Behavioural Public Policy (2022), 1–37
doi:10.1017/bpp.2022.14

CAMBRIDGE UNIVERSITY PRESS

ARTICLE

Who nudges whom? Expert opinions on behavioural field experiments with public partners

Katja Marie Fels

RWI - Leibniz-Institute for Economic Research, Ruhr University Bochum, Bochum, Germany
Email: katja.fels@rwi-essen.de

(Received 5 March 2021; revised 16 March 2022; accepted 30 March 2022)

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: First, public bodies have a considerable influence on study set-up and sample design. Second, high scientific standards are regularly not met in cooperative field experiments, mainly due to risk aversion in the public body. Third, 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.

Keywords: field experiments; behavioural public policy; Behavioural Insights Team (BIT); expert interviews; evidence-based policy advice

Introduction

Collaborative research projects involving policy makers and either academic or practical behavioural researchers are increasingly attracting attention. In 2010, 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 and 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 used for policy design (DellaVigna, 2009; Madrian, 2014) with applications ranging from automatic enrolment 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).[1]

This study presents the first empirical investigation into the current state of collaborative research. Throughout the article, 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. My empirical research endeavour has the objective to 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. Like Vivalt and Coville (2019), the author of this study made use of a unique opportunity to conduct a survey among conference participants. I chose the 'Behavioural Exchange 2019 (BX2019)', one of the largest conferences on behavioural insights worldwide. In addition to that, 12 in-depth interviews with selected experts allow for a more comprehensive interpretation of the patterns observed in the quantitative data. They moreover provide the basis to develop some practical recommendations of how to remedy the observed weaknesses. According to Mayer (2006), 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.

Based on the mixed-methods approach employed in this article, the study derives three main results. First, public employees 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 employees yield clearly different priorities than researchers. This has implications for the scope and focus of the research endeavour.

Second, this study reveals that the highest scientific standards are regularly not met in cooperative experiments. Individuals cooperating with a public body experience a high degree of risk aversion in their public partner. Tentative evidence indicates that behavioural insight team members more often adjust to the needs of the public body than academic researchers. Main reason for this seems to be that academic researchers have the freedom to call off an experiment if their requirements are not met whereas behavioural insight team members are bound by contracts.

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

https://doi.org/10.1017/bpp.2022.14 Published online by Cambridge University Press

Third, 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 article, 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 since data from the interviews suggests that scientific standards increase when academic researchers are involved as partners. 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 increase independence of public employees from political influence and a too strong agenda of the public institution.

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 and 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 cooperation 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.

Moreover, the article provides a contribution to the literature on which kind of evidence is used for policy advice (Sanderson, 2003; Sherman, 2003; Sutherland & Burgman, 2015; Head, 2016). Collaborative experiments between a public body and an external cooperation partner should be under special scrutiny since they are specifically designed to inform policy decisions. As Karlan and 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.' This study puts the topic of generating evidence-based policy recommendations by experiments involving a public partner on the agenda and documents that the evidence produced by these studies may systematically differ from other evidence available.

The remainder of the article is structured as follows: First, I outline the methodology of the quantitative and qualitative data collection. Then the first main result is presented by taking a closer look at the role of public employees as gatekeepers in collaborative research. The following section 'Scientific standards and risk aversion' reports data on the second main result regarding compliance with high scientific standards. Presenting the third main result, the next section focuses on transparency

https://doi.org/10.1017/bpp.2022.14 Published online by Cambridge University Press

and quality control in collaborative experiments. After discussing remedies for some of the perceived shortfalls, the final section concludes.

Data

The present study applies a mixed-methods approach, a class of research where the researcher combines quantitative and qualitative research techniques into a single study (Johnson & Onwuegbuzie, 2004) in order to ensure breadth and depth of understanding, and for corroboration (Johnson *et al.*, 2007). For this, I apply a two-stage explanatory design: the quantitative data informs the qualitative data selection process. This design has the advantage that observations from the quantitative data allow to specifically pinpoint relevant data in the qualitative data collection process (Almalki, 2016).

Quantitative survey

For the first stage in the mixed-methods approach, I designed an anonymous 5 min-questionnaire, which was answered by in total 70 participants of the 'Behavioural Exchange 2019 (BX2019)' – conference in London. Like Vivalt and Coville (2019), I made use of such a unique opportunity to conduct a survey among participants of a professional gathering. According to its organizers, at BX2019, about 1200 of 'the world's leading policy-makers, academics and practitioners gathered to explore new frontiers in behavioural science' (BIT, 2019).[2] The sampling of survey participants was conducted in a randomized manner. For this, I approached individuals at many different times and locations during the two conference days 5–6 September 2019 and distributed the pen-and-paper questionnaire in person. A potential sampling bias could arise if certain subgroups from my target population were not present among conference attendees (Krumpal, 2013). Yet as my study specifically focuses on academic researchers, behavioural insight team members and public servants, and all three groups are represented in the survey, this seems unlikely.[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. After the conference, the questionnaire was moreover sent out as online survey via www.onlineumfragen.com, a Swiss survey platform, in order to increase sample size.[4]

[2]No more detailed information about conference attendees is provided publicly.

[3]In the full sample, also private employees are included. They are a prevalent participant group at the BX2019 and hence also took part in the survey, since respondents were not pre-selected. However, with respect to the research question of this study, the three target groups are more relevant and are hence referred to throughout the article.

[4]All conference attendees who indicated their email address in the conference app were sent a personalized email (see Supplementary Appendix B) on 16 September 2019. The email asked to reply with 'YES' when the attendee was willing to participate in the survey, otherwise she would not be contacted again. In order to comply with German data protection law, a follow-up reminder to non-respondents was not feasible. Of 260 individuals contacted via email, 45 replied with 'YES' and received the link to the online survey. 27 of them used the link and answered the questions.

https://doi.org/10.1017/bpp.2022.14 Published online by Cambridge University Press

The questionnaire comprised 12 questions (see Supplementary Appendix A) and had been pilot-tested among researchers with experience in collaborative research. While questions 1 to 4 refer to general aspects of collaborative experiments which were posed to all survey participants, the rest of the survey focuses on own experiences. Consequently, the group of those who indicated in question 5 to have no own experience in collaborative research was at this stage directly navigated to the final question asking for their affiliation. During the analyses of those questions referring to own experiences in collaborative research, only respondents are included who have at least conducted 1–2 experiments themselves. For clarification regarding the respective numbers of relevant respondents, all tables presented in this article depict the absolute number of respondents next to the relative share.

Response rates to the survey were 28.3% (43 of 152 handed out survey forms) and 10.4% (27 of 260 individuals contacted via email), respectively. Unfortunately, I have no information about non-respondents. In case systematic differences between respondents and non-respondents occurred, the study's validity would be threatened by non-response bias (Krumpal, 2013). However, apart from time restrictions (which would affect all groups of conference attendees similarly), the most probable motivation to participate in the survey is that responders were particularly affected by the topic, while non-responders might not have seen much relevance of the questions for themselves. Since it is not the aim of the survey to provide a representative picture of opinions on collaborative research in general but rather to focus on personal experiences individuals made during their collaborative research, this potential self-selection does not seem to limit the data's scope to answer the research question. Moreover, to prevent social desirability bias, all survey questions have been thoroughly checked. None of them gives reason to suspect that 'due to self-presentation concerns, survey respondents underreport socially undesirable activities and overreport socially desirable ones' (Krumpal, 2013, p. 2025). By choosing two of the least intrusive data collection modes – a self-administered paper-and-pencil questionnaire during the conference and a web-survey afterwards – one of the main driving factors for social desirability bias was kept at the minimum.

Respondents' characteristics

This section presents respondents' characteristics from of the anonymous survey. Out of 70 survey respondents, 60 (86%) provided information about their affiliation (Table 1).[5] Some 32% of them are academic researchers, another 45% public employees – employed either in the capacity of a public servant or as a member of a behavioural insight team, and 23% work for a private company.[6]

[5]Two respondents classified themselves as academic researchers as well as public employees. In Table 1, they are listed as researchers. In the analysis part, they are included in both subsamples.

[6]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 (a fact that was included as a given in the answering options). Releasing this condition, I could classify them into one of the four categories listed in Table 1.

https://doi.org/10.1017/bpp.2022.14 Published online by Cambridge University Press

Table 1. Descriptive statistics

	Full sample	Academic researchers	Public employees	Private employees
Frequency	70	19 (31.7%)	27 (45.0%)	14 (23.3%)
No own experiments	23	9	7	6
1–2 own experiments	16	6	6	3
3–6 own experiments	15	3	6	4
More than 6 own experiments	10	1	7	1
No information provided	6	0	1	0

Note: Out of 70 respondents, only 60 provided information about their affiliation. The percentages in brackets refer to the share in this subgroup.

