Essays in Decision-Making and Organizational Behaviour in Public Bureaucracies

Ranil Dissanayake

Pembroke

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy in Public Policy

Trinity 2022

Abstract

Public servants, no matter how carefully they are selected or incentivized, are only imperfectly able to discharge the objectives of the organizations they work for. They may suffer from bounded rationality, or make systematically biased decisions. They may be pursuing goals at odds with the stated or implied organizational mandate. Or they may suffer from low engagement in their work, and be difficult to retain over time. Through three quantitative essays, I examine how decisions are made in public sector organizations, and in particular how the behaviour of agents who vary in their ability, motivation and the incentives they are responding to contribute to public service (dys)function. In my first paper, a survey experiment in a large UK public sector organization, I find exposure to politician preferences (orthogonal to organizational mandate) increases the rate at which junior decision-makers provide advice contrary to the evidence they are provided with, but no such effect is found for senior officials. In my second paper, an observational study drawing on a large and novel dataset on projects implemented by the same organization, I show that a peer review process induced substantial avoidance of the system around the threshold for eligibility, but available metrics of project quality show only weak evidence of an effect around this threshold. And in my third paper, I use a survey experiment with a sample of public servants from 28 countries to investigate the effect of varying the framing of the gains to organizational initiatives on likelihood of participation, finding few overall effects but some heterogeneity according to pre-existing variation in public servant characteristics. These three papers have implications for how public organizations should structure their work.

Wordcount: 34,575

Acknowledgements

There have been times during my doctoral studies that I felt a little like Soares in The Book of Disquiet:

> "I will in the future... be living quietly in a little house in the suburbs, enjoying a peaceful existence not writing the book I'm not writing now, and so as to continue not doing so, I will come up with different excuses from the ones I use now to avoid actually confronting myself."[1]

In my defence, the excuses I've used recently have included a pandemic, marriage, the birth of a child, the dissolution of the institution I planned to conduct my research with and a new job. That I've persisted owes a lot to how much I've enjoyed the process of doing this research, my conviction that the work has some importance (no-one could sustain a doctoral effort without such a belief), and an extremely strong support network. My thanks are primarily due to the last, and in particular four.

My brother may feign sleep whenever I discuss economics, public policy or statistics, but whenever I've needed any help or support he's been there, without any questions or hesitation. It's not an exaggeration to say I would not have started this project if I didn't know I could depend on him, a personal backstop with a quick and biting sense of humour.

It's not unusual to thank one's supervisor for their guidance, technical support and so on. But in my case deeper thanks are due. In 2011 I sent Stefan Dercon an email asking him to look at a note I'd written for the Department for International Development, where we both worked at the time (he was the recently appointed Chief Economist, I had just joined the rank and file). He invited me to his office to discuss it, and what I expected to be 15 minutes of criticism turned into an hour of deep discussion both on my paper and economics and development more generally. It's hard to overstate how much that meeting changed the course of my career, and life. I went on to work for Stefan in his role as Chief Economist of DFID, and we have maintained a close working relationship since then. His support, mentorship and conversation have made me an infinitely better economist, thinker and public servant.

Finally: no-one manages to live with a doctoral student for any prolonged period of time

[1] From the Margaret Jull Costa translation of Fernando Pessoa's Book of Disquiet

without superhuman levels of patience and empathy. And they don't help that student through the inevitable roadblocks without wisdom, intelligence and love. I'm incredibly lucky that my wife has been with me through this process, providing encouragement and advice throughout. More than that, I'm lucky that she—and the son we had together two years ago–made my life so much fun and so exciting despite the ups and downs research always entails. For their companionship, love and support, I dedicate this work to Natalie and Luca.

Contents

1 Introduction

> Rationality of the sort described by the behavioural model doesn't optimize, of course. Nor does it even guarantee that our decisions will be consistent. As a matter of fact, it is very easy to show that the choices made by an organism having these characteristics will often depend on the order in which alternatives are presented... human reason is less a tool for modeling and predicting the general equilibrium of the whole world system... than it is a tool for exploring specific partial needs and problems.
>
> *Herbert Simon (1983)*

This thesis investigates the behaviour and decision-making of the agents who constitute public sector organizations affect their function, often in unexpected ways. It aims to further our understanding of behavioural decision-making in the public sector, and consequently, its performance.

Since at least the work of Herbert Simon (and specifically Administrative Behaviour, first published in 1945) researchers in the fields of public administration and organizational theory have considered how limits to human rationality and computational ability impose limits on the ability of the organizations they work in to optimize their decision-making, their function and ultimately their impact in the world. Concurrently, another literature spearheaded by James Buchanan and Gordon Tullock consider a more malign prospect: that rational organizations and individuals in them may make choices that are bad for society in pursuit of selfish aims. More recently—but still over half a century ago—Amos Tversky and Daniel Kahneman identified ways in which humans are not just computationally

limited, but systematically biased in the way they make decisions. That is, they make the same kinds of mistake repeatedly, an insight at the heart of behavioural economics. Each of these three ideas has sustained extended lines of academic inquiry about how organizations function: how they can overcome the limitations of human decision-making and the possibility that their agents do not always pursue a course of action in line with organizational objectives. Much of this work in organizational decision making, biases and the limits to rationality has been undertaken at the individual level, or private sector organizations. But the same issues are of first-order importance in public sector organizations, which often have to make decisions under conditions that seem tailor-made for mistakes: under intense time pressure, with often unclear objectives, when staff are over-worked and burnt out, and under intense political pressure. And recent research by Sheheryar Banuri, Stefan Dercon and Varun Gauri have demonstrated that even highly-selected, elite public servants suffer from all of the classic biases from the behavioural economics literature.

Yet most public sector organizations function, and often function rather well. How do they do so when they are made up of the same imperfect material that populates the pages of books and articles about bounded rationality, behavioural biases and malign incentives that maximise the personal at the expense of the public? I suggest that to understand this, it is necessary to understand the specific pressures that public servants are subject to, the systems of governance, decision-making and collaboration that public sector organizations use to try and improve their performance, the ways that public servants respond to these pressures and systems, expected and unexpected both. These can include the pressures of responding to political and Ministerial engagement and pressures, the subject of my first paper. Or they can be processes designed to check and re-check decisions, as in quality assurance peer review, the subject of my second paper. They may also be ways of engaging and connecting staff, as in the peer networks that are the subject of my third paper.

This focus on the ways in which the specific pressures and organizational processes and systems of governance that apply in public service builds on an impressive body of literature on Government and mission-driven organizations, elaborated in the next section. Much of this literature considers the role of motivation and incentivization of individual

workers in the performance and outputs of public sector organizations—considering how intrinsic motivation relating to the job or mission of the organization can motivate better performance among public sector workers, or how extrinsic motivation can be used to induce effort and incentivize the outcomes desired by the organization or Government. These forms of motivation are not uncomplicatedly positive, always inducing better performance or better decisions: the literature abounds in examples of motivating forces that produce outcomes at odds with the formal aims the organization has set. Another, expanding, arm of the literature considers how different systems of organization and governance can be applied to improve public sector outcomes. Much of this literature looks at management processes (including monitoring, target setting and the like), but some scholars investigate technologies, such as the use of checklists or algorithms to guide decision-making.

My three papers make a novel contribution to this broad literature. Paper 1 investigates how a common source of sub-optimal decision-making in public service, political pressure, varies with seniority within a public sector organization. Specifically, I use a survey experiment in a large, highly-respected and selectively staffed Government department responsible for making spending decisions of around £10 to £15 billion a year to investigate if civil servants responsible for a large (£35 million) project would spend it on a less effective option when their decision is complicated by knowledge that a political figure prefers that option (even though their task is to achieve the highest value-for-money in public spending). I then investigate heterogeneity by organizational seniority, finding evidence that more senior decision-makers behave substantially differently under these conditions than their junior colleagues, showing greater willingness to contradict their political leadership.

Paper 2 looks at another 'bureaucratic' adaptation which slows down the decision-making process, the use of a priori peer review and quality assurance of spending decisions, and investigates how it affects organizational function, in the same setting. In this case, I find rather different results: the institution of the assurance process led to significant reorganization of the portfolio of programmes as decision-makers within the organization sought to avoid it. What's more, though only imperfect measures of project quality are

available, they show weak evidence of a quality difference between reviewed and not-reviewed projects. Evaluating a system of governance designed to improve organizational function requires that we account for the behavioural response of public servants affected by it.

While the first two papers focus on systems and structures that are instituted by the organization and imposed upon staff through the formal rules and regulations of the organization or its way of working, my third paper considers a wholly voluntary activity. Peer to peer networks are common across the public sector, and are often promoted as contributing to organizational performance or career prospects, whether or not they actually achieve these ends. But what makes public servants decide to join such networks, given that they are time consuming and public servants already report being over-worked and underpaid for the time they spend on their work? Using a survey experiment with public servants from a number of countries and across a number of sectors, I show that framing the gains of participation in a peer network in terms of career benefits, benefits to end-users of public services or emotional satisfaction and happiness at work do not increase reported likelihood of participation across all public servants surveyed, but there is some evidence of heterogeneity: specifically, framings around personal happiness and emotional satisfaction reduce the reported likelihood of participation among those with high work engagement and low extrinsic motivation. This paper suggests that heterogeneity in the motivation and engagement of public servants may affect how they respond to alternative ways of framing organizational intiatives, though equally, in at least some circumstances, these framings will have no overall effect.

The next section surveys the literature in which I locate these findings, before a final section setting out the policy implications of this work.

1.1 Literature Survey

My work in this thesis contributes to an extensive literature on how public sector organizations function, including how public servants make decisions and how they are organized within a public bureaucracy.

The roots of this literature are deep: in Administrative Behaviour, Herbert Simon sets out

a theory of organizations that remains influential, emphasizing the psychological aspects of how they function, and come to decisions. In it he details both the practical limits to rational processing that all individuals must face (incompleteness of knowledge, difficulties in dealing with uncertainty and computational limitations), but also suggests that it would only rarely be in an organization's interests to fully combat these limitations through heavy investment in knowledge, analysis and the like (Simon 1997, pp. 93-95). Indeed, Charles Lindblom later argued that organizations instead 'muddle through' towards better outcomes through an iterative rather than optimizing process (Lindblom 1959). In later work, Simon added the existence of biases and faulty heuristics to this list of constraints to rationality (Simon 1983). Though he didn't explicitly make the link[1], the prospect of decision-makers systematically making the same mistakes also rather undermines the idea that 'muddling through' might always imply improvement. Accordingly, both he and other theorists considered ways in which organizations or collective action can—in Kenneth Arrow's words—"extend the domain of individual rationality" (Arrow 1974; Simon et al. 1991), or at least muddle through more effectively. Stylistically, we can divide the literature that emerges from these antecedents into two arms. One arm is about people, and the other about the structure of organizations and how they set up their work (this is a simplified version of the framework for understanding people management set out by Ali et al. 2021).

