along the way, and then reviews our analysis afterwards to confirm if we've done it correctly, or missed some deeper level findings", Lindsey Maser (2020) from the City of Portland describes.

6.4 Discussion

The best way of understanding what really works and what does not, is an accumulation of knowledge. It is hence a matter of concern that this study's finds that in collaborative research the majority of experiments seem to be neither pre-registered nor published after completion. In the academic literature, *publication bias* is a well known phenomenon: If researchers have a greater tendency to submit, and editors a greater tendency to publish studies with significant results, the publicly available evidence will be systematically skewed (Franco et al., 2014). Pomeranz & Vila-Belda (2019) hence strongly advocate to clearly define the scope of collaborative studies ex ante and to achieve an agreement with the cooperation partner that all results within that scope will be published, independently of the findings. A similar approach is taken by Karlan & Appel (2018) who recommend to set up a *Memorandum of Understanding* that codifies the agreement between the cooperation partners, among other things also regarding the right to share the findings. However, as this study documents, it's not that much a problem for academic researchers but rather behavioural insight team members refer to restrictions on publication of results. Contracts or internal constraints forbidding them to publish the data are mentioned as the most common reasons.

7 Recommendations

The following section presents some exploratory suggestions for ways to remedy some of the identified current weaknesses of collaborative research. Yet, as they have not been thoroughly checked for their feasibility and implications, their actual merit would have to be probed in the field.

7.1 Increase quality and quantity by external cooperation

Experiments with public partners can only truly improve policy making overall if their results are shared publicly. Best practice examples will help to convince more public bodies to test new policies before implementation. Null results, on the other hand, will help save public money (Sutherland, 2020). In order to achieve this, a promising way forward would be to *establish a new behavioural insights working paper series*. On a commonly shared platform, all behavioural insight teams could upload their reports. Working papers could take the form of policy reports which include a statistical

appendix for academic readers. Alternatively, a database with pre-determined categories[15] could be set up to be filled in by researchers and behavioural insight team members for trial documentation.

The Financial Conduct Authority in the UK, operating under time constraints as any other public authority, developed a best-practise procedure how to get external quality checks and publish all trial results without creating too much overhead for staff members: project leaders send out internal reports to academic researchers and incorporate their comments, mainly on what additional information should be included (Adams, 2020). Of course, this by no means replaces a full peer-review process for an academic journal. But it allows for some external quality control while providing the public body with a feasible way to make trial findings available to the public.

According to Wilte Zijlstra (2020) from the Netherlandian Financial Conduct Authority, such a *publication cooperation* could even go a step further: "You can align incentives for us with incentives for academia. Academics have incentives to publish. We have the data. If it gets published, it's more impactful." Dina Pomeranz (2020) from the University of Zurich agrees: "There are a lot of missed opportunities for collaboration where both parties would be excited to collaborate more. Students and researchers spend months writing theses that few people ever read. If we could have more of this research energy being channeled to answer questions that somebody really wants an answer to, that would be great." What is lacking so far is a *matchmaking platform for interested researchers and public bodies*. It could be attached to the new working paper platform. Academic researchers with interest in cooperation could set up a profile indicating their expertise and contact details. Public bodies, on the other hand, could upload questions they are interested in investigating and researchers could apply to them. Sole prerequisite would be that the platform is run by an institution or a group of individuals who really wants to make cooperation with the public sector happen and therefore takes care of promoting the platform and incorporating usability feedback.

With respect to pre-registering trials, a remedy for the popular concern that neither the public nor the media should be aware of a trial being under way is already provided by existing databases like the AEA RCT registry (2021) and the Center for Open Science (2020): They allow users to upload protocols and get a digital object identifier (DOI) immediately while public access can be embargoed for as long as four years. Knowledge about this possibility needs to spread more widely. If more *time stamped pre-analysis plans* were set up, trial quality would be likely to improve and chances for publishing the results in an academic journal increase, which in turn increases incentives for academic researchers to be part of the trial. Some journals even introduced "results-blind-review": a conditional acceptance based on a pre-analysis plan (Christensen & Miguel, 2018) to improve incentives for pre-specification even more.

[15]Categories could comprise, among others, a description of the intervention, target population, outcome of interest, sample size, observation period, effect size.