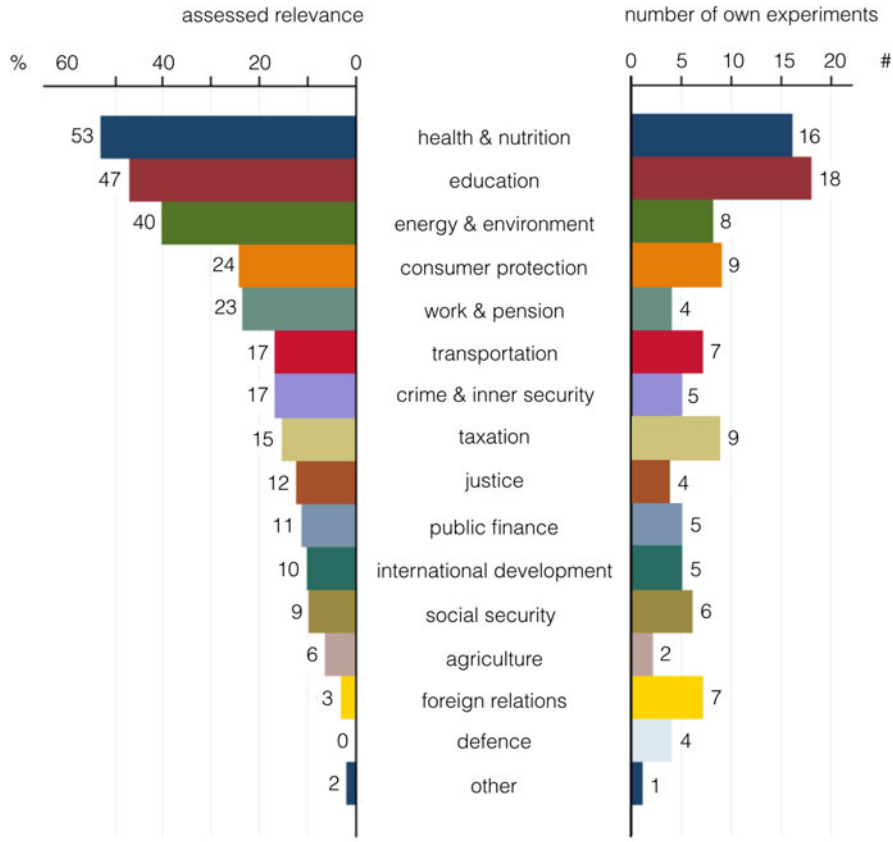
Table 1 also provides an overview of the number of experiments which respondents have conducted. It shows that among public employees a much smaller share than among academic researchers has no own experiences with experiments on behavioural insights. Almost half of the sample of public employees are considerably experienced: they have conducted at least three own experiments. This reflects that all employees from behavioural insight teams are included in this subgroup, and their job mainly consists of designing and implementing experiments. In addition to that, public employees most likely only attend an expert conference when it is practically relevant for their work, whereas for academics it is much more common to also choose a conference in order to get a good overview about current research on a topic even if oneself has not conducted own research in that field yet.

Most survey respondents indicate to have conducted experiments in Anglo-Saxon countries like the UK, the US, Australia, Canada or Ireland, with an overwhelming majority of experiments being located in the UK. This might be due to the fact that the BX2019-conference was organized by the BIT and many of the behavioural insight team members in my sample likely stem from the BIT. The second group of countries, in which experiments were conducted, includes 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 of own experiments, two fields clearly stand out: education (18) and health and nutrition (16) (see Figure 1, right column). In the second tier, taxation (9), consumer protection (9) and energy and environment (8) attracted frequent attention. The least often experiments were conducted in agriculture (2).

I also asked respondents to rank from 1 to 5 which policy field they consider most important to test behavioural insights. As the method of scaling, I chose the Borda count (Borda, 1784) which is applied worldwide in elections and sport competitions. This method is recommended for cases when one wants to produce a combined estimate of the best item, instead of discarding information by just counting which item was put on rank 1 most often. Under the Borda count, points are assigned to items

https://doi.org/10.1017/bpp.2022.14 Published online by Cambridge University Press



Notes: Left column: The figure depicts the percentage of maximal points each variable received in their ranking when asked which policy field is most relevant to test behavioural insights. The maximal sum of points from 70 valid responses was 350. Right column: The figure depicts how often respondents named the respective policy field in their top 5 of policy fields with own experiments.

Figure 1. Policy fields: relevance ranking and own experiments.

based on their ranking position: 1 point for last choice, 2 points for second-to-last choice and so on. Consequently, in my study, rank 1 was assigned 5 points, rank 2 received 4 points and so on. After the ranking, the point values from all survey participants were totalled, and the item with the largest point total was assigned the overall rank 1 of the full sample. As the last step, in order to make total points comparable between subgroups and the full sample, I standardized the number of total points by the maximum number of points that was achievable if every participant in the (sub) group had ranked this item on rank 1. The maximal sum of points from 70 valid responses in the full sample was 350.

Interestingly, for some policy fields, the relative frequency of own experiments stands in contrast to what respondents themselves assess as most relevant fields to test behavioural insights. While health and nutrition together with education make up the top 2 in both rankings, energy and environment is assessed as much more

https://doi.org/10.1017/bpp.2022.14 Published online by Cambridge University Press

relevant than it is mirrored by actual experiments. Especially when compared to the relevance assessment of consumer protection, which has the same number of own experiments but ranks much lower in the relevance assessment, energy and environment seems to be under-researched. The contrary is true for taxation and foreign relations. While they achieve the second tier of policy fields with most frequently conducted experiments, they have not been attributed much relevance in the assessment ranking. Based on relevance assessments, foreign relations score the next-to-last rank.[7] However, it is worth noticing that for some respondents, the formulation of the question might have installed a lower bound due to two reasons: First, respondents were asked to name their top 5 according to the frequency of own experiments. For those who conducted experiments in more than five policy fields, certain policy fields might be underrepresented in their answer. Second, if a respondent conducted more than one experiment in a certain field, it can still only be counted once since I have no data about the number of experiments in each respective policy field. The same is true for the question of which interventions have been tested.

For interventions tested in own experiments, a top 5 emerges (see Figure 2, right column): social norms (20), simplification (16), increase in ease and convenience (16), letter design (15), and – with a little distance – reminders (12). All these interventions 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 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, both in one trial (DellaVigna & Linos, 2020). However, when it comes to the relevance assessment under the Borda count, the respondents of this study clearly see default rules as the most important intervention to be tested. Defaults rank substantially before simplification, increase in ease and convenience, and social norms and can hence be considered under-researched with respect to the respondent's own priorities (see Figure 2, left column). Interestingly, the contrary is true for reminders and letter design: Both interventions do not achieve a high position in the assessment ranking but have been researched in own experiments quite frequently.[8]

Qualitative interviews

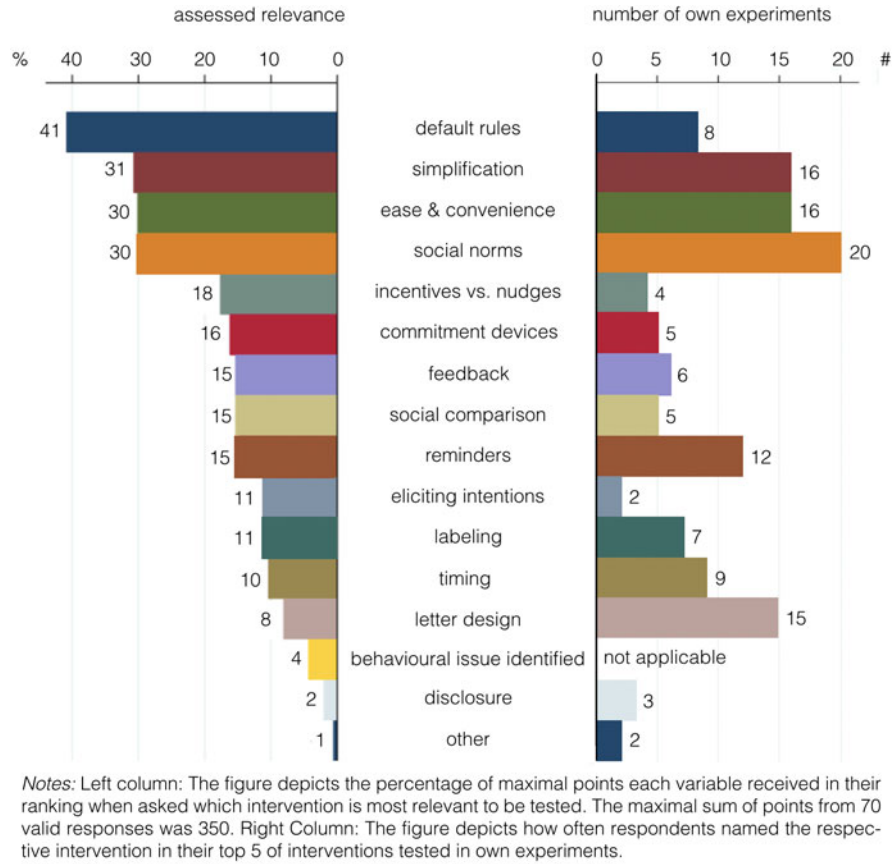
In the second stage of the mixed-methods approach, I collected insights from selected experts by semi-structured interviews. Such an approach is recommended when the study's aim is to gain sophisticated insights into aspects of social reality (Hoffmeyer-Zlotnik, 1992) and wants to capture nuances that cannot be detected by a quantitative approach (Glennerster *et al.*, 2018).