The first arm asks: who are the people who make up public bureaucracies? What motivates them? How able are they? How are they recruited? How do they perform once recruited? And how do these factors interact? One view is that workers (in all fields) are motivated by financial reward and career progression; that is they are extrinsically motivated. An extensive literature tests this idea both in private and public settings, but across settings, the results have been mixed. While Lazear found that a switch to performance-related pay increased performance in a private auto glass company, in experiments with students Gneezy and Rusticchini found that while within the set of students offered money as a reward for performance, performance was indeed increasing with the monetary reward offered, but those offered no reward outperformed those offered a payment (Lazear 2000;

[1] At least, not here

Gneezy and Rustichini 2000). These mixed findings extend to public sector studies. Some find substantial positive effects on performance (Gertler and Vermeersch 2012; Muralidharan and Sundararaman 2011), including when the source of extrinsic motivation is not pay per se but promotion (Karachiwalla and Park 2017) or recognition (Ashraf et al. 2014); others find no effect (Belle and Cantarelli 2015) and some find negative side effects (Weibel et al. 2009). Indeed, to the extent that extrinsic motivation may be pursued in ways orthogonal to the organizational mission—a view common in the literature which sees public service as a principal-agent problem which is too rarely adequately solved (Besley 2007; Buchanan and Tullock 1962)—its pursuit can lead to organizationally sub-optimal outcomes, as when regulators 'keep quiet' so as to avoid being seen to be wrong, harming future job prospects (Leaver 2009), or when bureaucrats discover that the optimal route to promotion is not excellence but political loyalty (Iyer and Mani 2012) or where there is no career return to the acquisition of organizationally valuable information and knowledge (Rogger and Somani 2018).

Another view on motivation in the public sector is that public servants are intrinsically motivated—they are driven to perform not by monetary or career rewards, or recognition, but by their desire to further the organizational mission, to increase public welfare or by the value of doing the job itself. Intrinsic motivation is multidimensional, with different authors meaning different things when they use the term (Ali et al. 2021), but in all forms the literature on its importance in the public sector has deep roots; indeed, Deci and Ryan point out that William James discussed the role of interest in motivation as early as 1890 (Deci 1975; Deci and Ryan 1985). Intrinsic motivation forms an important part of the theoretical literature on job match and optimal contract design (Acemoglu et al. 2007; Besley and Ghatak 2005; Dewatripont et al. 1999), as it does in the applied literature, as we will see below. Studies have found that intrinsic motivation matters for performance in a range of settings (Grant and Hofmann 2011; Kamenica 2012). Some forms of motivation appear in a 'grey area' between intrinsic and extrinsic motivation: in some cases studies where recognition for pro-social work motivates effort as in the aforementioned (Ashraf et al. 2014) experiment, it is classed as a point for intrinsic motivation. However, it should be noted in the widely used Work Preferences Inventory, measuring intrinsic and extrinsic

orientation of workers, peer recognition is counted as a form of extrinsic motivation (Amabile et al. 1994).

Building on these studies, an extensive body of work considers how, if workers can be motivated by intrinsic or extrinsic factors, they are optimally selected for ability and effort, and if these forms of motivation crowd each other out. Multiple studies have been concerned with how, at the margin, intrinsic and extrinsic motivation may be in conflict. The classic reference concerns the clash between intrinsic and extrinsic motivation among very young children (Lepper et al. 1973), but the same tension has been documented in public sector workers in a range of settings (Belle and Cantarelli 2015; Park and Word 2012). In a similar vein, an extensive literature considers whether appealing to extrinsic or intrinsic motivation in recruitment maximizes the ability of public sector workers recruited into the profession. In Zambia, a study of health workers found that making career incentives move salient at recruitment stage increased the performance of workers recruited into the health service (Bandiera and Lee 2015); what's more, any trade-off between the appeal to extrinsic motivation in recruitment and pro-sociality or intrinsic motivation in the job is apparent only among the least talented applicants; the marginal recruits are more talented and do not display lower pro-social attributes even when recruited through a process that stresses career opportunities (Ashraf et al. 2020). Similar results are observed in Mexico, where higher wages attract candidates with higher IQ, more desirable personality traits and stronger public service motivation (Dal Bo et al. 2013), though in an experiment with Indonesian students, Banuri and Keefer find that higher wages attract less pro-social applicants (Banuri and Keefer 2015a). It should also be noted, that for some functions in public bureaucracies (especially those Oliver Williamson classed as 'sovereign transactions') it may that ability is less important than other characteristics in recruitment, such as probity (Williamson 1999). More generally, the question of how to recruit for higher effort and better outcomes is increasingly well-tested in the experimental literature: in Indonesia again, pay-for-performance and pay-for-ability contracts perform similarly well (Banuri and Keefer 2015b), while a field experiment on the recruitment of teachers in Rwanda finds large effects on student outcomes from pay-for-performance schemes (Leaver et al. 2021).

While this work has been fruitful in setting out how public organizations can recruit the best individuals and get the most out of the individuals who make up their workforce, they are—except to the extent that compensation structure is an organizational characteristic, which it surely is—silent on how organizational structures and processes, including those of governance, can enhance the work of these individuals, holding constant their ability and the source of their motivation. The second broad arm of this literature considers this. A great deal of this literature is concerned with how real organizations are constructed, and these forms affect their functioning. In Complex Organizations, Charles Perrow summarizes an extensive literature on the merits of different organizational forms and structures, and what they mean for organizational function—though as often as not, these forms are not solely or even primarily aimed at the achievement of the explicit, formal objectives of the organization (Perrow 1979). Developing this general theme, a large and growing literature seeks to explain the quality of organizational performance by the managerial structures they use (Bloom and Van Reenen 2007, 2010; Gibbons and Henderson 2012). Managerial quality and structures has been found to explain organizational performance in the private sector (Bloom et al. 2013), in schools (Bloom et al. 2015a), in healthcare in developing countries (Dunsch et al. 2021) and in central government bureaucracies in multiple countries (Rasul and Rogger 2018; Rasul et al. 2017). In this literature the managerial structures and choices made by different forms is sometimes explicitly cast as a technology of production (Bloom et al. 2016).

But the idea that the structure or governance of organizational function or decision-making can improve organizational function has been applied beyond the merits of better managerial practices. In Cognition in the Wild, Edwin Hutchins's seminal study of organizational decision-making on a US Navy ship, he describes a very particular method of organizing the process of taking navigational readings that created what he termed 'cognitive redundancy' whereby information was recorded and processed by multiple agents to create a system which was resilient to costly error, a particularly important characteristic of a system in which decisions must often be taken under great pressure and at pace (Hutchins 1995). This approach created a fundamentally different way of generating knowledge to that which could be understood at the level of the individual

or even the team: "Humans create their cognitive powers by creating the environments in which they exercise those powers... [the] work organization [of]... the navigational team...shows why it is often difficult to apply the concepts that organize individual action to the organization of group action." (Hutchins 1995, p. xvi).

Cognitive redundancy is one organizational adaptation to the difficulties of decision-making; another is the use of organizationally mandated restrictions to individual choice and action. Such restrictions, in the form of checklists or algorithms have been proposed as a solution to biased or error-prone decision-making (Gawande 2010; Kahneman et al. 2019), and empirical studies in at least one healthcare setting have shown that they can work—provided the culture of the organization is receptive to such constraints (Martinez et al. 2015). But not all of the organizational choices made help effective organizational function or decision-making. Gibbons points out that the implication of Coase's seminal paper (Coase 1937) is that organizations can never truly become smoothly functioning optimizing machines of the type Weber imagined (Gibbons 2003). Instead, they may accrete into a system in which decision-making becomes a function of organizational dysfunction and interests, even if things are never quite as bad a Cohen, March and Olsen's apocalyptic vision of organizational decision-making as a 'garbage can' (Cohen et al. 1972; Gibbons 2003).

My research in this thesis attempts to contribute to, and bridge, these two arms of the literature. Each of my three papers examines the ways in which the behaviours and responses of the individuals who make up organizations-varied as they are-to the organizational and institutional systems and strictures they must operate within affect the performance of these organizations, for good and ill. In my first paper, I look at how different public servants respond to a common phenomenon in the policymaking process: the intervention of a political leader. I show how decision-making under these circumstances varies with seniority, and discuss how hierarchical structures in public services may fail to take advantage of the potential benefits of this difference. In my second paper, I show how an organizational adaptation nominally designed to improve decision-making quality (specifically project selection) caused bureaucrats to change how they operated, presumably to avoid additional scrutiny—without any clear effect on quality,

according to the metrics available. In my third paper, I look at things from the other side: what motivates individual public sector workers to take costly voluntary action in their work? How public servants, a heterogenous group respond to organizational initiatives is potentially an important factor in the impact these initiatives have, though in the example we study, we find few overall effects of framing. The next section briefly considers the policy implications of this work, which is further elaborated in the conclusion of the thesis.

1.2 Policy Implications

How well public sector organizations function is of first-order importance for public welfare, governing, as they do, so many aspects of our lives. In a recent essay, the public administration scholar Donald Moynihan wrote:

> The peculiar nature and constraints of government pose both a normative and design challenge that researchers cannot ignore. Public organizations lack market pressures for improvement. Democratic forces for change may miss the target... The applied study of government, properly understood, is able to not just bear witness to such problems but play a role in resolving them.[2]

My focus of research is aimed at both bearing witness to the problems of government and understanding how we can best resolve them. By understanding the specific interactions between the organizational choices and systems the institutions of public administration and policymaking adopt and the people who staff and make them up we can gain an insight not just into what helps public organizations function more or less effectively in pursuit of the social good, but how to improve further. It matters enormously whether organizational innovations and choices improve or impede the proper functioning of public administrations. It matters, too, whether this depends on which public servants these choices affect, or the balance of them in the organization. It matters enormously whether voluntary adaptions are attractive to overworked, underpaid and typically underappreciated staff. My research aims to shed light on each of these questions, making modest advances on what we already know.