7.2 Facilitate processes within the public bodies

Three main obstacles seem to keep public bodies from testing new policies in randomized field experiments more often: i) the fear of the unknown, ii) technical infrastructure, and iii) time constraints. The first obstacle could be addressed by more best-practise examples from other public bodies being publicly available. It might also be a feasible way to promote *conducting pilot trials first* in order to increase trust into the intervention and decrease the risk of unintended side-effects (Pomeranz & Vila-Belda, 2019; Karlan & Appel, 2018). The second hurdle will vanish gradually when public administrations improve their IT systems in the process of becoming a digital public administration.

Time constraints, the third obstacle, mainly occur because public servants pursuing an experiment have to "set in the whole machinery" (Tangen, 2020) of involving many different departments and people. A promising way forward would be to *develop internal guidelines* within the public body specifying the authorized ways of how to conduct a trial and who need to be informed about it in which order. Moreover, templates for a memorandum of understanding which clarifies the role and responsibilities of each partner (Karlan & Appel, 2018) would help build confidence within the organization to enter new partnerships. Once there are best-practise-examples from other public bodies available, they will moreover provide orientation for new public bodies entering this realm.

Also *cooperation between public institutions* could bring a leap forward in collaborative experiments. "In my ideal world, it would be great if we had federal programs, state programs and city level programs that could coordinate and support one another", Lindsey Maser (2020) from the city of Portland says. "Because I hear sometimes from federal or state local governments that they're not doing as much direct interaction with residents, whereas at the city level we're often interaction directly with residents – paying water bills, parking fines, business licenses, etc. On the other hand, at the city level, our population is smaller, so our impact is smaller even if the work to design, implement and evaluate an intervention is the same."

Public bodies will also benefit from *installing an internal behavioural insight unit*, be it as small as one or two staff members. Such a unit would have two functions: First, keeping a good overview about the administration's work and act as an interjunction to academic researchers looking for collaboration. "The most fruitful thing is for them to have as wide a knowledge as possible so that they can filter the things that are most interesting academically" (Gillitzer, 2020). Second, those dedicated staff members can promote applying behavioural insights to policy design and running field tests of new policies. They can offer inhouse trainings for administrative staff and take part in team meetings of different divisions of the organisation in order to bring the behavioural insight perspective to the table.

If such an investment in additional staff members does not seem feasible (Maser, 2020), the move towards more evidence-based policy making could be *partly funded by non-profit organizations or private foundations* like the Bloomberg Philanthropies (2020). They started the "What works Cities"-initiative in 2015 to financially support cities' use of data and evidence. Alternatively, *a sunset clause with a profitability criterion* could be set up like it was done when establishing the BIT: The team

would have been shut down after two years if it had not, among other criteria, achieved at least a tenfold return on cost (Sanders et al., 2018).

7.3 Emancipate behavioural insight teams

More possibilities of applying for a *co-funding by foundations and non-profit organizations* would also benefit the necessary emancipation of behavioural insight team members from the strong influence of their public cooperation partner: "If you've got core funding, you've much more freedom to walk away from things. So if the Minister of Education in country X doesn't want to work with us, maybe then we go to country Y", Ruth Persian (2020) from the BIT describes.

In addition to that, *compulsory pre-analysis plans* would protect behavioural insight team members from a too strong own agenda of the public partner by clearly laying out the methodology to the research community. "One time we were being asked whether it was possible to shorten the timeline on a project. But we were also working with an academic partner. The academic was able to provide advice on how long the experiment needed to be in the field to be confident in the results. That helped the organization make an informed decision", Paul Adams (2020) remembers his experience as member of the behavioural insight team of the Financial Conduct Authority.

If a pre-analysis plan was set up and uploaded to an external platform, this would also provide strong arguments for publication of the findings, even if these turn out to be a null effect or negative. "I would love all senior policy makers to be open to field experiments that show null results or negative results. I think that was what is great about the FCA – they genuinely want to know what works. Senior policymakers are often judged by what they do rather than what they don't do. So it's very difficult to change a culture where for their next job interview, they're going to be asked: What did you implement? And then it's really hard to just say: oh, we spent two years investigating this and then decide that it's actually the wrong thing to do, so we are not going to do anything", Paul Adams (2020) says.