Non-probabilistic purposive sampling

The sampling followed a non-probabilistic purposive sampling approach: 'particular settings, persons or events are deliberately selected for the important information they

[7]For the relevance assessment differentiated by subgroups, see Figure 4.
[8]For the relevance assessment differentiated by subgroups, see Figure 3.

https://doi.org/10.1017/bpp.2022.14 Published online by Cambridge University Press



Notes: Left column: The figure depicts the percentage of maximal points each variable received in their ranking when asked which intervention is most relevant to be tested. The maximal sum of points from 70 valid responses was 350. Right Column: The figure depicts how often respondents named the respective intervention in their top 5 of interventions tested in own experiments.

Figure 2. Interventions: relevance ranking and own experiments.

can provide that cannot be gotten as well from other choices' (Maxwell, 2012, p. 235).[9] Since the aim was to fully understand the patterns discovered by the anonymous survey and to shed some light on the reasons and mechanisms behind the apparently different experiences of academic researchers, behavioural insight team members and public servants in collaborative research, I purposefully selected interview partners who satisfied the following criteria: (i) they belong to one of the three target groups and (ii) they have personal experience with at least one collaborative field experiment testing behavioural insights applied to policy design. With the exception of two non-responders and one refusal due to time reasons, all of the contacted experts either agreed to be interviewed or referred me to a colleague who would speak to me instead.

[9]As Morse (1991) points out, for most qualitative research a random sampling approach is not suitable since it both violates the quantitative principle of an adequate sample size in order to achieve representativeness and the qualitative principle of appropriateness achieved by purposeful selection of suitable study participants.

https://doi.org/10.1017/bpp.2022.14 Published online by Cambridge University Press

The final sample consists of 12 interview partners who belong in similar shares to the three target groups (see Table 7 in Supplementary Appendix C).

As recommended in the literature, I started the sampling inductively and increased sample size until theoretical saturation was achieved (Guest *et al.*, 2006). Theoretical saturation refers to the point at which additional data does not bring about new properties for a category. When a researcher sees similar observations over and over again, empirical confidence grows that this category is saturated (Glaser & Strauss, 1967). The observation that saturation occurs at a sample size of 12 is supported by an experimental test by Guest *et al.* (2006): They find that 'clearly, the full range of thematic discovery occurred almost completely within the first twelve interviews' (Guest *et al.*, 2006, p. 66).[10] In my mixed-methods study, the interviews provided the second stage of data collection to corroborate and explain findings discovered by the quantitative survey. The included interview partners, even though with different professional backgrounds, can hence be considered as a particular group of experts, namely individuals who gained personal experience in collaborative research.

Semi-structured interviews

For the interviews, I used a semi-structured approach: a blend of closed- and open-ended questions (Adams *et al.*, 2015), sometimes accompanied by ad-hoc follow-up questions to let the interviewee further elaborate on aspects that come up during the interview (Ebbecke, 2008). Such a flexibility is recommended in order to capture the full range of the topic (Bock, 1992). The questions were asked along one of three interview guides, each of which was specifically developed for the respective target group (researchers, behavioural insight team members, public servants) (see Supplementary Appendix D). To facilitate documentation, all interviews were recorded and transcribed. The interviews addressed the following topics: personal details, motivation for collaborative research, influence of public employees, the role of behavioural insight teams (only for researchers and behavioural insight team members), means of quality control and potential ways forward.

Thematic analysis

I analysed the interview transcripts by using a categorizing strategy and followed the six phases of thematic analysis as suggested by Braun and Clarke (2006): familiarizing oneself with the data, generating initial codes, searching for themes, reviewing themes, defining and naming themes and producing the report. Initial codes were the broad topics discovered by the survey data. As themes, I identified aspects that captured something important in the data in relation to the research question. A theme would ideally be mentioned several times across the dataset but did not have to in order to be relevant (Braun & Clarke, 2006). Goal of this procedure is to identify, analyse and report patterns across the dataset rather than within an individual interview. Categorizing helps to develop a general understanding of what is going on (Maxwell, 2012).

[10]Guest *et al.* (2006) interviewed in total 30 participants and documented changes to their codebook step-by-step. 73% of codes were identified within the first 6 interviews, about 92% of codes within the first 12 interviews.

https://doi.org/10.1017/bpp.2022.14 Published online by Cambridge University Press

The identified themes guided the write up of those parts of the study which refer to the qualitative interviews. As recommended by Braun and Clarke (2006), I chose vivid, compelling extract examples from each theme to be included in the article. All direct quotes have been authorized by the interviewees. Each of them consented to be named with full name and position.

For validation, the full article was sent to the interview partners after analysis and write-up was completed. None of them requested any changes. According to Maxwell (2012), such a respondent validation is the single most important way of ruling out misinterpretation as threat to validity.

Public employees as influential gatekeepers

In this section, the first main result is presented. The data of this study reveal that public employees yield a great influence on collaborative research, specifically on developing the research question and selecting the sample. At the same time, the study documents that public employees follow different priorities in collaborative research than academic researchers. This has implications for the scope and focus of the research endeavour.

Influence on study design

Development of the research question

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 2). In contrast, 40% reported 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 and Vila-Belda (2019) which focused 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.

Table 2. 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 researcher & public servant	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

Notes: Values may sum up to less or more than 100% due to rounding. Of the 70 survey participants, only those with own experiences in collaborative research were asked these questions.

https://doi.org/10.1017/bpp.2022.14 Published online by Cambridge University Press

All qualitative interviews 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.' This view is shared by Jaap Drooglever (2021) from the Ministry of Internal Affairs and Kingdom Relations: 'The municipalities decide in the end. They are responsible if something goes wrong, so that means that they look very closely on what the set-up is and if it fits their standards.' Academic researchers like Christian Gillitzer (2020) made similar experiences in collaborative research: 'This was more of a partnership where the ideas and proposals were directed by the Australian Taxation Office: they had a business need and something they wanted to be evaluated. Our scope and role were to refine and design the intervention such that it could be tested scientifically.' 'In my experience, it works best when they come to you and they want to do something', Johannes Haushofer (2021) from Stockholm University confirms. 'There is openness for suggestions, but it's within a pretty narrow parameter space.'

While many researchers in academia aim to influence public policy by publishing their study results, 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.

Yet, there also seems to be some scope for researchers to increase their 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.'

Selection of the sample

With respect to selecting the sample, the quantitative data again document a great influence of the public body. As Table 2 depicts, only one fifth of respondents (21%) state that the researchers were free to choose any sample from the target population. In contrast, public servants at least pre-selected the sample in about 80% of studies.

This is corroborated by some qualitative interviews. '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 incentive for the researchers to go in and 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 recounts his collaboration with the Australian Taxation Office. 'One of the dilemmas is that there is a trade off between really new designs to experiment and the scale at which you can do it and therefore the scientific control you have. If you do something really innovative, but it's only in one or two spaces, maybe that is how you make a big step from "how we did it" to "how we can do

https://doi.org/10.1017/bpp.2022.14 Published online by Cambridge University Press

it". But on the other hand, you don't have a lot of evidence supporting whether that was really effective', also Jaap Drooglever (2021) from the Ministry of Internal Affairs and Kingdom Relations reports.

On the administrative side, the interviewed public servants were despite their adherent interest and experiences in behavioural experiments very aware of their organization's limited expertise in experimental design and the current state of research. 'What we did at first was to realise that we were not the specialists, that we were interested in the field and can read about it. But I think it's important to say: Other people are the experts, and then to use that expertise', Jaap Drooglever (2021) summarizes. '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', Lindsey 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.'

Differing priorities

The great influence of the public body on study design might have meaningful consequences if priorities of public employees 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).