There are three specific policy implications that arise from my thesis, which I will elaborate

[2] https://donmoynihan.substack.com/p/how-to-think-about-social-science

further in the conclusion, and one broad implication from the full sweep of the work.

First, the circumstances under which decision-making take place can significantly affect the decision-making of public servants, but these effects can be heterogeneous. We find senior decision-makers less likely to be swayed by political preferences that do not accord with the evidence available. This suggests organizational design that leverages differences in abilities under such pressure may improve function. And the recent tendency of political figures to force out senior civil servants for speaking truth to power undermines one of the more valuable adaptations the civil service has adopted.[3]

Secondly, innovations and processes designed to improve function must be tested and assessed, without assuming they will be implemented in the best possible manner, with the best possible intentions. The process of peer review and quality assurance I study in my second paper was a reasonable innovation to improve the decision-making of a department that suffers from cognitive and informational imperfections. However, its implementation provoked a behavioural response from the public servants working in the department: they very clearly adapted their work to avoid the scrutiny this process provided. Such gaming is not unusual: similar results have been observed across the UK Government in response to audit, and in procurement systems in the EU (Elston and Zhang 2022; Coviello et al. 2021). The review system might still be worthwhile if the behavioural response is sufficiently small compared to the benefit it brings in project quality; the available quality metrics provides only weak evidence that reviewed projects were of higher quality, and this might be due to either selection or the causal impact of review. The policy implication, that such systems should be tested, and organization should invest in the information required to do so, is clear but not always welcome to such organizations.

Thirdly, when organizational adaptations and systems are both voluntary and often costly to the public servants working in them, it may matter how their benefits are framed. Though in our investigation, there are overall effects of framing for only a subset of public servants, we find differential responsiveness to specific framings according to pre-existing characteristics of respondents.

[3]See, for example, the recent sacking of Tom Scholar as Permanent Secretary of the Treasury: https://www.theguardian.com/politics/2022/sep/14/kwasi-kwarteng-sacking-tom-scholar-marks-shift-away-from-impartial-advice

Across all of these papers stretches one major policy implication: the choices made in structuring the work and workforce in public administrations have profound implications for their function. Public services and public organizations may be inevitably imperfect, tarnished by the original sin of all organizations that Coase identified, but how they operate, especially under pressure and conditions that seem tailor-made to generate errors, can be improved. Doing so, however, requires active experimentation, assessment and learning. My thesis makes a contribution to this process.

Notes for the reader: This dissertation comprises an introduction, three stand-alone papers, and a conclusion, followed by references for the introduction and conclusion. The total word count is approximately 33,500 words.

2

The Importance of Being Earnest: Incentives and Hierarchy in Public Sector Decision-Making

The Importance of Being Earnest: Incentives and Hierarchy in Public Sector Decision-Making

Ranil Dissanayake *

May 2023

Abstract

Policymakers struggle to interpret evidence correctly and navigate complex and sometimes conflicting incentives. In this survey experiment in a public bureaucracy in the UK, policymakers choose between two investments on the basis of statistical information on their prospects; some receive information that their Minister has previously expressed support for the less effective option (treatment 1) or the more effective option (treatment 2), though the bureaucrat's task is always to select the more effective choice. This constitutes a 'bad' incentive if it is costly to disagree with those in power. Bad incentives affect junior staff, inducing a 16 percentage point increase in selection of the less effective option. Senior staff are unaffected. This has implications for how decision-making might best be organized in public bureaucracies.

Keywords: Bureaucracy, Public Organization, Organizational Behaviour

1 Introduction

The decision-making and advisory capacity of civil servants carry enormous welfare implications. The middle and senior ranks of the civil service are consulted on virtually every matter of national importance in the UK, from foreign policy to taxation structure. In some departments civil servants may take discretionary spending decisions of up to GBP5 million (DFID 2016, 2020). However, even in the best-functioning bureaucracies, decision-makers make mistakes and can be subject to incentives that run counter to public welfare. This article investigates decision making under ideal circumstances and when the incentives facing public servants are complicated by political intervention, and examines how one organisational feature – hierarchy – affects decision-making quality.

There is a long history of empirical and theoretical research into bureaucratic performance, dating at least to the 1950s in the work of Herbert Simon (Simon 1997; Simon et al. 1991). Simon took a largely benign view of public bureaucracies, arguing that their function was limited by the cognitive capacity of their constituent members, and extended (or indeed sometimes further limited) by the organisational choices made in structuring these bureaucracies. Other scholars were more jaundiced. Much public choice theory is concerned with understanding the competing

*Center for Global Development and Blavatnik School of Government, University of Oxford

private and public incentives of bureaucrats, investigating how private returns, ranging from salary and promotion opportunities to opportunities for predatory rent-seeking, influence the decisions made by public agents, both political and bureaucratic (Besley 2007; Buchanan and Tullock 1962).

Empirical work has confirmed aspects of both of these views. Though little research yet exists in a public bureaucratic setting, it is becoming apparent that in all organisations and most settings decision-makers make mistakes, and there is high variance in how different decision-makers within the same organisation make choices over the same set of options (Kahneman et al. 2016). This is very much in line with the spirit of Simon's work on bounded rationality. It is also clear that performance of public servants (assessed in terms of organisational objectives) is also affected by their personal incentives to optimise their career prospects (Iyer and Mani 2012; Leaver 2009; Rogger and Somani 2018), though with considerable heterogeneity across 'types' of agents. Further research has demonstrated an additional cognitive limitation, one not predicted by the early work in organisational behaviour: like most other populations, civil servants fall prey to sunk cost bias, confirmation bias and other decision-making biases common in the behavioural economics literature (Banuri et al. 2019).

Related research has considered how best to motivate better bureaucratic and organisational performance. Theoretical foundations for optimally selecting and incentivising bureaucrats and public-spirited agents have been established (Acemoglu 2010; Besley and Ghatak 2005; Dewatripont et al. 1999). Empirical testing has yielded mixed results, but the general point that there are returns to better selection and incentivisation of bureaucrats is clear (Ashraf et al. 2014; Bandiera and Lee 2015; Banuri and Keefer 2015). A related seam of empirical work has demonstrated the importance of management structures on performance (Rasul et al. 2017; Rasul and Rogger 2018).

However, virtually all of this work considers some form of project completion as the measure of bureaucratic effectiveness, rather than the quality of decisions they make. This means there is still relatively little quantitative research into how the cognitive performance of bureaucrats affects the quality of policy making as opposed to delivery. It also neglects the role of commonplace structural choices in bureaucratic organisation of the type Simon focused on in amplifying or mitigating the performance (or lack thereof) of bureaucrats. Research has looked at the role of selecting or incentivising all bureaucrats; an alternative is to consider leveraging variation within the organisation to improve decision-making. In some sense hierarchical systems are an attempt do just this – giving different roles to people with a different vector of qualities. However, the ways in which cognition and decision-making quality vary with hierarchy has not been the subject of much empirical attention, though work has considered behavioural biases among CEOs or their contribution to firm value – presumably in part driven by their decision making quality (Lieberson and O'Connor 1972; Malmendier and Tate 2015), and an extensive literature considers effect of steepness of hierarchy on attitudinal outcomes and final performance in a range of (mainly private sector) organisations (see Anderson and Brown 2010 for a good survey).

This article begins to fill these gaps. I implement a survey experiment with civil servants in the UK's international development department, in which they are required to interpret some simple quantitative evidence and make a choice between two investments. For a treatment sample, this choice is complicated by providing them with additional information expressing a preference stated by the most senior political figure associated with the Department studied,

though in the context of the experiment this preference should not affect their interpretation of the evidence provided. I find junior officials are much less likely to recommend the investment backed by the evidence provided if the Minister has expressed a preference for the less effective investment, but this effect disappears among more senior officials. These results are robust to alternative functional forms and definitions of seniority.

This study contributes to this literature in three ways. Firstly, it is one of very few studies that experimentally examines decision-making in a well-functioning, highly selective and well-motivated public sector bureaucracy, or among middle and senior managers. Secondly, it investigates the effect on decision-making of incentives that are misaligned with the organisational objective. And thirdly, it demonstrates how these effects vary with seniority in the organisation, with implications for decision-making in hierarchical organizations.

The article proceeds as follows. The next section describes the setting. Section 3 describes the survey experiment. Section 4 discusses sampling and attrition. Section 5 provides the empirical strategy. Section 6 presents the results of the experiment. Section 7 provides some analysis of the mechanisms underlying the main result. Section 8 concludes with discussion considering the implications of these findings for how decision-making structures should be set in bureaucracies, and the limits to existing approaches for error minimisation.

2 Setting

The study setting was a UK Government bureaucracy with a strong international reputation. Quantitative empirical research into decision-making in such settings is still rare. The Department for International Development was the UK Government Department charged with managing the UK's GBP14 billion aid budget.[1] It employed 2,754 staff (2,678.4 full time equivalents).[2] Though it was merged with the UK's Foreign and Commonwealth Office into the Foreign, Commonwealth and Development Office in 2020, the Department had a strong reputation and played an important role in international development, both directly through its bilateral footprint and its substantial contributions to multilateral bodies (Gavas and Calleja 2020).

In the years running up to the experiment DFID rapidly expanded its staffing. Between 2010 and 2017, full-time equivalent staff increased by 53%. Some of this uplift was recruited through the highly selective civil service fast stream, which involves two online questionnaires, an e-tray exercise testing decision-making skills, a video interview before additional tests, sifting and finally a half-day assessment centre consisting of a leadership exercise, a group exercise and a written-analysis exercise before a possible final assessment.[3] However, the majority were recruited from across the civil service or externally. DFID had an unusually high proportion of staff operating in middle and senior-management levels compared to other civil service departments—at the time of the experiment some 59% of DFID staff were senior civil servants or in the top advisory grades, compared to a civil service median of 25%, and just 37%

[1] At the time of this research DFID was a standalone Government Department, with Cabinet-level Ministerial representation. It has since been merged with the Foreign and Commonwealth Office to become the Foreign, Commonwealth and Development Office.

[2] DFID published comprehensive workforce data on a monthly basis at https://www.gov.uk/government/collections/dfid-workforce-management-information-public-body. These figures come from the February 2019 update, the latest available at the time of writing.