In general, the public should support a strong culture of transparency and against contracts which allow implementation partners to selectively hold back findings from publication. This might be achieved by *allowing exclusive "behind the scences"-reporting on behavioural insight team's work* by trusted journalists. Suitable candidates for this are journalists who themselves come with a strong background in econometrics and statistical inference; an expertise that nowadays is much more common at universities. Yet also in the research community more researchers should be comfortable to share "own juicy failures from which everyone can learn" (Karlan & Appel, 2018, p.136). Books like "Failing in the field" are paving the way.

8 Conclusion

Who and what drives experiments with public partners is an important question because policy decisions are based on the findings of these experiments. The present study is the first to empirically

investigate this topic. It analyzes a unique dataset with anonymously collected insights of public servants, behavioural insight team members and academic researchers, and combines these with in-depth expert interviews.

What becomes clear is that experimental research in cooperation with a public partner differs from other economic research in many respects. In particular, public servants exert a huge influence on study design and sample selection. At the same time, they have different priorities than academic researchers regarding the choice of policy fields and interventions to be tested in experiments. Together, this suggests that public servants shape the research agenda in a systematically different way than academic thinking would. Analogue to the literature about clinical research, field experiments with public partners could hence be classified as *pragmatic trials* (Patsopoulos, 2011).[16]

The strong influence of the public body can be both, an opportunity or a risk. As an opportunity, public servants open up new perspectives and shape the field of questions under investigation. Additionally, their investment ensures a high policy impact of the findings and because they know a lot about the context and the population the intervention is tested in they might also be the ones to uncover if something in the results is flawed (Cartwright, 2007). As risks, confirmation bias and a too narrow scope of collaborative research need to be taken into account. Given that quality control — as manifested in pre-analysis plans, publication, and medium and longterm effects — is reportedly low in collaborative research, these are reasons for concern.

As another major finding, the present study documents that behavioural insight team members act under a particular pressure. According to the anonymous survey of this study, an overwhelming majority of behavioural insight team members feels that they had to move away from ideal scientific standards in order to accommodate the requirements of the cooperation partner more than half of the time. They moreover experience much more often than academic researchers that risk aversion on the side of their public partners limits their research opportunities.

The main limitation of this study is that the findings are based on a small sample. Nevertheless, being the very first empirical research endeavour that investigates the current state of collaborative field experiments, this work represents the beginning rather than the end of a discussion. Its findings shall serve as an indication and stimulus for the reader where to dig deeper with future research. A first step into this direction was taken by discussing the patterns which emerged from the data with selected public servants, behavioural insight team members and academic researchers in in-depth interviews. While this, again, is based on a small sample of interview partners, it provides a first validity check of the descriptive results.

Some of the weaknesses in collaborative research which have been identified by this study can be addressed. Given the time constraints under which public bodies operate, one feasible option would be to align interests with academic researchers. Public bodies should increasingly allow access to their data and take academic researchers on board for trial design. A matchmaking platform for interested

[16]There are two categories of clinical trials: pragmatic trials evaluate the effectiveness of an intervention in real-life routine conditions, while explanatory trials test whether an interventions works under optimal conditions (Patsopoulos, 2011).

academics and public bodies would help facilitate cooperation. Moreover, a new working paper series for collaborative research projects could be established. This would allow to share knowledge among public bodies worldwide and to provide key statistical information for academic readers.

Additionally, behavioural insight teams should be emancipated. This could be achieved by applying for co-funding from foundations or non-profit-organizations. More funds with the aim to support evidence-based policymaking are needed. As another measure, behavioural insight teams could make use of time-stamped pre-analysis plans. These plans satisfy the requirement of confidentiality before the end of an experiment while providing a helpful means of quality control at the same time. In order to increase awareness for RCTs testing public policy and to promote a strong culture of transparency, one way is to allow trusted journalists access for exclusive behind-the-scences reporting.

All in all, more field experiments which test the application of behavioural insights to policy design could and should be conducted. Research interest is high, and many researchers are willing to invest time and labor in policy relevant field experiments. As this in turn will improve decision making in public bodies, there seems to be a win-win-situation. Given the findings of this study, it is just important to be aware of pitfalls and to make sure that structures are in place which do not allow any partner to systematically nudge the other into a particular direction.

Bibliography

Adams, P. (2020). Interview conducted on 1 July 2020. Unpublished transcript.