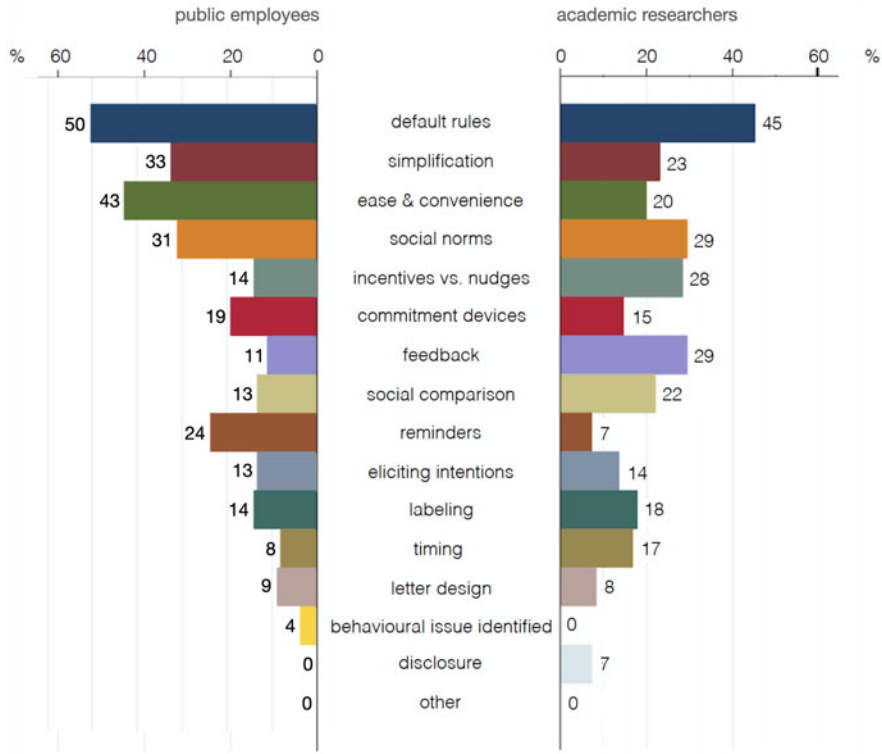
Interventions to be tested

This study shows that academic researchers and employees from a public body do yield such different priorities regarding the interventions to be tested (see Figure 3). While both groups agree that changing the default rule, one of the most effective but also most controversial nudges (see Reisch & Sunstein, 2016; Sunstein *et al.* 2018; Jachimowicz *et al.*, 2019), is the most important intervention, assessments differ for the other interventions in their respective top 5.

For public employees, 'increase in ease and convenience' is the second most important intervention (43% of maximal points), whereas among academic researchers this intervention does not even achieve their top 5. The same is true for reminders, which receive 24% of maximal points (rank 5) among public employees but score very low with academic researchers (7% of maximal points).

Academic researchers, on the other hand, assess several interventions as much more relevant than public employees: feedback (29% vs 11%) as well as 'financial incentives versus nudges' (28% vs 14%). Both interventions make it into the researchers' top 5 but are not ranked highly by public employees. In addition to that, timing (17% vs 8%) and disclosure (7% vs 0%) are considered more important by researchers than by public employees. Given the great influence of the public body on study design, this will likely lead to these interventions being under-researched in collaborative experiments. This interpretation was confirmed by several qualitative interviews. One interviewee put it in a nutshell: 'We did have early on a conference where we met

https://doi.org/10.1017/bpp.2022.14 Published online by Cambridge University Press



Notes: 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 of public employees, the maximal sum of points from 27 valid responses was 135. In the subsample of academic researchers, the maximal sum of points from 19 valid responses was 95. Depicted numbers are rounded.

Figure 3. Ranking of interventions – subsamples public employees and academic researchers.

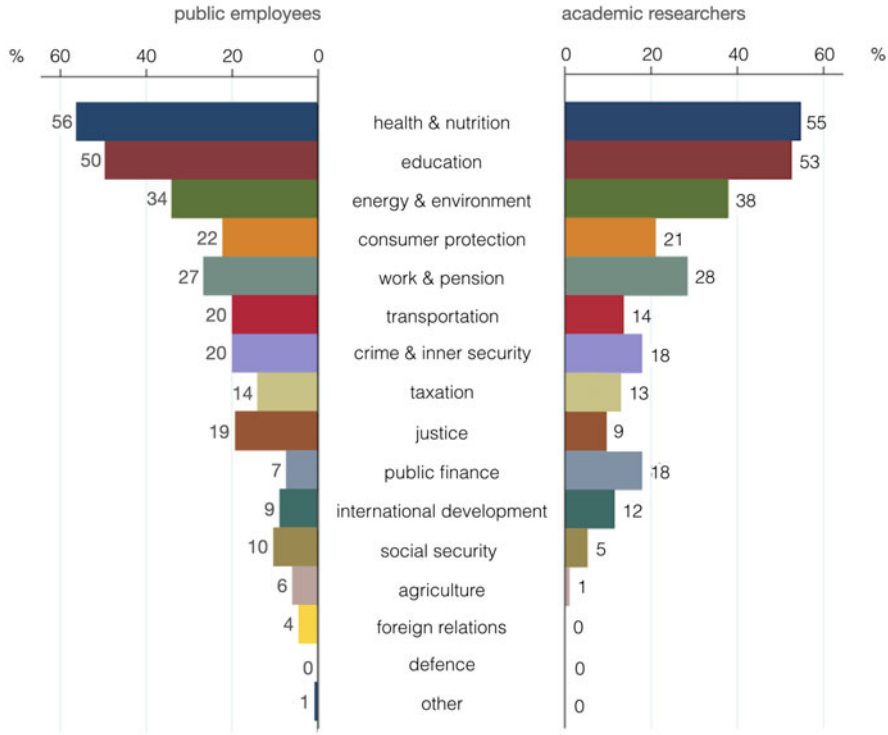
with many senior people from the public partner. There we came on with proposals. And they were receptive. But ultimately the things that went ahead were dictated primarily by their business needs.'[11]

Policy fields of interest

Interestingly, with respect to policy fields, less pronounced differences can be observed. Researchers and public employees overall agree that health and nutrition as well as education rank on the first two spots, followed with some distance by energy and environment, work and pension, and consumer protection (see Figure 4). This mirrors the pattern observed in the full sample with the sole modification that in the full sample data, consumer protection ranks before work and pension (see Figure 1, left column). When compared with the AEA RCT registry, the most popular database in economics to pre-register an RCT, one can also observe

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

https://doi.org/10.1017/bpp.2022.14 Published online by Cambridge University Press



Notes: 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. In the subsample of public employees, the maximal sum of points from 27 valid responses was 135. In the subsample of academic researchers, the maximal sum of points from 19 valid responses was 95. Depicted numbers are rounded.

Figure 4. Ranking of policy fields – subsamples public employees and academic researchers.

that the pattern closely fits those policy fields in which most RCTs have actually been conducted.[12]

Yet there are also some differences between the two groups. While public employees assess transportation as an important policy field, this field is not deemed very important by academic researchers (20% vs 14%). On the other hand, public finance (7% vs 18%) is considered relatively more important by academic researchers than by public employees.

Overall, there seems unity that defence, foreign relations and agriculture are not very relevant for testing the application of behavioural insights. However, in practice, respondents from the sample indicated to have conducted own experiments in these policy fields: foreign relations were nominated seven times, agriculture two times and defence at least once (see Figure 1).[13]

[12]For a more detailed comparison, see Supplementary Appendix E.

[13]Comparing the data on policy fields of own experiments with the relevance ranking of the corresponding subsample indicates that the ranking does not follow own experiences but shows a much more differentiated pattern.

https://doi.org/10.1017/bpp.2022.14 Published online by Cambridge University Press

Table 3. Greatest advantage of collaborative research

	% of full sample	% of academic researchers	% of public employees
Increased relevance of research	63.7 (44)	66.7 (12)	55.6 (15)
Access to new types of data	10.1 (7)	11.1 (2)	7.4 (2)
New starting points for research	15.9 (11)	16.7 (3)	25.9 (7)
Direct benefits to the wider public[a]	7.3 (5)	5.6 (1)	7.4 (2)
Others	2.9 (2)	0	3.7 (1)

Notes: The full sample also comprises of those respondents who did not indicate their affiliation. Values may sum up to less or more than 100% due to rounding. Absolute numbers in brackets.
[a]In the open text-field option, some respondents enlisted reasons that could be summarized into a new category: creating direct benefits of research for policy institutions and the wider public. This was not provided answering category in the questionnaire.

Motivation for collaborative research

Several interviewees explicitly mention that the perspectives of academic researchers differ from those of the public body while in the quantitative data differences in the motivation for collaborative research are not clearly observable (Table 3). This might be due to the fact that the in-depth interviews were able to uncover some aspects which were not listed as answering options in the survey. '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 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) from BIT explains. Johannes Haushofer (2021) describes the same phenomenon from the academic point of view: 'Public partners are often interested in a quick turnaround and quick results, and they're maybe less interested in understanding mechanisms. They just want to know if it works. As a researcher, you care a lot about why it works. You want a really clean design. That's often not a priority of the public partner.' Wilte Zijlstra (2020) from the internal behavioural insight team at the Authority for the Financial Markets in the Netherlands adds: '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, these categories are used to distinguish *pragmatic trials* from *explanatory trials* with the former reportedly being most attractive for policy makers (Patsopoulos, 2011). On the upside, the documented great influence of the public body ensures that the research endeavour is designed in a way that it will create

https://doi.org/10.1017/bpp.2022.14 Published online by Cambridge University Press

substantial political interest and impact policy decisions (Karlan & Appel, 2018). On the downside, it gives rise to *confirmation bias*.

This bias occurs when the motivation for evaluating a programme or a new policy is to produce external evidence for observations that have been already made.[14] 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 behaviour.