[3] https://www.faststream.gov.uk/faqs/index.html

and 46% in comparable Whitehall departments, the Foreign and Commonwealth Office and Treasury, respectively.[4] Those joining the Department at these more senior levels were usually expected to have a number of years' experience, a Master's degree in a subject relevant to the role, substantial technical expertise in the area relating to the role and a strong history of work experience. Staff recruited into one of DFID's many professional cadres (including economists, statisticians, governance advisers, health advisers and education advisers) were required to have a Master's degree in a related subject.[5]

Once hired, all staff were bound and protected by the Civil Service Code, which provides both formal protections and aspirational statements of conduct to civil servants. The code is protected by the Civil Service Commission.[6] It aims to protect their ability to provide impartial advice to Ministers and make decisions in the public interest. The civil service code states:

> As a civil servant, you are appointed on merit on the basis of fair and open competition and are expected to carry out your role with dedication and a commitment to the Civil Service and its core values: integrity, honesty, objectivity and impartiality. In this code:
>
> - 'integrity' is putting the obligations of public service above your own personal interests
> - 'honesty' is being truthful and open
> - 'objectivity' is basing your advice and decisions on rigorous analysis of the evidence
> - 'impartiality' is acting solely according to the merits of the case and serving equally well governments of different political persuasions
>
> These core values support good government and ensure the achievement of the highest possible standards in all that the Civil Service does. This in turn helps the Civil Service to gain and retain the respect of ministers, Parliament, the public and its customers.[7]

The civil service code may be seen as part of the 'explicit' public service bargain between civil servants and their principals (Ministers, the public, or their organization). How it affects civil servant behaviour in practice will also depend on the 'tacit' aspects of the public service bargain struck between civil servants and principals, which may also change over time (Elston 2016).

The Department had a hierarchical structure, with work produced in teams typically made up of staff of the same grade or closely bunched grades and submitted and 'cleared' up the hierarchical chain until a final decision was taken, though the level of seniority at which a policy or spending decision could be approved varied by importance of the decision. Throughout this paper, hierarchy is defined as grade seniority in the civil service structure. While some

[4] Data are taken from DFID's Workforce Management Information (https://www.gov.uk/government/collections/dfid-workforce-management-information-public-body) and the Civil Service statistics (https://www.ons.gov.uk/employmentandlabourmarket/peopleinwork/publicsectorpersonnel/datasets/civilservicestatistics)

[5] For an example job advert, see: https://www.greatugandajobs.com/jobs/job-detail/job-a2-social-development-advisor-job-at-department nav-15

[6] "If a civil servant is asked to do something which conflicts with the values in the Code... Their department should investigate their concern. If the Civil Servant remains dissatisfied following the outcome of the investigation, they may bring a complaint to the Civil Service Commission. In some cases, the Commission may also hear a complaint direct." Quotes from https://civilservicecommission.independent.gov.uk/code/

[7] https://www.gov.uk/government/publications/civil-service-code/the-civil-service-code

people do hold high grades without line management or spending responsibility, there is a clear correlation between decision-making responsibility and grade.

Grade seniority in DFID was not an automatic consequence of tenure: staff did not receive automatic promotion with time served in the organization. Instead, there were three possible paths into senior or middle management positions. The first was internal promotion within DFID. The second was transfer either at level or on promotion from another civil service department. The third was direct recruitment from outside the civil service. At the time the survey was undertaken, all civil service recruitment and promotion was done on an open competition basis by default [8]. In each case, promotion or appointment to a senior role followed an active application process—that is staff had to self-select into the pool of applicants for the role. Applicants would be shortlisted by a selection panel, and then interviewed and assessed on the basis of a structured scoring criteria based on a set of 'civil service competencies' which were explicitly set out, with the requirements to be satisfied at various levels of seniority detailed in civil service guidance. Which competencies were assessed depended on the specific characteristics of the role.[9] Often, successful promotion required demonstrating by example capabilities typical of the grade to which applicants aspired. That is, for an applicant to a Director-level job, even if they are currently in a Deputy Director role, they would need to demonstrate that they have shown the capabilities expected of a Director in their role as Deputy Director.

Staff reported a high degree of intrinsic motivation, but dissatisfaction with pay.[10] After many years of below-inflation pay rises, at the time the experiment was conducted (2017), achieving a real-terms pay rise was typically only possible through promotion or a posting overseas.

3 Method

The study was conducted in December 2017. I took advantage of the regular but infrequent 'DFID Evidence Survey', a representative sample of DFID civil servants with at least some decision-making responsibility. The survey is undertaken every two years and took around thirty minutes to complete. It collected basic information about roles, seniority, professional and technical qualifications and self-assessments about their level of ability in interpreting evidence. It then moves on to a series of open questions about the use of evidence in DFID, including about the quality of evidence used in decision-making and whether evidence-based decision-making is valued in the Department. The surveys were sent out using 'Survey Monkey' with a message requesting that staff fill out the survey, and a link to a video of the 'survey champion', a DFID senior manager, talking about the survey. A full list of staff was obtained from Human Resources, and administrative grades (those below 'B2') and political appointees were excluded. The remaining list was stratified by civil service grade and then by professional specialism. We took a random sample

[8] A very small number of roles were filled by direct appointment. "Direct appointments (including promotions) are now only allowed in a few exceptional cases" (Sasse and Norris 2019)

[9] The general civil service competency framework is set out in detail here: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/436073/cscf_fulla4potrait_2013-2017_v2d.pdf. Additional competency guidance, specific to DFID in this period is set out here: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/436073/cscf_fulla4potrait_2013-2017_v2d.pdf. Since this time, the civil service competency framework has been replaced by a similar but distinct approach, using so-called 'Success Profiles'

[10] Over the last ten years, the annual Civil Service People Survey's show that satisfaction and identification with the Department's mission and objectives was usually between 86-92% while satisfaction with pay and conditions hovered around 36-42%. See: https://www.gov.uk/government/publications/department-for-international-development-civil-service-people-survey-2018

from each stratum, with the number sampled calculated based on their proportion in the overall population. Further details of sampling, attrition and data are presented in the next section.

Within this survey, I implemented a survey experiment setting a moderately challenging cognitive task for respondents and tested how well their decisions held up against compromised incentives. The aim was to understand how well civil servants process evidence; if signals from powerful political leaders disrupt this; and how the hierarchical decision-making structure of the civil service amplifies or mitigates these effects. Even though the setting features highly selective recruitment, we would still expect error rates to be non-negligible as this has been observed in other highly motivated settings, with both systematic (Thaler 1987) and non-systematic errors observed (Kahneman et al. 2016). Errors of any kind matter in civil service settings, since the advice of civil servants is typically used as a benchmark against which to hold elected Ministers to account; and Ministers correspondingly place great weight on the advice provided by their civil servants. The hierarchical structure of decision-making ensures that decisions are usually 'cleared' by more senior staff after the initial analysis is undertaken by more junior civil servants. Understanding if this process is likely to improve or hamper decision-making is important.

The experiment took the form of a vignette. Vignettes are an increasingly popular method for learning about organisational behaviour, often looking at questions of discrimination (Neumark 2018). In this vignette, the respondent is asked to assess options for a new education programme. They are explicitly informed that the department's objective is to improve learning outcomes. They are told that they have commissioned research in the place in which the programme will be run, performing an A/B comparison of two kinds of intervention by means of a randomised control trial (RCT), and given the results of this RCT. They are then asked to choose which of the two interventions they would select, noting that they will personally sign the submission of the business case to the Secretary of State. In the control group, this is the extent of the problem. The numerical skills required to correctly select between the two options are well below the threshold of a generalist civil servant. The full control group vignette is reproduced below.

> *"You are responsible for a business case for a new GBP35 million education programme with the objective of improving student learning outcomes and will sign the covering submission requesting approval by the Secretary of State. Learning outcomes are a ministerial priority and manifesto commitment.*
>
> *In preparation for your new business case, you commissioned a study that provides strong, experimental evidence in a number of regions of the country you work in on the long-term effects of both curriculum development and teacher performance incentives. The results of this study are summarised below. (For logistical reasons, the total number of students in each of the two groups is not exactly the same, but this does not prevent the assessment of the results.)*

	Students with improved scores	*Students whose scores show no change*
Curriculum Development	*2230*	*750*
Teacher incentives	*1070*	*210*

> *In light of the information provided, which of the following recommendations would you make in your*

business case and covering submission:

1. *To use the programme to improve curriculum development across the country.*
2. *To use the programme to roll out a system of teacher performance incentives across the country."*

The treatments were identical except for the addition of a new paragraph, providing information on Ministerial preferences while reiterating that improved learning outcomes are the objective:

> *The Secretary of State has been giving a series of interviews in which she has stressed the importance of curriculum development [OR teacher incentives] for improving learning outcomes, which she has said is one of her key objectives.*

This paragraph was placed directly after the first paragraph of the control treatment text. The paragraph provides information on what the most senior Minister in the department has publicly announced as her preferred method of achieving learning improvements. However, it also reiterates that achieving success in this area is one of her key objectives.

This additional text created two treatment groups. In treatment group 1 the Minister been supporting the less effective intervention in interviews. In treatment group 2 the Minister backed the more effective intervention (which intervention was more effective was also randomised). Including a treatment for the Minister backing the more effective option aids interpretation of the results: it allows us to disambiguate the incentive effect on decision-making of knowing that selecting the better option means disagreeing with Minister from the information effect of simply knowing that another person has backed one or the other option. In the former case we would expect treatment 1 to have a larger effect than treatment 2; in the latter case we would expect the treatments to have the same effect. We return to this distinction below.

The vignette reflects the kind of choice often made in 'spending' departments in the UK, where decisions have to be made to between alternative programmes or investments, or to select the most cost-effective way of achieving some specified output. In DFID, almost all spending decisions were made on the basis of some form of cost benefit analysis comparing different approaches to achieve pre-specified objectives (DFID 2011). As such, the vignette captures a salient form of decision-making, and learning more about how mistakes can be made in this kind of decision-making is valuable for improving Government spending decisions.