AEA RCT registry (2021). Registered trials. `https://www.socialscienceregistry.org/`. Last accessed: 2021-03-05.

Allcott, H. (2011). Social norms and energy conservation. *Journal of Public Economics*, *95*(9-10), 1082–1095.

Andor, M. A., & Fels, K. M. (2018). Behavioral economics and energy conservation–a systematic review of non-price interventions and their causal effects. *Ecological Economics*, *148*, 178–210.

Ariely, D. (2019). Interview conducted on 6 September 2019. Unpublished transcript.

Benartzi, S., Beshears, J., Milkman, K. L., Sunstein, C. R., Thaler, R. H., Shankar, M., Tucker-Ray, W., Congdon, W. J., & Galing, S. (2017). Should governments invest more in nudging? *Psychological Science*, *28*(8), 1041–1055.

Bloomberg Philanthropies (2020). About What Works Cities. `https://whatworkscities.bloomberg.org/about/`. Accessed: 2020-07-13.

Bobek, D. D., Hageman, A. M., & Kelliher, C. F. (2013). Analyzing the role of social norms in tax compliance behavior. *Journal of Business Ethics*, *115*(3), 451–468.

Bock, M. (1992). Das halbstrukturierte-leitfadenorientierte Tiefeninterview. In *Analyse verbaler Daten*, (pp. 90–109). Springer.

Bott, K. M., Cappelen, A. W., Sørensen, E. Ø., & Tungodden, B. (2019). You've got mail: A randomized field experiment on tax evasion. *Management Science*.

Cartwright, N. (2007). *Hunting causes and using them: Approaches in philosophy and economics*. Cambridge University Press.

Cartwright, N. (2010). What are randomised controlled trials good for? *Philosophical Studies*, *147*(1), 59.

Cartwright, N., & Hardie, J. (2012). *Evidence-based policy: A practical guide to doing it better*. Oxford University Press.

Center for Open Science (2020). Is my preregistration private? `https://www.cos.io/our-services/prereg/`. Accessed: 2020-07-13.

Christensen, G., & Miguel, E. (2018). Transparency, reproducibility, and the credibility of economics research. *Journal of Economic Literature*, *56*(3), 920–80.

Cialdini, R. B., Reno, R. R., & Kallgren, C. A. (1990). A focus theory of normative conduct: recycling the concept of norms to reduce littering in public places. *Journal of personality and social psychology*, *58*(6), 1015.

Czibor, E., Jimenez-Gomez, D., & List, J. A. (2019). The dozen things experimental economists should do (more of). *Southern Economic Journal*, *86*(2), 371–432.

Deaton, A., & Cartwright, N. (2018). Understanding and misunderstanding randomized controlled trials. *Social Science & Medicine*, *210*, 2–21.

Delaney, L. (2018). Behavioural Insights Team: ethical, professional and historical considerations. *Behavioural Public Policy*, *2*(2), 183–189.

DellaVigna, S. (2009). Psychology and economics: Evidence from the field. *Journal of Economic Literature*, *47*(2), 315–72.

DellaVigna, S., & Linos, E. (2020). RCTs to scale: comprehensive evidence from two nudge units. *Working Paper, UC Berkeley*.

DellaVigna, S., & Pope, D. (2018). What motivates effort? evidence and expert forecasts. *The Review of Economic Studies*, *85*(2), 1029–1069.

Dubner, S. (2017). "How to Launch a Behavior-Change Revolution" - Interview with Daniel Kahneman. Online Podcast Freakonomics.

Ebbecke, K. M. (2008). *Politics, Pilot Testing and the Power of Argument. How People's Feedback and the "Look, it is working"-Argument Help Policymakers to Communicate Controversial Reform Ideas*. Master's thesis, University of Dortmund.

Einfeld, C. (2019). Nudge and evidence based policy: fertile ground. *Evidence & Policy: A Journal of Research, Debate and Practice*, *15*(4), 509–524.

Fecher, B., Fräßdorf, M., & Wagner, G. G. (2016). Perceptions and practices of replication by social and behavioral scientists: Making replications a mandatory element of curricula would be useful. *DIW Berlin Discussion Paper*.

Ferraro, P. J., Miranda, J. J., & Price, M. K. (2011). The persistence of treatment effects with norm-based policy instruments: evidence from a randomized environmental policy experiment. *American Economic Review*, *101*(3), 318–22.