Constraints within the public administration

To gain a deeper understanding of the constraints that public employees are facing, the qualitative interviews also addressed the question what reservations interview partners met within their public administration while pursuing an experimental test of interventions. In the following, their answers are presented along the themes that occurred during the interviews.

Time constraints

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 trade-off: '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.' Johannes Haushofer (2021) from Stockholm University sees the reason in election cycles: 'Experimentation is often hard for policymakers because they face pressures that don't lend themselves to that kind of approach. They need to get reelected next year. And the three years it takes to run an RCT is too long for them and too risky.'

Ruth Persian (2020) from the BIT made the following observations: '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.' This view is shared by Helen Aki (2021) from the Ministry of Justice in New Zealand: 'People just trust the available knowledge and just want run with it. They don't want to invest the effort and the time that goes with having to set up trials and the extra work for people doing that. When you're actually running trials, sometimes it can take six months, 12 months to get a result. And that's often not good enough. People want to know now.'

[14]The phenomenon is broadly discussed in hypothesis testing (see, for example, McMillan & White, 1993; Oswald & Grosjean, 2004).

https://doi.org/10.1017/bpp.2022.14 Published online by Cambridge University Press

Since 'testing takes more time and is not expected yet' (Maser, 2020), incentives for public servants are 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). Helen Aki (2021) knows examples of that: 'Often the types of work that we do within the justice sector, it's all about trying to make it better for the defendant that's going through the process, or the victim. It doesn't actually make it better for the staff member. Sometimes it can make their job harder.'

Involvement of many different actors and departments

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) sees a main task of 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.'

High internal turnover

However, having won the trust of certain public servants does not mean that the experiment will go along as planned. 'You need to have a champion on the inside. That can be risky to the extent that if you don't know that person really well, they might change their mind or they might leave their job and then someone else comes in who doesn't have the same priorities', Johannes Haushofer (2021) from Stockholm University explains. Christian Gillitzer (2020) from Sydney University recounts similar experiences: '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.' This can 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".'

https://doi.org/10.1017/bpp.2022.14 Published online by Cambridge University Press

Outdated IT systems

Another substantial constraint for collaborative experiments seems to be outdated IT systems within the administration. Three out of four 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 (Maser, 2020; Tangen, 2020; Aki, 2021). 'Our systems are not set up in a way that enables good experimental design. So we want to change something but the IT system only allows one thing or another, so you can't set up RCTs', Helen Aki (2021) from the Ministry of Justice in New Zealand describes. This creates yet another internal 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 recounts. '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.'

Ethical concerns

On a more principal level, also 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. Johannes Haushofer (2021) from Stockholm University met similar reservations: 'Randomization is something that many policy makers shy away from, it seems unethical at first glance. Not knowing the researchers is another obstacle, if you don't know whether to trust these people and if they're going to deliver on the things that they promised.' 'For the municipalities, the practical result in the neighbourhood is the most important. These are their inhabitants. There are not guinea pigs. They are real people. So they have to be and want to be really respectful toward them', Jaap Drooglever (2021) from the Ministry of Internal Affairs and Kingdom Relations in the Netherlands says. 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).

Scientific standards and risk aversion

As second main finding, the data from the quantitative survey and the qualitative interviews reveal that the highest scientific standards are regularly not met in cooperative experiments. Individuals cooperating with a public body experience a high degree of risk aversion in their public partner. Especially members from behavioural insight teams seem to perceive a pressure to accommodate the needs of the public body, either due to employee or to contract obligations.

https://doi.org/10.1017/bpp.2022.14 Published online by Cambridge University Press

Scientific standards

The anonymous survey of this study unveils that in cooperative research the highest scientific standards are regularly not met. A majority of respondents indicates that they had to move away from an ideal scientific approach at least *50% of the time*. One quarter of respondents experienced this even *in the majority of cases* or *always* (see Table 4).

Moreover, there is tentative evidence that public employees, in particular behavioural insight team members, feel the pressure to accommodate the needs of their cooperation partner at the cost of high scientific standards more strongly than academic researchers. While the subsample of academic researchers is evenly split between those who report that they had to move away from an ideal scientific approach *never* or only *in the minority of cases*, and those who did this at least *50% of the time*, the proportions within the subsample of public employees are clearly different. More than 70% of public employees indicate to have experienced such a pressure at least *50% of the time*, while only 30% report this for *a minority of cases*

Table 4. Scientific standards and risk aversion

	% of full sample	% of academic researchers	% of public employees[a]
Moving away from an ideal scientific approach			
Never	11.8 (4)	10.0 (1)	12.5 (2)
In the minority of cases	23.5 (8)	40.0 (4)	18.8 (3)
50% of the time	38.2 (13)	30.0 (3)	37.5 (6)
In the majority of cases	23.5 (8)	10.0 (1)	33.3 (5)
Always	2.9 (1)	10.0 (1)	0
Experiencing risk aversion in the public body			
Never high risk-aversion[b]	3.0 (1)	0	0
Less frequent	24.2 (8)	12.5 (1)	25.0 (4)
Equally frequent	42.4 (14)	75.0 (6)	25.0 (4)
More frequent	12.1 (4)	12.5 (1)	12.5 (2)
Always high risk-aversion[b]	18.2 (6)	0	37.5 (6)

Notes: Respondents were asked (i) how often they had the impression they had to move away from an ideal scientific approach in order to accommodate the requirements of their cooperation partner, and (ii) how frequently they had the impression that their research opportunities were limited by a high degree of risk aversion in the public cooperation partner in comparison to experiments with other partners. The full sample comprises of 70 respondents, ten of whom did not indicate their affiliation. Only survey respondents with own experiences in collaborative research were asked these questions. Values may sum up to less or more than 100% due to rounding. Absolute numbers in brackets.
[a]In the online survey, respondents who indicated to be a public servant were not asked these questions. Answers in the subsample of public employees hence mainly stem from members of a behavioural insight unit.
[b]The answering options 'never' and 'always' include the statement that the respondent only conducted experiments with public partners.

https://doi.org/10.1017/bpp.2022.14 Published online by Cambridge University Press

or *never*. The answers of public employees to this question are mainly driven by members of behavioural insight teams, since a sample split in the online survey prevented respondents who had indicated that they were a public servant from getting asked that question at all.

For behavioural insight teams, two types 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 working for public bodies all over the world.[15] Since both types of behavioural insight units are on the payroll of the public body, they are more closely connected to its interests than external actors like academic researchers.

This observation was also confirmed by the qualitative interviews. Paul Adams (2020) recounts that during his time in the behavioural insight team at the Financial Conduct Authority in the UK they were likely to go along with the public body's constraints, for example by implementing 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. So our approach changed a little bit over time. I think early on we were happy to be more flexible.' Another interviewee of a behavioural insights unit, who wants this quote to be used anonymously, adds: 'Obviously we cherish our autonomy. But I also know I might have to collaborate with this person again sometime.'

Academic researchers, on the other hand, seem to experience more freedom when it comes to setting the terms and conditions of the research project. Dan Ariely (2019), founding member of the Center for Advanced Hindsight, says: '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.' A similar view is put forward by Dina Pomeranz (2020) from Zurich University: 'Basically, for quantitative research we need a certain number of observations. If they don't have that, we cannot do the study.' Also for Johannes Haushofer (2021) from Stockholm University, there are non-negotiables: 'If someone's not willing to randomize, mostly that's the end of the conversation for me.'

For behavioural insight teams like the BIT, it is more about explaining the statistical needs in plain language and gathering understanding: '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) summarizes.

When the interviewees were explicitly asked whether they were free to walk away from an experiment 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

[15]For a more detailed discussion of the two different types of behavioural insight teams, see Supplementary Appendix F.

https://doi.org/10.1017/bpp.2022.14 Published online by Cambridge University Press

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.'

Behavioural insight team members also 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.

Wilte Zijlstra (2020) from the behavioural insight team of the Dutch Authority for the Financial Markets, even sees himself as an advocate 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.' Jaap Drooglever (2021) from the Ministry of Internal Affairs and Kingdom Relations adds: 'Most of the municipalities are really far from thinking in the way the behavioural insight team does. These are two different worlds of thinking. But it can be addressed: by a lot of dialogue.'

Quite fittingly many interviewees mentioned 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 hence not surprising that three out of six interviewees working on behavioural insights within a public body held positions as communications professionals in their organisation before (Maser, Tangen, Zijlstra).

From a positive perspective, 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 servants 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.'

A systematically different research agenda of cooperative research could arise from that. As Delaney (2018) sketches out in his recent article 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.

Risk aversion of public bodies

As the most important reason for the pressure on adjusting scientific standards, the qualitative interviews identified a high degree of risk aversion in the public body, while the quantitative data on this aspect is not as clear cut. When respondents were asked how often they had the impression that their research opportunities were limited by a high

https://doi.org/10.1017/bpp.2022.14 Published online by Cambridge University Press

degree of risk aversion in their public partner in comparison to experiments with other partners, only a slight majority experienced this *more frequent* or *always* (see Table 4). However, the answering pattern might have been influenced by the formulation of the question which did not ask for the absolute frequency in which respondents experienced high risk aversion but for a comparison to experiments with other partners. If some respondents had only conducted experiments with public partners before and did not feel represented by the two answering options which incorporated that fact (namely: *never* and *always*), they would very likely be inclined to pick the answering option *equally frequent* since they have no comparison.