It is worth making explicit a few points about the construction of this vignette. Firstly, selecting between the two options is simply a matter of comparing the ratio of successful outcomes (students with improved scores) to unsuccessful outcomes (students whose scores show no change). In the example above, the ratio for curriculum development is about 3:1 and that for teacher incentives is about 5:1. We randomised which of the two interventions was better, so some respondents would see teacher incentives with the better ratio, and others would see curriculum development with the better ratio. Anyone who had successfully completed the recruitment process outlined above should be able to calculate and compare these ratios. The table presented is essentially exactly the same as that used by Banuri, Dercon and Gauri (2019) to investigate confirmation bias, itself an extension of that used in the original Kahan paper on confirmation bias (Kahan et al. 2017). The only difference is that the numbers in these two papers are all multiplied

by 10 to give an appropriate sense of scale for an international development RCT. In each of these previous uses of the 2x2 table, the correct option was the one with the higher ratio of good outcome to bad outcome, which had a smaller sample size and thus lower numbers of individuals treated. The language I use on sample size ("the total number of [students] in each of the two groups is not exactly the same, but this does not prevent the assessment of the results") is identical to that used by Kahan et. al. and Banuri et. al.

Secondly, the sentence "learning outcomes are a ministerial priority and manifesto commitment" is of particular significance to civil servants. Manifesto commitments are promises made by the Government in power while campaigning for election, and it is considered an extremely serious matter to break them. They are the collective responsibilities of the Government. They supersede the preferences of any individual Minister, since all Ministers campaign on the basis of the Manifesto, and if elected are expected to adhere to these promises.

Thirdly, the size of the business case, GBP35 million was carefully chosen. It is a large business case and as such respondents should infer that the (hypothetical) stakes are high. However, it falls just below the threshold for independent scrutiny by the Department's Quality Assurance Unit, which is set at GBP40 million (DFID 2016). This means that while a decision must be taken carefully (which is of course true no matter what the size of the business case), there is a relatively low chance of external scrutiny on the decision taken by the respondent.

Finally, the fact that the respondent is told that they will sign the covering submission that presents the final business case to the Minister indicates that they will be personally identifiable as the author of the decision taken.

The survey was piloted with an out-of-sample group of staff. They answered all the questions, timed how long it took them to complete the survey, and provided feedback on the construction of all the questions. In their feedback on the vignette, they all reported a clear understanding of the objective facing the civil servant: to find the most effective means of promoting learning outcomes. To this end, all reported an attempt to infer which intervention was more effective. All but one of the respondents reported some attempt to calculate the ratios of the two interventions, though not all did so correctly, and that they recognised the element of personal responsibility invoked in the vignette. The survey was taken at work, during working hours; such surveys are not uncommon and approximate the real working conditions of staff, including the time pressures they experience.[11]

The vignette is constructed to provide an insight into how policymakers process evidence under a somewhat stylized version of their usual working environment. The sequence of events in the control group is as follows:

1. Civil servants receive a signal about their objective (in this case, the manifesto commitment and ministerial priority, learning outcomes).

2. The civil servant receives information about the best way of achieving this outcome from the options available in a format that requires some additional processing.

3. The civil servant interprets this information.

4. The civil servant makes a recommendation to the final decision-maker, in this case the Secretary of State.

[11] We asked how they went about comparing the two interventions, and what, if any, importance they attached to the fact that they would be author of the covering submission for the business case.

For the treatment groups, the sequence is identical, except they receive an additional signal after step one, which complicates their calculation:

1. Civil servants receive a signal about their objective (in this case, the manifesto commitment and ministerial priority, learning outcomes).
2. The civil servant receives a signal about the Minister's public announcements about how to achieve learning outcomes.
3. The civil servant receives information about the best way of achieving this outcome from the options available in a format that requires some additional processing.
4. The civil servant interprets this information.
5. The civil servant makes a recommendation to the final decision-maker, in this case the Secretary of State.

For the control group, the set-up and interpretation is simple. Given an objective, the civil servant must interpret information and select the best way of achieving this objective. Since we randomize which of the two possible interventions is more effective, we are effectively observing the rate at which civil servants interpret the information provided correctly. In the treatment groups, the new stage two complicates matters. A civil servant may interpret this as signalling information the Minister holds that is private, or different to the information they receive in step 3. Another possibility is that they interpret this as a signal of Ministerial preferences, which tells them something about how different options will be received by the Minister. Such a preference could reasonably be interpreted as strongly-held, if the Minister has made public pronouncements in favour of one option. Agreeing with the Minister's preferences may please them, while disagreeing with the Minister's preference may displease them. Pleasing or displeasing Ministers may have perceived career consequences (though in practice Ministers in the UK have almost no say over civil service promotions or career paths), or may simply carry personal costs, if it is difficult to displease those in power.

In the context of this experimental vignette, the informational signal of the Ministers preference should have relatively little weight relative to the knowledge that learning outcomes are a Government manifesto commitment and the evidence commissioned, which has a much more direct bearing on the optimal choice for achieving the stated Government and Ministerial objectives. However, two points should be noted. First, even within the experimental set-up, a distinction should be drawn between the informational signal the Minister's preferences carry and the effect it can be expected to have in the context of a power-based relationship in which the Minister has much more power than the typical civil servant. In such a situation, we may expect civil servants to pay attention to Ministerial utterances and, potentially, to act upon them even if contra-indicated by the evidence. Such a situation is clearly anticipated in civil service recruitment and promotion, given that the competency framework for senior civil servants explicitly notes that to be effective they should not simply tell Ministers 'what they want to hear'. This suggests one way in which senior and junior staff may respond differently to the experimental vignette. Secondly, in the real world, it would not normally be so straightforward to ignore the informational signal provided by Ministerial utterances: they may carry new or additional information about Government or Ministerial priorities, and given that the Minister learns about

their brief (and brings experience to it), may reflect different or superior knowledge of a specific situation.

While hierarchy was not explicitly invoked in the vignette, the survey collected information on the seniority of respondents to allow investigation of how it affects outcomes. One possibility is that those at more senior levels are more likely to be influenced by a desire to please Ministers because they have more regular interaction with them, or are more likely to be in the running for the few highly senior jobs that involve some Ministerial say in appointments. Another possibility is that senior staff are less likely to act to please Ministers because they more closely adhere to the civil service code (which becomes more salient as seniority increasingly exposes staff to political interactions) or because they are more secure in their jobs, having already secured influence, pay and status. Interacting treatment with seniority will provide insight into this.

4 Sampling, Attrition and Data

We sampled 989 of DFID's staff to survey. The sampling strategy followed a stratified randomisation design, to match the previous Evidence Survey in 2013. There was no possibility of changing this sampling strategy, as consistency and comparability was a key requirement for the team implementing the survey.

The population of the survey was from all DFID staff of B2 grades and above (i.e. those likely to make some programming or policy decisions and use evidence in their work), across DFID's two headquarters (one on Whitehall in London and the other in Abercrombie House in Scotland), and its Country Offices. A complete list of this population was stratified by grade (B2 – B1 – B1D – A2L – A2 – A1 – SCS) and further sorted by professional specialism. The number in each stratum was calculated based on the proportion of the relevant cadre and grade in the overall population. Some non-advisory roles and political positions were filtered out of the survey. The survey sample was randomly selected from each stratum, and then assigned randomly to treatments 1, 2 or control groups (without any stratification) and sent the survey link to complete within a 6-week deadline.

We sent two reminders and used a publicity campaign using internal communications and media to boost response rate.

Despite these efforts, our response rate was only 61% (604 individuals) - nevertheless, high for typical civil service surveys. At this stage attritors will not have seen any part of the survey and will be unaware of the existence of the experimental vignette, let alone to which treatment group they are assigned: we are simply measuring whether or not they ever even opened the survey instrument (attrition once the survey was opened is treated separately, below). Nevertheless it is possible that certain kinds of civil servant were systematically less likely to respond to the survey, which would compromise external validity – the extent to which the results presented here can be generalised to the entire population of civil servants in DFID.

To see if this was the case, I ran a regression to see if any respondent characteristics predicted their likelihood of opening the survey at all. Using the linear probability model on a binary 'opened survey/did not open survey' dependent variable I find that few respondent characteristics are predictive of whether or not the survey is opened. Table 1

summarises. (The data set from which sampling was undertaken was more detailed than that collected by the survey – as a result, we have data on tenure in DFID which is precise to the year for the analysis of response rates, but not for the survey experiment).

Table 1: Survey participation regressed against respondent characteristics

	Dependent variable: Opened survey (0/1)		
	Coefficient	Robust SE	p-value
Tenure	-.004*	.002	0.06
Grade	-.021	.013	0.11
Role			
Senior Management	.289**	.116	0.01
Policy Professionals	-.069	.111	0.54
Programme Managers	-.076	.106	0.47
Econ, Evaluation and Statistics	.040	.113	0.72
Human Development	-.064	.122	0.60
Governance and Conflict	-.063	.123	0.61
Climate, Humanitarian and Livelihoods	.0463	.126	0.71
Private Sector and Infrastructure	-.030	.128	0.82
Directorate			
Country Programmes	.076	.101	0.45
Economic Development	.020	.104	0.85
Middle East and North Africa	.099	.108	0.36
Policy and Global Issues	.122	.094	0.20
Top Management Group	.048	.176	0.79
HQ Dummy	.050	.048	0.29
Constant	.654***	.087	0.00

Note: *** p<0.01, ** p<0.05, * p<0.1.

No respondent characteristics predict attrition except tenure and whether or not they were senior manager. For tenure the effect is small – equivalent to a 0.4% reduction in likelihood of opening the survey per year of service. With a mean tenure of around 7 years in the sample, this suggests a slight underrepresentation of more experienced staff in the results, but the size of this effect means it is of little practical significance. Senior managers were also slightly more likely to respond to the survey, which does suggest that they are over-represented in the sample compared to junior staff, but seniority is controlled for in the main specifications, which should minimise the extent to which this compromises generalisation to the whole population.

Among survey participants, a small number opened the survey but did not complete it. Of the 604 who opened the survey 536 completed it (89%). Of those who did not complete the survey, none progressed as far as observing the experimental vignette.[12] A regression of attrition on treatment arm yields no significant relationships, indicating that attrition is not predicted by treatment arm.[13] Similarly, chi-square tests confirm no differential attrition by treatment arm either at either response or completion level (i.e. there is neither differential attrition at the stage of opening the

[12]Since certain pages required mandatory information to be completed before the respondent could move on to the next page, it's possible to assess whether or not it was possible for the respondent to have observed the vignette by looking at how much data they did enter. Almost all failed to enter any data at all, and those that started the survey all exited it before reaching the main body of the survey, let alone the vignette.