Financial Conduct Authority (2020). Occasional Papers. `https://www.fca.org.uk/publications/search-results?start=1&sort_by=dmetaZ&np_category=research-occasional%20papers`.

Franco, A., Malhotra, N., & Simonovits, G. (2014). Publication bias in the social sciences: Unlocking the file drawer. *Science*, *345*(6203), 1502–1505.

Frey, B. S., & Meier, S. (2004). Social comparisons and pro-social behavior: Testing" conditional cooperation" in a field experiment. *American Economic Review*, *94*(5), 1717–1722.

Gillitzer, C. (2020). Interview conducted on 13 July 2020. Unpublished transcript.

Gillitzer, C., & Sinning, M. (2020). Nudging businesses to pay their taxes: Does timing matter? *Journal of Economic Behavior & Organization*, *169*, 284–300.

Glennerster, R., Walsh, C., & Diaz-Martin, L. (2018). A practical guide to measuring women's and girls' empowerment in impact evaluations. *Gender Sector, Abdul Latif Jameel Poverty Action Lab*.

Gueron, J. M. (2017). The politics and practice of social experiments: Seeds of a revolution. In *Handbook of Economic Field Experiments*, vol. 1, (pp. 27–69). Elsevier.

Hallsworth, M. (2014). The use of field experiments to increase tax compliance. *Oxford Review of Economic Policy*, *30*(4), 658–679.

Hallsworth, M., List, J. A., Metcalfe, R. D., & Vlaev, I. (2017). The behavioralist as tax collector: Using natural field experiments to enhance tax compliance. *Journal of Public Economics*, *148*, 14–31.

Harrison, G. W. (2014). Cautionary notes on the use of field experiments to address policy issues. *Oxford Review of Economic Policy*, *30*(4), 753–763.

Hoffmeyer-Zlotnik, J. H. (1992). Einleitung: Handhabung verbaler Daten in der Sozialforschung. In J. H. Hoffmeyer-Zlotnik (Ed.) *Analyse verbaler Daten. Über den Umgang mit qualitativen Daten*, (pp. 1–8). Opladen: Westdt. Verl.

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS medicine*, *2*(8), e124.

Jachimowicz, J. M., Duncan, S., Weber, E. U., & Johnson, E. J. (2019). When and why defaults influence decisions: A meta-analysis of default effects. *Behavioural Public Policy*, *3*(2), 159–186.

Johnson, R. B., Onwuegbuzie, A. J., & Turner, L. A. (2007). Toward a definition of mixed methods research. *Journal of mixed methods research*, *1*(2), 112–133.

Karlan, D., & Appel, J. (2018). *Failing in the field: What we can learn when field research goes wrong*. Princeton University Press.

Levitt, S. D., & List, J. A. (2009). Field experiments in economics: The past, the present, and the future. *European Economic Review*, *53*(1), 1–18.

Lewin, K., et al. (1947). Group decision and social change. *Readings in social psychology*, *3*(1), 197–211.

Lewis, M. A., & Neighbors, C. (2006). Social norms approaches using descriptive drinking norms education: A review of the research on personalized normative feedback. *Journal of American College Health*, *54*(4), 213–218.

Madrian, B. C. (2014). Applying insights from behavioral economics to policy design. *Annu. Rev. Econ.*, *6*(1), 663–688.

Madrian, B. C., & Shea, D. F. (2001). The power of suggestion: Inertia in 401 (k) participation and savings behavior. *The Quarterly Journal of Economics*, *116*(4), 1149–1187.

Maser, L. (2020). Interview conducted on 2 July 2020. Unpublished transcript.

Mayer, H. O. (2006). *Interview und schriftliche Befragung*. München: R. Oldenbourg Verlag.

McMillan, J. J., & White, R. A. (1993). Auditors' belief revisions and evidence search: The effect of hypothesis frame, confirmation bias, and professional skepticism. *Accounting Review*, (pp. 443–465).

Miguel, E., Camerer, C., Casey, K., Cohen, J., Esterling, K. M., Gerber, A., Glennerster, R., Green, D. P., Humphreys, M., Imbens, G., et al. (2014). Promoting transparency in social science research. *Science*, *343*(6166), 30–31.