However, interestingly, again a difference in answering patterns of academic researchers and public employees can be observed. Even though subsample sizes are small, it is remarkable that the greatest share of public employees indicate that they *always* experienced high risk aversion. In contrast, not a single academic researcher chose that option. This might be due to the fact that, as discussed above, researchers feel free to walk away from a cooperation if their required standards are not met.

Notably, during the qualitative interviews, the topic of risk aversion in the public body was frequently raised. '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) from Duke University says. Dina Pomeranz (2020) from 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 adds: '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.'

Experiments might also 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.' A similar view is put forward by Johannes Haushofer (2021) from Stockholm University: 'For public servants, it takes a lot of courage to run a study. Because trials can produce bad results for your program, it can fall apart, and then you've spent a lot of money and nothing to show for it. It's a risky proposition. And so if you do it with partners that you know and trust, that can make it easier.'

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

https://doi.org/10.1017/bpp.2022.14 Published online by Cambridge University Press

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 trust building. 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 does not only need to build trust into the researcher but also into the intervention. For this, best practice 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.' A view that is also shared by Jaap Drooglever (2021) from the Ministry of Internal Affairs and Kingdom Relations in the Netherlands: 'Experiments give more arguments for the internal process to say: OK, we have to do it another way because the result from this experiment is that we have to redesign something.'

A decisive role in the process plays the media, several interviewees emphasized (Ariely, 2019; Maser, 2020; Pomeranz, 2020; Tangen, 2020; Zijlstra, 2020; Aki, 2021). 'As soon as anything gets out, you then tend to get a lot of questions and that creates a lot of work and essentially media requests. And so organisations in the public sector are generally relatively risk averse in terms of putting things up publicly', one interviewee describes. 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).

Transparency and quality control in collaborative experiments

Given the documented pre-selection of topics under investigation in collaborative research and the apparent pressures on research partners involved, it seems more important than ever for evidence-based policy advice that high quality evidence provides the groundwork to back up recommendations (Schmidt, 2014; Pomeranz, 2017; Smets, 2020). Yet, as third main finding, this study documents that transparency and quality control in collaborative research tends to be low as manifested in the application of preregistry, publication of results, and the comparison of short- and long-term effects.

https://doi.org/10.1017/bpp.2022.14 Published online by Cambridge University Press

Pre-registry

In the scientific community, publishing a pre-analysis plan for experimental research has become a highly recommended means of quality control (World Bank, 2020).[16] Yet in collaborative experiments with a public partner, it does not seem to be commonly used. In the anonymous survey, almost 45% of study respondents indicate that they have *never* pre-registered any of their collaborative experiments before (see Table 5). Another 14% did this *less frequently* than with other partners. For 18 of the 21 respondents, who chose either one of these two answers, the affiliation is known. The vast majority of them are public employees who have been shown to yield a great influence on the study design of cooperative experiments. It hence seems that another distinct feature of cooperative experiments is a tendency to not pre-register.[17] Interestingly, however, the small minority of respondents who states to *always* pre-register are not the academic researchers: none of them indicates to have done so in contrast to two public employees.

When comparing the answering options *less frequent*, *equally frequent* and *more frequent*, most respondents indicate that they pre-register cooperative experiments *equally frequent* (36%) or *less frequent* (14%) than experiments with other partners. None of them indicates that this was *more frequent* the case (0%). For those answering options, all answers can be accounted for by affiliation. The analysis shows that researchers and public employees contribute to the results in the same absolute numbers. It has to be noted, however, that among public employees almost exclusively members from behavioural insight teams are present, since for public servants these answering options were not available in the online survey.

Several qualitative interviews confirmed the tendency to not upload a pre-analysis plan for cooperative experiments. According to the interviewees, three main hurdles prevent from pre-registering their experiments at established platforms: time pressure, confidentiality issues and the aim of keeping experimental subjects and the media uninformed that a trial is underway (Adams, 2020; Persian, 2020; Sutherland, 2020; Tangen, 2020; Zijlstra, 2020; Aki, 2021; Drooglever, 2021). Wilte Zijlstra (2020) from the behavioural insight team at the Dutch Authority for Financial Conduct weighs the pros and cons: 'Setting up a pre-analysis plan helps you with your design. But, again, it costs time. And you are operating under a deadline.' 'We use our advisory board for the function of a pre-analysis check. We will consult them about the way we have designed experiments and whether they approve of it or not', Jaap Drooglever (2021) from the Dutch Ministry of Internal Affairs and Kingdom Relations describes. Yet, he also clarifies that they would not make anything public before the experiment starts: 'I don't think we will upload a pre-analysis plan publicly because I don't know how who will react. But there are no real barriers to do so.'

[16]For an overview of what a pre-analysis plan should comprise, see, for example, Christensen and Miguel (2018, p. 42).

[17]As one of the reviewers pointed out, cooperative experiments might also specifically attract researchers who have a preference for not pre-registering. Such a self-selection on part of the researchers, however, would only reinforce the prevalent tendency on side of the public body and hence sharpen this distinct feature of cooperative research.

https://doi.org/10.1017/bpp.2022.14 Published online by Cambridge University Press

Table 5. Frequency of pre-registering

	% of full sample	% of researchers	% of public employees	% of private employees
Never pre-register[a]	44.4 (16)	33.3 (3)	50.0 (8)	25.0 (2)
Less frequent	13.9 (5)	11.1 (1)	6.3 (1)	37.5 (3)
Equally frequent	36.1 (13)	55.6 (5)	31.3 (5)	37.5 (3)
More frequent	0	0	0	0
Always pre-register[a]	5.6 (2)	0	12.5 (2)	0

Notes: Respondents were asked how frequently they pre-registered their experiments with public partners compared to experiments with other partners. For public servants, the question in the online survey was slightly modified, namely: 'How frequently did you register a pre-analysis plan for your experiments (e.g., in the AER RCT registry)?' The answering options were: *never*, *in the minority of cases*, *50% of the time*, *in the majority of cases*, *always*. All three public servants picked *never* and are hence listed in this table in the corresponding category. The full sample also comprises of those respondents who did not indicate their affiliation. Values may sum up to less or more than 100% due to rounding. Of the 70 survey participants, only those with own experiences in collaborative research were asked these questions. Absolute numbers in brackets.

[a]The answering options 'never' and 'always' include the statement that the respondent only conducted experiments with public partners.

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.' Persian 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.'

Publication of results

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 finding is in line with the results of a recent meta-study by DellaVigna and 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 problems follow from this: (i) trials that are not published at all produce 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 prevent any quality control since the reliability of their results cannot be assessed appropriately.

'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', Alex Sutherland (2020) from the BIT describes the tradeoff. Ruth Persian (2020) also puts attention to personal factors: 'That's all great if you actually get something

https://doi.org/10.1017/bpp.2022.14 Published online by Cambridge University Press

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.' Helen Aki (2021), too, describes that she has to prioritize tasks given her limited time resources: 'We don't publish all the trials we conduct. And the honest reason is the time it takes to really write a report that is suitable for a public audience and get through the approval processes and everything else. Ultimately we're a government organisation and we're here to make change for New Zealanders. And so it's hard to make the time to do the kind of work for publishing.' Wilte Zijlstra (2020) from the behavioural insight team at the Dutch Authority for Financial Conduct made similar experiences: 'If you want to publish externally, you think about: how will this be interpreted and understood? So it's a cost-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".'

Jaap Drooglever (2021) from the Ministry of Internal Affairs and Kingdom Relations has quite the opposite impression. He perceives tight boundaries around the possibilities of holding results back from being made public: 'To keep certain information as classified, you need really good arguments, based on law. And I can't really see such arguments for this pilot, on the contrary: We want to share this way of working, and our results. So other than the privacy aspect of a pilot, there are no real barriers to be as transparent as we can.'

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). That's why researchers strongly advocate to clearly codify an agreement like a memorandum of understanding (MOU) with the cooperation partner beforehand that all findings can be shared publicly (Karlan & Appel, 2018; Pomeranz & Vila-Belda, 2019).

It is hence not surprising that 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 of the public partner: '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 public partner's perspective, the disengagement might be efficient and rational (Karlan & Appel, 2018). Yet in sum it contributes to a situation where the

https://doi.org/10.1017/bpp.2022.14 Published online by Cambridge University Press

Table 6. Maximum period of observation

	Frequency	Percent
Less than 4h	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 less or more than 100% due to rounding. Of the 70 survey participants, only those with own experiences in collaborative research were asked these questions. Absolute numbers in brackets.

publication of results from collaborative experiments is not considered the default and hence another essential requirement for high-quality evidence informing policy decisions is not met.