[13]This is true whether a logistic or LPM model is used to estimate the regression. See Appenix A

survey at all nor among those who completed the survey and the experimental vignette). The 989 sampled and those responding were divided into the control and treatment groups as in Table 2:

Table 2: Control and treatment sample sizes

	N (assignment)	N (respond)	N (completed)
Control	330	199	173
Treatment 1: Minister backs the less effective option	327	203	188
Treatment 2: Minister backs the more effective option	332	202	175

Calculating statistical power *a priori* was guesswork to some extent, given that there was no publicly available data from which to judge the baseline ability to interpret evidence, effect size for treatment or standard deviation in interpreting evidence for the population. However, assuming around 75% of the population select the correct option in the control group, the study was powered to detect relatively small effect sizes – of 5 percentage points - 80% of the time if the standard deviation was as high as 17.

In addition to the vignette through which the survey experiment was implemented, the Evidence Survey collected a range of fully anonymised information useful as control variables and to explore causal pathways through additional analysis. The data collected included:

- Respondent grade (i.e. the level of seniority achieved in the Department)
- Respondent cadre (technical specialisation, if any)
- Respondent tenure (in DFID, a categorical variable, with three choices: less than five years, five to ten years and more than ten years)
- Gender
- Their self-assessed ability to interpret quantitative evidence
- Whether they were recruited locally (if international staff) or in the UK[14]
- Their self-assessed level of comfort in challenging fellow civil servants
- Their self-assessed level of comfort in challenging Ministers

Though much of this information is valuable as control variables and for teasing out possible causal pathways, most of the data has the drawback of being categorical rather than continuous. This was unavoidable. Previous iterations of the Evidence Survey used such variables and the implementing team needed to ensure comparability. Choices have had to be taken in coding this categorical and ordinal data, and in particular, two should be highlighted.

Firstly, I code civil service grade into a simple dummy variable, taking the value 1 if the respondent is a middle or senior manager and 0 if they are at any grade below this. One alternative to doing so would be to either treat grade as a continuous variable, coded from 0 to 6. This has the major drawback that there isn't a clear linear relationship in

[14]The recruitment process for locally recruited staff is less intensive than for UK-recruited staff.

responsibility with grade. It is not clear that grade 2 is only half as powerful as grade 4; while a respondent at grade level 6 on this scale (the 'Senior Civil Service') may have substantially more power than one at grade 5. Another alternative would be to split the grades into three strata, separating junior staff, middle managers and senior managers. Reassuringly, using any of these strategies preserves the main results of the paper, suggesting that data coding choices and functional form assumptions are not driving them. Appendix A presents two tables of all six specifications used in the main analysis with each alternative definition of seniority in turn, which demonstrate the the results are robust to using either of these definitions of seniority.

Secondly, self-assessed ability to interpret quantitative evidence was also collected as a five-point likert scale, with 1 being no ability at all and 5 being very high ability. Again, there is no clear interpretation of this five point scale – is 2 twice as effective as 1, for example? What's more, the full scale used for self-assessed ability is not particularly informative about differences in actual ability, at least when judged against the cognitive task in this experiment. In the control group, there is a clear difference in the rate at which the correct option is chosen between those self-assessing as levels 4 and 5 in and the rest, but no real difference between 4 and 5. As a result, I have chosen to code the information in this variable as a dummy for high ability, given the value 1 for those self-assessing their interpretive skills at 4 or 5 and 0 for the rest.

The experimental data collected was as follows:

- Dummy variables signifying assignation to treatment 1 (Minister back the less effective option) and treatment 2 (Minister backs the more effective option), with the reference category those assigned to the control group
- A 0/1 dummy variable for whether or not the respondent selects the more effective option in their business case.

The next section sets out the empirical strategy used to analyse the results.

5 Empirical Strategy

Analysis of the results uses the Linear Probability Model (LPM). Its major advantage is in the straightforward interpretation of the coefficients of interest, which can simply be read as percentage point changes in the likelihood of the event of interest. In the case of this experiment, coefficients on variables of interest represent the increase (or decrease) in the likelihood of selecting the more effective option from a one-unit increase in the variable in question. The basic models estimated are as follows:

$$Y_i = \beta_0 + \beta_1 Treatment1 + \beta_2 Treatment2 + \epsilon \tag{1}$$

$$Y_i = \beta_0 + \beta_1 Treatment1 + \beta_2 Treatment2 + \boldsymbol{\beta}'\mathbf{X}' + \epsilon \tag{2}$$

Y_i is the variable of interest, the probability of selecting the more effective option. β_0 is an intercept term, interpreted as the baseline probability of selecting the correct option (i.e. the probability holding all control variables at 0). β_1

and β_2 are the effects of the two experimental treatments (the Minister supporting the more or less effective option, respectively. $\boldsymbol{\beta}'\mathbf{X}'$ is a matrix of control variables used to improve accuracy of the model. ϵ is the error term.
To investigate how these effects vary with position in the civil service structure, a third specification is estimated investigating the interaction between treatment and position in the departmental hierarchy.

$$Y_i = \beta_0 + \beta_1 Treatment1 + \beta_2 Treatment2 + \beta_3 Treatment1 * Seniority + \beta_4 Treatment2 * Seniority + \boldsymbol{\beta}'\mathbf{X}' + \epsilon \quad (3)$$

This is the preferred specification. Empirically, 'seniority' is defined as a simple 0/1 dummy variable capturing whether a staff member is part of the middle and senior management grades of the civil service or not. In this specification, the X variables will also include seniority. The interaction terms will give the impact of the two treatment arms among senior as opposed to junior staff. This gives the effect of treatment at different levels of seniority.
What seniority carries in the presence of controls is worth discussing explicitly. Controls include tenure in the organization and self-assessed ability in handling quantitative evidence (and in one additional specification included to investigate mechanisms, self-reported willingness to challenge Ministers). As the section on promotion and hiring explains, there was no automatic process of career progression in DFID. Controlling for tenure means that the seniority variable excludes the effect of experience and habituation (it is possible that experience imparts knowledge of how best to handle difficult interactions in the organization regardless of grade). Controlling for ability with quantitative evidence excludes any additional capacity senior staff have to interpret the numbers presented in the experimental vignette compared to junior staff (either selective, if selection processes prize technical ability, or causal, if seniority brings with it more opportunities for learning and development). This does mean that one way in which seniority might be expected to be associated with better outcomes is excluded from what seniority captures (we examine this by interacting ability with treatment in another additional specification included to investigate mechanisms). This means that in our primary specifications, the seniority variable captures a subset of what might make senior officials different to junior ones. That includes what personal factors lead them to seek out more senior positions in the first instance (ambition, for example, or self-confidence; or a desire to be involved at the highest level of decision-making); other characteristics that are selected for in the process of promotion or recruitment (which might include probity, or independence of thought, or indeed willingness to challenge others in the pursuit of better performance, or communication or other non-cognitive skills); and capabilities that are conferred causally by achieving seniority (which might include, for example, an ability to directly challenge other senior figures due to organizational status, or specific types of experience that accrue to only those who reach certain levels of the hierarchy). In other words, we capture a bundle of characteristics that senior people have or acquire, excepting their ability to handle the specific calculation in the experimental vignette (which is of secondary importance to the present research) and such characteristics associated solely with length of tenure in the organization.
Inferences are as follows:

1. β_0 in specification 1 provides the frequency of errors in the cognitive test, with no incentives invoked and without controlling for individual characteristics. This is essentially the mean of the dependent variable in the control

group.

2. β_1 gives the causal impact on decision-making quality of information that the Secretary of State has backed the less effective option. Any effect could reflect a belief that promotion prospects are harmed by contradicting the Secretary of State[15], simple disutility from disagreeing with Ministers or that the Secretary of State holds relevant private information.

3. β_2 gives the causal impact of information that the Secretary of State backs the more effective option, indicating if agreeing with Ministers has a career or other utility benefit, or that the Secretary of State holds relevant private information.

4. β_3 and β_4 give the causal impact of seniority on the two treatment effects above: does the information have a different impact by seniority?

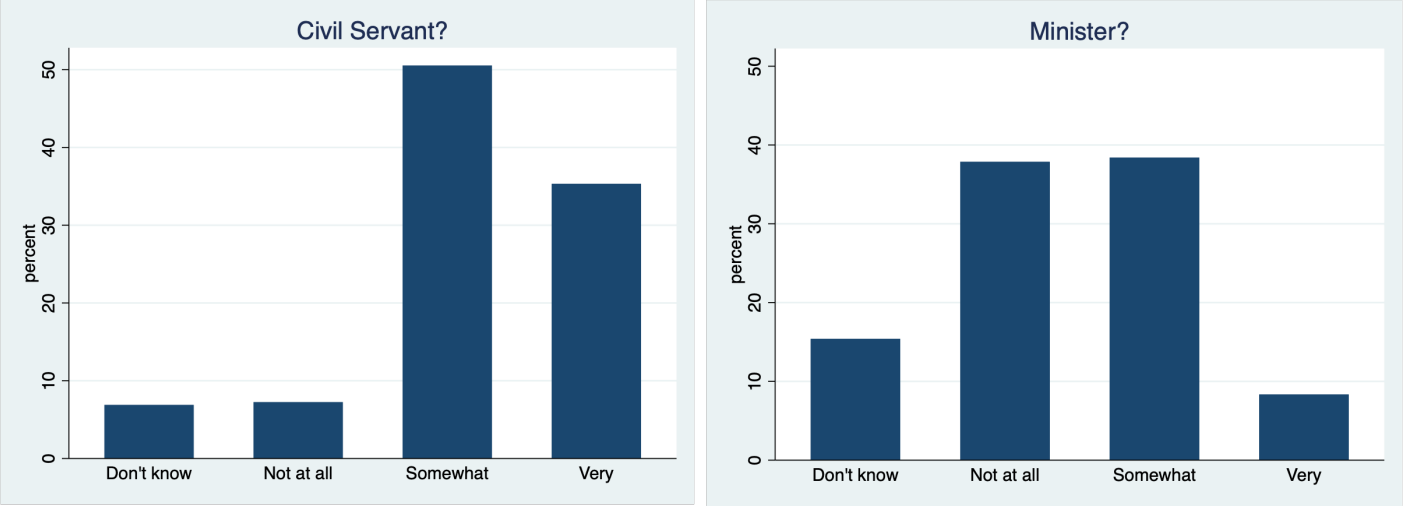
We also run exploratory analysis to investigate the mechanisms through which any observed effects occur. These are specified below.