OECD (2020). Behavioural insights. `https://www.oecd.org/gov/regulatory-policy/behavioural-insights.htm`. Accessed: 2020-07-14.

Oswald, M. E., & Grosjean, S. (2004). Confirmation bias. In *Cognitive illusions: A handbook on fallacies and biases in thinking, judgement and memory*, vol. 79. Psychology Press.

Patsopoulos, N. A. (2011). A pragmatic view on pragmatic trials. *Dialogues in Clinical Neuroscience*, *13*(2), 217.

Persian, R. (2020). Interview conducted on 25 June 2020. Unpublished transcript.

Pomeranz, D. (2015). No taxation without information: Deterrence and self-enforcement in the value added tax. *American Economic Review*, *105*(8), 2539–69.

Pomeranz, D. (2017). Impact evaluation methods in public economics: A brief introduction to randomized evaluations and comparison with other methods. *Public Finance Review*, *45*(1), 10–43.

Pomeranz, D. (2020). Interview conducted on 14 July 2020. Unpublished transcript.

Pomeranz, D., & Vila-Belda, J. (2019). Taking state-capacity research to the field: Insights from collaborations with tax authorities. *Annual Review of Economics*, *11*, 755–781.

Pritchett, L., & Sandefur, J. (2014). Context matters for size: why external validity claims and development practice do not mix. *Journal of Globalization and Development*, *4*(2), 161–197.

Reisch, L. A., & Sunstein, C. R. (2016). Do europeans like nudges? *Judgment and Decision making*, *11*(4), 310–325.

Sanders, M., Snijders, V., & Hallsworth, M. (2018). Behavioural science and policy: where are we now and where are we going? *Behavioural Public Policy*, *2*(2), 144–167.

Schmidt, C. M. (2014). Wirkungstreffer erzielen - Die Rolle der evidenzbasierten Politikberatung in einer aufgeklärten Gesellschaft. *Perspektiven der Wirtschaftspolitik*, *15*(3), 219.

Smets, L. (2020). Supporting policy reform from the outside. *The World Bank Research Observer*, *35*(1), 19–43.

Sunstein, C. R., Reisch, L. A., & Rauber, J. (2018). A worldwide consensus on nudging? not quite, but almost. *Regulation & Governance*, *12*(1), 3–22.

Sutherland, A. (2020). Interview conducted on 29 June 2020. Unpublished transcript.

Tangen, T. (2020). Interview conducted on 30 June 2020. Unpublished transcript.

Thaler, R. H., & Sunstein, C. R. (2008). *Nudge: Improving decisions about health, wealth, and happiness*. Yale University Press.

Vivalt, E., & Coville, A. (2017). How do policymakers update? *Unpublished manuscript, Berkeley, CA: University of California, Berkeley*.

World Bank (2020). Pre-Analysis Plan. `https://dimewiki.worldbank.org/wiki/Pre-Analysis_Plan`. Accessed: 2020-07-09.

Zijlstra, W. (2020). Interview conducted on 8 July 2020. Unpublished transcript.

A Appendix: 5-Minute-Survey

5-Minute-Survey

Dear participant of the Behavioural Exchange 2019,

This is an anonymous survey about our common interest: research motivated by behavioral insights. My current study focusses on a review of experimental research conducted in cooperation with a public partner. I would highly appreciate if you take the time to answer this questionnaire. Please put the completed document into one of the boxes marked with "5-Minute-Survey" or hand it directly to me.

Thank you very much in advance!

Katja Fels
RWI - Leibniz Institute for Economic Research, Germany

1. In your view, what is the greatest advantage of cooperative research with a public partner?

- ☐ Increased political and practical relevance of research
- ☐ Access to new types of data
- ☐ New starting points for research (knowledge gap identified by practitioners)
- ☐ Other (please specify): _____

2. CONSIDER THE FOLLOWING FIELDS OF PUBLIC POLICY:
a) Which fields would you consider most relevant for conducting experiments testing behavioral insights?
b) In which fields of public policy did you test a behavioral insights intervention?

(Please choose your top 5, (a) starting with 1 for the highest priority, (b) starting with 1 for the field with the most experiments. If you haven't conducted any experiments with a public partner, please answer only a.)