Short- and long-term effects

A third 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 means, too, does not seem to be commonly used. Quite the contrary: the vast majority of study respondents (73%) indicate that the maximum time of observation to measure a (long-term) effect in any of their collaborative experiments was 12 months (see Table 6). Only 9% measured an effect after more than 24 months. Even though these results do not provide evidence that the period of observation in cooperative studies is shorter than in other research, it does show that in the field of cooperative research another criterion expected from high quality evidence is not satisfied.

However, as the qualitative interviews point out, time might bring about some improvements. According to Dan Ariely (2019) from Duke University, the concentration on short-term effects could be a phenomenon mirroring the relative young age of testing behavioural interventions with a public partner: 'We started with low-hanging fruits to show success. There was a focus on short-term effects. Now the field will develop into longer term experiments.' Yet long-term studies also need a public partner to go along. While some researchers mention that policy makers find long run effects of great importance (Czibor *et al.*, 2019), others 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. Jaap Drooglever (2021) from the Ministry of Internal Affairs and Kingdom Relations made similar experiences: 'We asked those

https://doi.org/10.1017/bpp.2022.14 Published online by Cambridge University Press

municipalities, which we know are very open to really innovative ways of working. And that is important because you are doing something which isn't usual business. So it has to be conducted in an organisation where there's space for it.' It seems that for public bodies the same is true what Karlan and 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.'

Internal quality control

On the upside, the qualitative interviews also revealed that some institutions have installed their own processes of quality control. At the BIT, an internal research team reviews all research protocols and their proper set-up (Persian, 2020). Being the Chief Scientist and Director of Research and Evaluation of 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).

Other interviewees describe that their ministries installed advisory boards with external academics and stakeholders (Aki, 2021; Drooglever, 2021) or have contracts with external research consultants like the BIT for quality control (Maser, 2020; Aki, 2021). '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. That there are several feasible ways of quality control, is also acknowledged by Johannes Haushofer (2021) from Stockholm University: 'I think having academics involved is one approach of ensuring high scientific standards, or having a high quality outside organization involved, like an NGO that evaluates studies, or even a company.' However, what becomes clear is that it should always be made sure that experts from outside the public body are part of the quality control process.

Discussion and recommendations

The insights gained from the survey and the qualitative interviews suggest several possible ways to remedy some of the identified weaknesses of collaborative research. In

https://doi.org/10.1017/bpp.2022.14 Published online by Cambridge University Press

the following section, they are discussed in the light of which constraints could be addressed within the public administration and how increased transparency and external cooperation might provide means for improvement.

Alleviate constraints within the public administration

Public employees seem to face three main constraints when testing new policies in randomized field experiments: (i) fear of the unknown, (ii) technical infrastructure and (iii) time constraints. The first obstacle could be addressed by more best practice 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 (Karlan & Appel, 2018; Pomeranz & Vila-Belda, 2019). 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, partly 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; Haushofer, 2021) would help build confidence within the organization to enter new partnerships. Once there are best practice examples from other public bodies available, they will also provide orientation for new public bodies entering this realm.

In addition to that, *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.' An important factor could also be that *at a central level a coordinating role in the government* establishes testing as an integral part of policy making: 'It would be great if some administrative leads in the government would really champion the importance of testing and evaluation and trial design, so that it starts to become an expectation, such that from the senior leader level we expect to see this in policy proposals.'[18]

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

[18]The interviewee wants this quote to be cited anonymously.

https://doi.org/10.1017/bpp.2022.14 Published online by Cambridge University Press

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, 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).

More possibilities of applying for a co-funding by foundations and non-profit organizations would also benefit public employees like members from behavioural insight teams to become more independent from the interests and risk aversion of the public body: '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 researchers and public employees from a too strong political influence by clearly laying out the methodology to the research community beforehand. '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 (FCA).

If a pre-analysis plan was set up and uploaded to an external platform, this would also provide strong arguments for publishing 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.

Increase quality control by transparency and 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

https://doi.org/10.1017/bpp.2022.14 Published online by Cambridge University Press

upload their reports. Working papers could take the form of policy reports which include a statistical appendix for academic readers. Alternatively, a database with predetermined categories[19] could be set up to be filled in by researchers and behavioural insight team members for trial documentation. For those public bodies which still shy away from publicly sharing their data, a recent initiative by the OECD Expert Meeting on Behavioural Insights might be a good compromise: The initiative wants to develop an online pre-registration portal to share and collect case studies within the closed community of behavioural insight practitioners working in policy making. The goal is identical with the one put forward by this study: 'to implement standards of transparency on experiments, publication of all results, as well as on the quality of evidence used' (OECD, 2021, p. 6). The content of the portal shall enter into a continuous meta-analysis as well as into a power calculation tool which can improve calculation of sample sizes needed in the future.

Another best practice procedure how to get external quality checks and publish all trial results without creating too much overhead for staff members was developed by the Financial Conduct Authority in the UK: 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 Dutch 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.' Also, other public servants mentioned an openness of their institution for such cooperation (Aki, 2021; Drooglever, 2021). From the academic side, Dina Pomeranz (2020) from the University of Zurich sees a lot of potential in such an approach: '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. The OECD might be the right institution for this.

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

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

https://doi.org/10.1017/bpp.2022.14 Published online by Cambridge University Press

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 likely 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.

In general, the public should support a strong culture of transparency and oppose contracts which allow implementation partners to selectively hold back findings from publication. This might be achieved by *allowing exclusive 'behind the scences'-reporting on collaborative experiments* 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.

On a more general level, also a *better statistical education of the population* would help (see recent initiatives like the Data-Literacy-Charta in Germany (Schüller *et al.*, 2021)). As Johannes Haushofer (2021) puts it: 'I do think that more education about experimentation and statistics in school would be really helpful. We're learning that now with the COVID vaccines that a lot of the hesitancy around vaccinations comes from a lack of education. So partly that's scientific, partly that's statistical education. I think having more of that would lead to a greater willingness among policymakers to experiment, and to greater acceptance among the public.'

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 analyses a novel dataset with anonymously collected insights of public employees and academic researchers, and combines these with in-depth expert interviews with behavioural insight team members, public servants and academic researchers.

What becomes clear is that experimental research in cooperation with a public partner differs from other economic research in many respects. In particular, the public body exerts a huge influence on study design and sample selection. At the same time, public employees have different priorities than academic researchers regarding the choice of policy fields and interventions to be tested in experiments. Together, this suggests that public employees shape the research agenda in a systematically different way than academic thinking would.

The strong influence of the public body can be both, an opportunity and a risk. As opportunity, public employees open up new perspectives and shape the field of questions under investigation. Additionally, their investment ensures a high policy impact of the findings. Because they are experts about the context the intervention is tested

https://doi.org/10.1017/bpp.2022.14 Published online by Cambridge University Press

in, they might also be more able 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 long-term effects – is reportedly low in collaborative research, these are reasons for concern.

The present study also documents that the highest scientific standards are regularly not met in cooperative research. Main driving factor seems to be a high degree of risk aversion in the public body. Interestingly, tentative evidence indicates that members from behavioural insight teams experience more frequently the pressure to accommodate the needs of the public body than academic researchers. New structures should be put in place to prevent that. Potential remedies for some of the identified weaknesses of collaborative research could be an increased co-funding by foundations or non-profit organizations, the use of time-embargoed pre-analysis plans, a matchmaking platform for researchers and public bodies to facilitate cooperation, and a new working paper series. The latter should explicitly meet the needs of all partners in collaborative research by sharing knowledge in the form of policy briefs while at the same time providing sufficient statistical background information to allow for a scientific quality control.

Two limitations of this study are worth mentioning. First, the findings of the anonymous survey are based on a small sample. Second, since participants of the BX2019-conference could self-select into survey participation, and members from the British BIT were over-represented at that conference, the external validity of the study's findings is limited when it comes to an overall assessment of collaborative research. Yet being the very first empirical research endeavour that investigates the current state of collaborative field experiments at a meta-level, this work represents the beginning rather than the end of a discussion. The study's aim is to 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 quantitative data with purposefully 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 results.

In sum, 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 labour 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.

Supplementary material. To view supplementary material for this article, please visit https://doi.org/10.1017/bpp.2022.14.

Acknowledgements. My gratitude goes to the interviewees of this study: Paul Adams, Helen Aki, Dan Ariely, Jaap Drooglever, Christian Gillitzer, Johannes Haushofer, 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, Stephan Sommer and two anonymous referees for comments on study design or an earlier version of this article. Thank you to Alex Bartel for his reliable research assistance.

Funding. 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.

References

Adams, P. (2020), Interview Conducted on 1 July 2020. Unpublished Transcript.

Adams, W. C. (2015), 'Conducting Semi-Structured Interviews', in K. E. Newcomer, H. P. Hatry, and J. S. Wholey (eds), *Handbook of Practical Program Evaluation*, 4th edn, Hoboken, New Jersey, U.S.: Wiley Online Library, 492–505.