6 Results

Descriptive data from the survey suggest that the idea that civil servants would rather avoid challenging their Ministers directly is sound. Though the data doesn't directly measure the disutility of challenging Ministers, but we can compare how 'comfortable' respondents report being when challenging Ministers compared to civil servants.

Clearly, there is a substantial difference here: almost all respondents are 'somewhat' or 'very' comfortable challenging

Figure 1: How comfortable are you challenging the use of evidence by a...



civil servants on poor use of evidence, but almost as many report being 'not at all' as 'somewhat' comfortable challenging Ministers. This visual inspection is confirmed by a Kolmogorov-Smirnov test of equality of distributions, which

[15] Note that this need not operate via political interference, since senior civil servants may penalise 'difficult staff' even if Ministers make no such representations to them

strongly rejects the null hypothesis that these two variables follow the same distribution ($p<0.0001$).[16] This is not direct evidence that civil servants will not challenge Ministers, since many may still do what they feel uncomfortable doing, but it does suggest that challenging a Minister carries a personal disutility, while the data are also consistent with perceptions of a career penalty to challenging Ministers. In the UK context, such a penalty is likely to be at best indirect and weak, since Ministers have almost no influence on promotion or career prospects of all but the very most senior civil servants. Any such channel would need to operate either through mistaken beliefs about promotion processes or, a belief that other civil servants (who may well have a say on their colleagues' promotion prospects, by sitting on promotion and job search panels) will penalise colleagues for challenging Ministers too directly. Empirical studies do suggest that career incentives affect performance, though the presence of such a channel in this context is by no means certain (Bandiera and Lee 2015; Iyer and Mani 2012; Leaver 2009).

Table 3 presents the results of the basic LPM regression without controls (specification 1), with additional basic controls (specification 2), with basic controls and interactions with seniority (specification 3), with basic controls, interactions and additionally controlling for self-reported comfort challenging Ministers (specification 4), and with basic controls, interactions between seniority and treatment and interactions between self-assessed ability in quantitative tasks and treatment (specification 5). These last two are exploratory regressions aimed at better understanding the mechanisms through which treatment effects the likelihood of selecting the more effective option and discussed further below.

[16] The two-sample Kolmogorov-Smirnov is a non-parametric test to evaluate the null hypothesis that the two variables are drawn from the same distribution. It returns a test statistic of $D=0.3913$, which strongly rejects the null hypothesis ($p<2.2e-16$)

Table 3: Effect of information on Ministers' preferences on selection of the most effective investment option

DV: Selected the more effective option	Specification (all models are LPM)					
Model	(1)	(2)	(3)	(4)	(5)	(6)
T1: Minister backs less effective option	0.011	0.002	-0.158*	-0.157*	-0.182*	-0.160*
	(0.05)	(0.045)	(0.095)	(0.095)	(0.095)	(0.095)
T2: Minister backs more effective option	0.026	0.007	-0.111	-0.112	-0.148	-0.112
	(0.047)	(0.046)	(0.103)	(0.103)	(0.104)	(0.103)
T1 x Seniority			0.222**	0.224**	0.171	0.217**
			(0.107)	(0.107)	(0.114)	(0.107)
T2 x Seniority			0.161	0.163	0.097	0.16
			(0.115)	(0.115)	(0.121)	(0.115)
Seniority		0.161***	0.031	0.027	0.075	0.031
		(0.056)	(0.086)	(0.086)	(0.089)	(0.086)
Ability		0.06	0.061	0.055	-0.052	0.056
		(0.038)	(0.039)	(0.04)	(0.08)	(0.04)
Comfort challenging Ministers				0.063		0.072
				(0.059)		(0.091)
T1 x Comfort Challenging Ministers						0.038
						(0.14)
T2 x Comfort Challenging Ministers						-0.048
						(0.136)
T1 x Ability					0.139	
					(0.95)	
T2 x Ability					0.189*	
					(0.097)	
Tenure (>1 <5 years)		0.148***	0.154***	0.157***	0.158***	0.160***
		(0.056)	(0.056)	(0.056)	(0.056)	(0.056)
Tenure (>5<10 years)		0.063	0.067	0.072	0.069	0.07
		(0.059)	(0.059)	(0.059)	(0.059)	(0.059)
Tenure (>10 years)		0.068	0.07	0.072	0.064	0.072
		(0.056)	(0.056)	(0.056)	(0.057)	(0.057)
Constant	0.734***	0.482***	0.575***	0.531***	0.599***	0.570***
	(0.033)	(0.078)	(0.091)	(0.101)	(0.091)	(0.092)
Controls	N	Y	Y	Y	Y	Y
Comfort challenging Ministers controlled?	N	N	N	Y	N	Y
R-squared	0.0006	0.1103	0.1193	0.1278	0.1264	0.1213
N	536	536	536	536	536	536

Notes: Robust standard errors in parentheses * Statistically significant at the 10% level ** Statistically significant at the 5% level *** Statistically significant at the 1% level Basic controls: a dummy variable taking the value 1 if the respondent is a middle or senior manager and 0 if not; a dummy variable coding whether or not the staff member was recruited in country as opposed to being put through the full recruitment process in the UK; a dummy variable for whether or not the respondent self-assesses as having high ability in interpreting quantitative evidence (levels 4 or 5 from the 5 point likert scale); a dummy variable for whether teacher incentives was the 'correct' choice; and gender, tenure and professional cadre.

The constant in specification 1 gives the percentage of respondents who selected the more effective option in the control group, when they had no information on Ministerial preferences and no incentives to select anything other than the better option.[17] This is exactly equivalent to the mean of the dependent variable in the control group, and is 73%. In specification 2, I control for seniority in the organisation, as well as a range of other characteristics including self-assessed ability to interpret quantitative evidence, technical specialism of the respondent, recruitment path (home or international) and gender. Recall that seniority, in the presence of controls for quantitative ability and time spent

[17]In each specification, the constant is interpreted as the baseline rate at which surveyed staff selected the more effective of the two learning interventions, with all other variables held at 0. Only the coefficients on the two treatment arms, and interactions with them, can be interpreted as causal since assignation to treatment was random. Other coefficients may have explanatory value, but we cannot be sure that any relationship is causal.

in the organization, carries primarily those other characteristics that differentiate those staff who apply for and are selected to senior roles from those at junior levels, which may include ambition and self-confidence (which may account for application for senior roles), or probity, communication skills and so on. Seniority increases selection of the better option. Self-assessed ability to interpret quantitative evidence is not significantly related to the rate at which the more effective option is selected.

Neither specification 1 or 2 find any effect from either treatment on the rate at which officials select the more effective of the two interventions. Once we introduce heterogeneity by seniority/grade (specification 3), however, large effects are found for treatment 1—when the Minister signals support of the less effective option. In specification 3, the coefficient on treatment 1 alone can be interpreted as the effect of treatment when seniority is 0, that is the effect of Treatment 1 on junior staff. This coefficient is significant (at the 10% level) and negative, meaning junior staff are less likely to submit the option supported by the evidence provided to the Minister when the Minister has previously signalled support of the less effective option. The coefficient on seniority alone loses significance, but the coefficient on seniority interacted with Treatment 1 is both positive and significant. This suggests that the impact of seniority on selection of the more effective option is insignificant in the control group, but that seniority has a large positive and significant effect on the rate at which the more effective option is presented to the Minister in treatment 1. The coefficients on Treatment 2—where the Minister signals support of the more effective option—alone and on its interaction with seniority are not significant. All of the results presented in Table 3 are preserved when the Linear Probability Model (reported here) is replaced by a logistic regression (reported in Appendix B), both in terms of statistical significance and relative magnitude (confirmed by computing the average marginal effect of each reported variable and comparing them with the coefficients obtained from the LPM model). The LPM results are preferred for ease of interpretation of the coefficients. Appendix A demonstrates robustness to alternative definitions of seniority.

In other words, while (junior) civil servants are less likely to select the correct option when the Minister has supported the incorrect one in the media, there is no corresponding boost to the rate at which the correct option is chosen when the Minister has supported the correct one. Indeed, though the effect of treatment 2 is not significant, the point estimate goes in the same direction as the effect of treatment 1, and a Wald test fails to reject the null hypothesis that the coefficients on treatment 1 and treatment 2 are equal (p=0.63). That is, even when the Minister supports the correct option, the point estimate suggests that junior civil servants are less likely to communicate a correct reading of the evidence to the Minister in their advice.

This asymmetry in treatment effect requires some discussion. Though the effect is not statistically significant, it is striking that even when the Minister backs the more effective option, junior civil servants appear less likely to select it in their advice. One possible explanation is that invoking Ministerial involvement makes the cognitive task harder, irrespective of what the Minister actually contributes. Such an effect would plausibly be stronger among junior staff, given that they are less likely to have regular contact with Ministers and are thus less used to engaging with them (or because in a power-based relationship, the difference in power is much sharper between junior staff and Ministers than between senior staff and Ministers). That treatment 1 has a larger point estimate, and the interaction with seniority has a statistically stronger effect may suggest that on top of this 'scrambling' effect, there is an additional perceived

cost to disagreeing with Ministers that junior staff respond to, but senior staff do not—an effect that is concentrated among those who would otherwise have selected the correct answer. In Treatment 1, some of these junior respondents who have correctly interpreted the evidence in front of them despite the 'scrambling' effect of Ministerial involvement, and would otherwise select the correct answer, switch to backing the less effective option in order to avoid disagreeing with the Minister's statements. In Treatment 2, these respondents already agree with the Minister (since the Minister is backing the more effective option and they have correctly interpreted the evidence), and thus do not switch. Those who would in either case have backed the less effective option do not switch—the 'scrambling' effect or cognitive difficulty of the task dominates.

Thus, it seems that junior officials who would otherwise select the correct option may be changing their selection to avoid disagreeing with a Minister in treatment 1, for reasons we discuss further below.

This effect in treatment 1 is substantial. At junior grades, respondents are 16 percentage points less likely to select the more effective option when the Minister has publicly spoken in support of an alternative approach. This effect is statistically significant at the 10% level (p=0.095) and practically significant – it is the equivalent of introducing additional errors to more than a sixth of decisions.