FIELD OF PUBLIC POLICY	MOST RELEVANT (rank 1.-5.)	MOST OWN EXPERIMENTS (rank 1.-5.)
AGRICULTURE		
CONSUMER PROTECTION		
CRIME AND INNER SECURITY		
DEFENCE		
EDUCATION		
ENERGY AND ENVIRONMENT		
FOREIGN RELATIONS		
HEALTH AND NUTRITION		
INTERNATIONAL DEVELOPMENT		
JUSTICE		
PUBLIC FINANCE		
SOCIAL SECURITY		
TAXATION		
TRANSPORTATION		
WORK AND PENSION		
OTHER (please specify):		

5-Minute-Survey

3. CONSIDER THE FOLLOWING NUDGES:

a) If a public partner offered you to implement any intervention you like, which interventions would you consider most relevant to be tested?

b) Which of these interventions did you test in an experiment with a public partner?

(Please choose your top 5, (a) starting with 1 for the highest priority, (b) starting with 1 for the intervention with the most own experiments. If you haven't conducted any experiments with a public partner, please answer only a.)

TYPE OF INTERVENTION	MOST RELEVANT (rank 1.-5.)	MOST OWN EXPERIMENTS (rank 1.-5.)
COMMITMENT DEVICES		
DEFAULT RULES		
DISCLOSURE		
ELICITING IMPLEMENTATION INTENTIONS		
FEEDBACK		
INCREASE IN EASE AND CONVENIENCE		
LABELING (WARNINGS, GRAPHICS ETC.)		
LETTER DESIGN		
MONETARY INCENTIVES VERSUS NUDGES		
REMINDERS		
SIMPLIFICATION		
SOCIAL COMPARISON		
SOCIAL NORMS		
TIMING		
OTHER (please specify):		
OTHER (please specify):		

4. How many field experiments have you conducted in collaboration with a public partner?

- ☐ None (please move forward to question 12)
- ☐ 1-2
- ☐ 3-4
- ☐ 5-6
- ☐ More than 6

Public partners include:

- Government departments
- Government agencies and public bodies such as the Taxation Office, the Teaching Regulation Agency, or the Animal and Plant Health Agency
- Public institutions such as schools & universities

5. In which countries did your experimental research in cooperation with a public partner take place?
(If more than one, please rank the top five according to the number of experiments.)

1. _____
2. _____
3. _____
4. _____
5. _____

5-Minute-Survey

6. In this collaborative research, how was the research question derived?
(If you conducted more than one experiment with a public partner, please indicate the most frequent case.)

- ☐ A knowledge gap identified by the researcher/s was the starting point for developing the research question.
- ☐ A knowledge gap identified by the public partner was the starting point for developing the research question.
- ☐ Consultations between the researcher/s and the public partner were the starting point for developing the research question.

7. Who selected the sample?
(If you conducted more than one experiment with a public partner, please indicate the most frequent case.)

- ☐ The researcher/s were free to choose any sample from the target population.
- ☐ The researcher/s chose the experimental sample from a sub-population, which the public partner selected beforehand.
- ☐ The sample was chosen by the public partner.

8. In comparison to experiments with other partners, how frequently did you have the impression your research opportunities were limited by a high degree of risk aversion in your public cooperation partner?

- ☐ More frequent
- ☐ Less frequent
- ☐ Equally frequent
- ☐ I only conduct experiments with public partners and experience high risk aversion.
- ☐ I only conduct experiments with public partners and do not experience high risk aversion.

9. In comparison to experiments with other partners, how frequently did you register a pre-analysis plan for your experiment/s with a public partner (e.g. in the AER RCT registry)?

- ☐ More frequent
- ☐ Less frequent
- ☐ Equally frequent
- ☐ I only conduct experiments with public partners and pre-register them.
- ☐ I only conduct experiments with public partners and do not pre-register them.

10. When considering all your experiments, what was the maximum period of observation to measure a (long term) effect?

_____

11. How often did you have the impression you had to move away from an ideal scientific approach in order to accommodate the requirements of your cooperation partner?