AEA RCT Registry (2021), Registered Trials. Retrieved from: https://www.socialscienceregistry.org/ (accessed on March 5, 2021).

Aki, H. (2021), Interview Conducted on 12 August 2021. Unpublished Transcript.

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

Almalki, S. (2016), 'Integrating quantitative and qualitative data in mixed methods research – challenges and benefits', *Journal of Education and Learning*, **5**(3): 288–296.

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

BIT (2019), Behavioural Exchange Conference 2019. Conference App.

Bloomberg Philanthropies (2020), About What Works Cities. Retrieved from: https://whatworkscities.bloomberg.org/about/ (accessed on July 3, 2020).

Bock, M. (1992), 'Das Halbstrukturierte-leitfadenorientierte Tiefeninterview', in J. H. P. Hoffmeyer-Zlotnik (ed.), *Analyse Verbaler Daten*, Wiesbaden: VS Verlag für Sozialwissenschaften, 90–109.

Borda, J. (1784), *Mémoire sur les élections au scrutin*. Paris: Comptes Rendus de l'Académie des Sciences.

Bott, K. M., A. W. Cappelen, E. Ø. Sørensen and B. Tungodden (2019), 'You've got mail: a randomized field experiment on tax evasion', *Management Science*, **66**(7): 2801–3294.

Braun, V. and V. Clarke (2006), 'Using thematic analysis in psychology', *Qualitative Research in Psychology*, **3**(2): 77–101.

Cartwright, N. (2007), *Hunting Causes and Using Them: Approaches in Philosophy and Economics*. Cambridge: Cambridge University Press.

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

Cartwright, N. and J. Hardie (2012), *Evidence-Based Policy: A Practical Guide to Doing It Better*. Oxford: Oxford University Press.

Center for Open Science (2020), Is My Preregistration Private? Retrieved from: https://www.cos.io/our-services/prereg/ (accessed on July 13, 2020).

Christensen, G. and E. Miguel (2018), 'Transparency, reproducibility, and the credibility of economics research', *Journal of Economic Literature*, **56**(3): 920–980.

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

Deaton, A. and N. Cartwright (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. and E. Linos (2020), RCTs to Scale: Comprehensive Evidence from Two Nudge Units. Working Paper, UC Berkeley.

https://doi.org/10.1017/bpp.2022.14 Published online by Cambridge University Press

Drooglever, J. (2021), Interview Conducted on 17 August 2021. Unpublished Transcript.
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.
Financial Conduct Authority (2020), Occasional Papers. Retrieved from: https://www.fca.org.uk/publications/search-resu
Franco, A., N. Malhotra and G. Simonovits (2014), 'Publication bias in the social sciences: unlocking the file drawer', *Science*, **345**(6203): 1502–1505.
Gillitzer, C. (2020), Interview Conducted on 13 July 2020. Unpublished Transcript.
Gillitzer, C. and M. Sinning (2020), 'Nudging businesses to pay their taxes: does timing matter?' *Journal of Economic Behavior & Organization*, **169**: 284–300.
Glaser, B. and A. Strauss (1967), *The Discovery of Grounded Theory: Strategies for Qualitative Research*. London: Aldine Publishing Company.
Glennerster, R., C. Walsh and L. Diaz-Martin (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 A. V. Banerjee, and E. Duflo (eds), *Handbook of Economic Field Experiments*, vol. **1**. Amsterdam: Elsevier, 27–69.
Guest, G., A. Bunce and L. Johnson (2006), 'How many interviews are enough? An experiment with data saturation and variability', *Field Methods*, **18**(1): 59–82.
Hallsworth, M. (2014), 'The use of field experiments to increase tax compliance', *Oxford Review of Economic Policy*, **30**(4): 658–679.
Hallsworth, M., J. A. List, R. D. Metcalfe and I. Vlaev (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.
Haushofer, J. (2021), Interview Conducted on 11 August 2021. Unpublished Transcript.
Head, B. W. (2016), 'Toward more "evidence-informed" policy making?' *Public Administration Review*, **76** (3): 472–484.
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*, Opladen: Westdt. Verl, 1–8.
Jachimowicz, J. M., S. Duncan, E. U. Weber and E. J. Johnson (2019), 'When and why defaults influence decisions: a meta-analysis of default effects', *Behavioural Public Policy*, **3**(2): 159–186.
Johnson, R. B. and A. J. Onwuegbuzie (2004), 'Mixed methods research: a research paradigm whose time has come', *Educational Researcher*, **33**(7): 14–26.
Johnson, R. B., A. J. Onwuegbuzie and L. A. Turner (2007), 'Toward a definition of mixed methods research', *Journal of Mixed Methods Research*, **1**(2): 112–133.
Karlan, D. and J. Appel (2018), *Failing in the Field: What We Can Learn When Field Research Goes Wrong*. Princeton, New Jersey: Princeton University Press.
Krumpal, I. (2013), 'Determinants of social desirability bias in sensitive surveys: a literature review', *Quality & Quantity*, **47**(4): 2025–2047.
Levitt, S. D. and J. A. List (2009), 'Field experiments in economics: the past, the present, and the future', *European Economic Review*, **53**(1): 1–18.
Madrian, B. C. (2014), 'Applying insights from behavioral economics to policy design', *Annual Review of Economics*, **6**(1): 663–688.
Madrian, B. C. and D. F. Shea (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.
Maxwell, J. A. (2012), *Qualitative Research Design: An Interactive Approach*. Newbury Park, California: Sage Publications.
Mayer, H. O. (2006), *Interview und schriftliche Befragung*. München: R. Oldenbourg Verlag.

https://doi.org/10.1017/bpp.2022.14 Published online by Cambridge University Press

McMillan, J. J. and R. A. White (1993), 'Auditors' belief revisions and evidence search: the effect of hypothesis frame, confirmation bias, and professional skepticism', *Accounting Review*, **68**(3): 443–465.
Morse, J. M. (1991), *Qualitative Nursing Research: A Contemporary Dialogue*. Newbury Park, California: Sage Publications.
OECD (2020), Behavioural Insights. Retrieved from: https://www.oecd.org/gov/regulatory-policy/behavioural-insights (accessed on July 14, 2020).
OECD (2021), OECD Expert Meeting on Behavioral Insights, 21 January 2021, Internal protocol.
Oswald, M. E. and S. Grosjean (2004), 'Confirmation Bias', in R. F. Pohl (ed.), *Cognitive Illusions: A Handbook on Fallacies and Biases in Thinking, Judgement and Memory*, vol. **79**, London: Psychology Press, 79–96.
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. and J. Vila-Belda (2019), 'Taking state-capacity research to the field: insights from collaborations with tax authorities', *Annual Review of Economics*, **11**: 755–781.
Pritchett, L. and J. Sandefur (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. and C. R. Sunstein (2016), 'Do Europeans like nudges?' *Judgment and Decision making*, **11**(4): 310–325.
Sanders, M., V. Snijders and M. Hallsworth (2018), 'Behavioural science and policy: where are we now and where are we going?' *Behavioural Public Policy*, **2**(2): 144–167.
Sanderson, I. (2003), 'Is it 'what works' that matters? Evaluation and evidence-based policy-making', *Research Papers in Education*, **18**(4): 331–345.
Schmidt, C. M. (2014), 'Wirkungstreffer erzielen - Die Rolle der evidenzbasierten Politikberatung in einer aufgeklärten Gesellschaft', *Perspektiven der Wirtschaftspolitik*, **15**(3): 219.
Schüller, K., H. Koch and F. Rampelt (2021), Data-Literacy-Charta. Retrieved from: https://www.stifterverband.org/charta-data-literacy (accessed on September 5, 2021).
Sherman, L. W. (2003), 'Misleading evidence and evidence-led policy: making social science more experimental', *The Annals of the American Academy of Political and Social Science*, **589**(1): 6–19.
Smets, L. (2020), 'Supporting policy reform from the outside', *The World Bank Research Observer*, **35**(1): 19–43.
Sunstein, C. R., L. A. Reisch and J. Rauber (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.
Sutherland, W. J. and M. Burgman (2015), 'Policy advice: use experts wisely', *Nature News*, **526**(7573): 317.
Tangen, T. (2020), Interview Conducted on 30 June 2020. Unpublished Transcript.
Thaler, R. H. and C. R. Sunstein (2008), *Nudge: Improving Decisions About Health, Wealth, and Happiness*. New Haven, Connecticut: Yale University Press.
Vivalt, E. and A. Coville (2019), *How Do Policymakers Update?* Berkeley, CA: University of California. mimeo.
World Bank (2020), Pre-Analysis Plan. Retrieved from: https://dimewiki.worldbank.org/wiki/Pre-Analysis-Plan (accessed on July 9, 2020).
Zijlstra, W. (2020), Interview Conducted on 8 July 2020. Unpublished Transcript.

Cite this article: Fels KM (2022). Who nudges whom? Expert opinions on behavioural field experiments with public partners. *Behavioural Public Policy* 1–37. https://doi.org/10.1017/bpp.2022.14