- ☐ Never
- ☐ In the minority of cases
- ☐ 50 % of the time
- ☐ In the majority of cases
- ☐ Always

12. Which of the following applies to you?

- ☐ I am a researcher from a university or a public research institute collaborating with public partners in order to run experiments on public policy issues.
- ☐ I am an employee of a government department or of a behavioral insights unit.
- ☐ I am an employee or official from a privately financed institution running studies on behavioral insights.
- ☐ None of the above. I am _____

B Appendix: Interview guides

B.1 Questions for public servants

- Personal details
  - What is your professional background?
  - What are your personal experiences in experimental research between researchers and public collaboration partners?
- Motivation for cooperative research:
  - In your view, what are the advantages and disadvantages of collaborative studies between researchers and the public sector?
  - What do you, as a public servant, hope to get out from these studies?
- Influence of public servants:
  - In your experience, do you feel you have a "gatekeeper function" in collaborative studies?
  - Why would public servants be interested in having a strong influence on the design of studies?
  - Is this more a risk or an opportunity for the study? How often have you cancelled a collaborative experiment because of concerns?
    - What were the specific reason for this?
    - Do you feel you have the freedom to stop an experiment at any time?
- Pre-registry / refereed publication
  - How could high scientific standard be ensured in collaborative experiments?
  - One idea is to upload a pre-analysis plan before running the experiment, which specifies the research question, outcome variables and sometimes even hypotheses that are tested. Is this something you could imagine doing with future experiments?
  - Another idea is to require all experiments to be written up and submitted to a peer-reviewed journal. Which hurdles do keep the experimental partners from doing that and how could these be addressed?
- Ways forward Do you have any suggestions what would need to change structurally in order to improve collaborative research?

B.2 Questions for academic researchers

- Personal details

  - What is your professional background?
  - What are your personal experiences in experimental research with public collaboration partners?

- Motivation for cooperative research
  - In your view, what are the advantages and disadvantages of collaborative studies between researchers and the public sector?
  - What do you, as a researcher, hope to get out from these studies?
  - Do you think your priorities are different than those of the public body?

- Difference to other experimental research
  - When you compare the public sector to other collaboration partners in experimental research – are public collaboration partners different and if yes, in what way?

- Influence of public servants
  - In your experience, do you feel that public servants have a "gatekeeper function" in collaborative experiments?
  - Is this more a risk or an opportunity for the study?

- Difference between experiences of researchers and behavioural insight team members
  - How often have you cancelled a cooperative experiment because of concerns?
    * What were the specific reason for this?
    * Do you feel you have the freedom to stop an experiment at any time?
    * Is your position different to a member of a behavioural insight team?
    * What are the advantages/disadvantages of your role?

- Pre-registry / refereed publication
  - How could high scientific standard be ensured in collaborative experiments?
  - One idea is to upload a pre-analysis plan before running the experiment, which specifies the research question, outcome variables and sometimes even hypotheses that are tested. Is this something you could imagine doing with future experiments?
  - Another idea is to require all experiments to be written up and submitted to a peer-reviewed journal. Which hurdles do keep the experimental partners from doing that and how could these be addressed?

- Ways forward
  - Do you have any suggestions what would need to change structurally in order to improve collaborative research?

B.3 Questions for members of behavioural insight teams

- Personal Details
  - What is your professional background?
  - What are your personal experiences in experimental research with public cooperation partners?
- Motivation for collaborative research
  - In your view, what are the advantages and disadvantages of collaborative studies with the public sector?
  - What do you, as a behavioural insight team member, hope to get out from these studies?
- Influence of public servants
  - In your experience, do you feel that public servants have a "gatekeeper function" in collaborative experiments?
  - Is this more a risk or an opportunity for the study?
  - What are the main reservations of the public body you are facing when implementing an experiment?
- Caught between research and politics
  - What are the differences between you and an external researcher when you implement an experiment?
  - What advantages does your role bring about? What disadvantages?
  - How often have you cancelled a cooperative experiment because of concerns?
    - What were the specific reason for this?
    - Do you feel you have the freedom to stop an experiment at any time?
- Pre-registry / refereed publication
  - How could high scientific standard be ensured in cooperative experiments?
  - One idea is to upload a pre-analysis plan before running the experiment, which specifies the research question, outcome variables and sometimes even hypotheses that are tested. Is this something you could imagine doing with future experiments?
  - Another idea is to require all experiments to be written up and submitted to a peer-reviewed journal. Which hurdles do keep the experimental partners from doing that and how could these be addressed?
- Ways forward

– Do you have any suggestions what would need to change structurally in order to improve collaborative research?