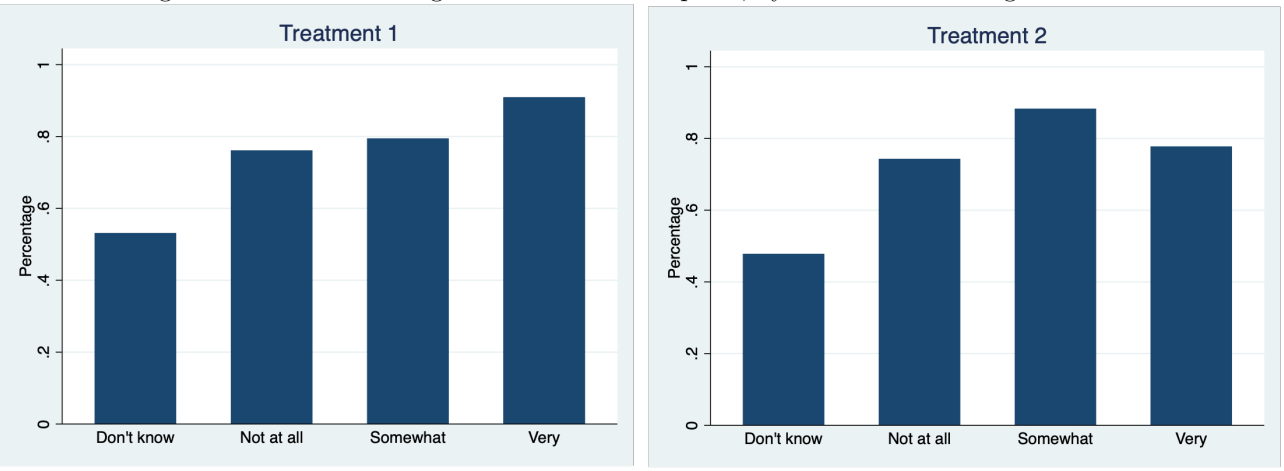
However, this effect is mitigated by the interaction of the treatment with position in the hierarchy. Specifically, the regression shows that middle and senior managers are 22 percentage points more likely to select the more effective option than their junior colleagues under treatment 1 (p=0.038). It appears that senior civil servants are less likely to be influenced by political preferences or disoriented by Ministerial engagement in the face of evidence than their more junior colleagues. Further, we cannot reject that $\beta_1 + \beta_3$ jointly sum to zero (p=0.2015). In other words, for senior officials, knowledge that the Minister has backed a less effective option does not affect their decision-making quality either way, neither inducing them to make more mistakes nor inspiring additional attention and hence greater accuracy. This is a highly encouraging finding, suggesting that the decision-makers most close to Ministers are those most likely to challenge them when necessary, but are not unduly confrontational. Though these findings are at first blush in line with work by Lakshmi Iyer and Anandi Mani, who suggest in the Indian context that junior staff may choose to maximise their promotion prospects through loyalty to political leaders rather than through ability or bureaucratic quality (Iyer and Mani 2012), it is important to note that this mechanism is not very plausible in this context. In the UK civil service, Ministers have little control over civil service appointments, and still less over junior appointments. It is possible that junior staff operate under the belief that disagreeing with Ministers is bad for their career, but this is speculative, and unlikely to survive extended exposure to civil service recruitment and promotion processes. It also has echoes of the mechanism proposed by Elston and Bevan (2021) to counter high-risk policy choices made by Ministers, though they emphasise the need to question policy objectives, and not simply the means used to achieve them (this experiment focuses on the latter).

This encouraging finding does not appear to be an artefact of the experiment, driven by respondents guessing the intention of the survey and adjusting their answers accordingly. Respondents might 'game' the experiment if they guess that, at mention of the Minister, the vignette is testing willingness to challenge them. If so, they may contradict the Minister in their response regardless of what they would do in real life. The results outlined above might be the

result of senior staff gaming the experiment more effectively. One way of testing whether or not this is the case is looking for signs of gaming by comparing the results of treatment 1 and treatment 2 among more senior staff, or among those self-reporting as most willing to challenge Ministers (as we will see later, willingness to challenge Ministers is closely correlated with seniority).

If gaming were prevalent among the higher grades, we would expect that those most comfortable challenging Ministers

Figure 2: Rate of selecting the more effective option, by comfort in challenge Ministers



respond to information about the Minister's preference by contradicting them. As such, they may select the opposite option to the Minister irrespective of treatment, leading to a higher rate of correct choices in Treatment 1 and a *lower* rate in treatment 2.[18]

The raw data suggests some such effect is observed (Figure 2). However, it is very small, and a Fisher's exact test finds no significant difference between the rate of selecting the more effective option among those very comfortable challenging Ministers across arms ($p = 0.665$). The observed difference in Figures 3 and 4 may therefore be a simple artefact of the data. Similar results are obtained when we look at seniority directly: there is a small dip in accuracy among the most senior civil servants in Treatment 2, but the effect is small and statistically insignificant. The key result, that needing to contradict the Minister lowers accuracy among junior staff, but has no effect on those most senior and likely to be in contact with them, does not therefore appear to be driven by gaming of the experiment.

7 Mechanisms

There are two immediately obvious reasons we might see a diminishing effect of treatment with seniority. The first is that the public sector is effective in promoting people with higher cognitive ability (at least in the interpretation of quantitative evidence, tested here), and this cognitive ability is especially important under pressure (as extra information about the Minister's preference might constitute, even if it has no incentive value). Self-assessed quantitative ability is controlled for in our main specification; however it may matter most only under pressure, which might be cap-

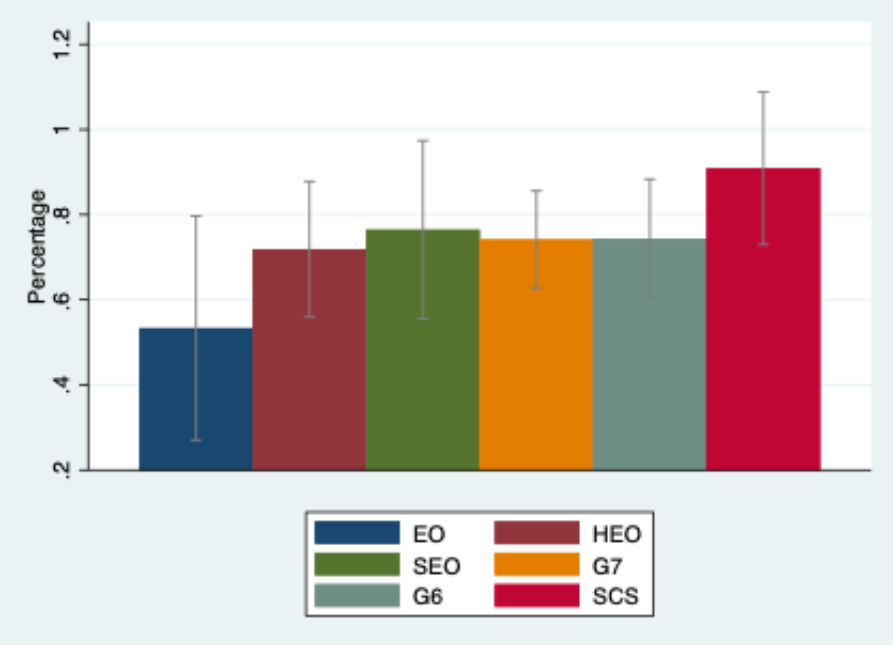
[18] This also helps explain why treatment 2 has no significant impact: it appears to only matter for those who know they are not at all or only somewhat comfortable challenging Ministers.

tured by the interaction between seniority and treatment. The second is that there is a relationship between position in the hierarchy and ability to challenge Ministers. This relationship could be causal, with promotion providing the institutional mandate, job security or training to contradict Ministers when necessary. It might also be selective, with the civil service identifying the most independent-minded individuals to promote, so individuals who are inherently more willing to challenge are most likely to be promoted.

If the civil service is able to identify and promote the most cognitively able individuals, we should see that when no incentive complications are introduced, more senior respondents do better at selecting the most effective option.

The raw data do not bear this out. In Figure 3, there appears little difference in selection of the correct option in the

Figure 3: Rate of selecting the more effective option, by comfort in challenge Ministers

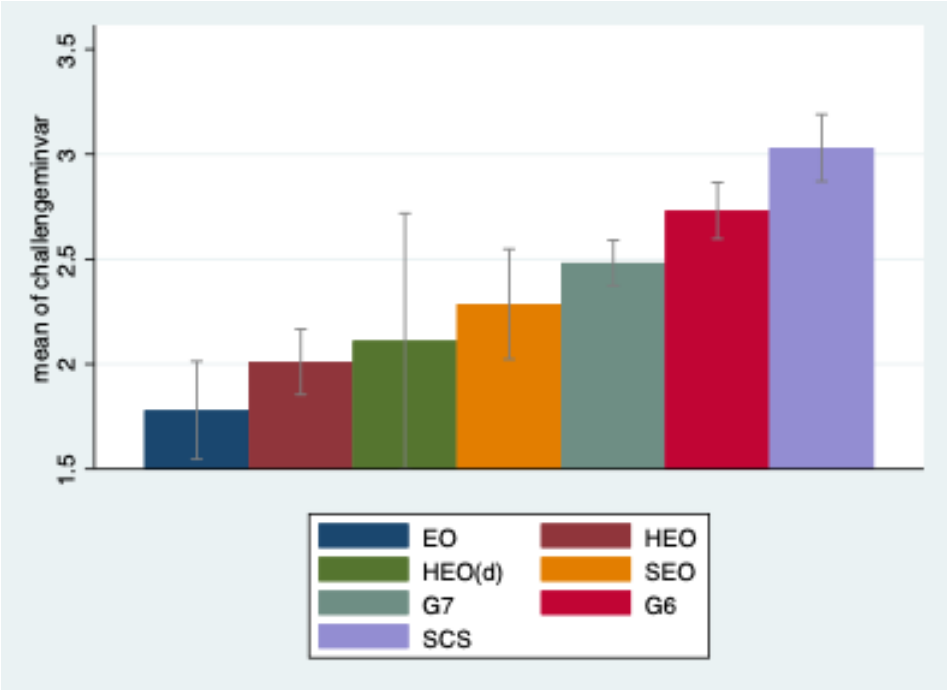


control group by seniority, as indicated by the almost entirely overlapping confidence intervals. Indeed, a regression of selecting the better option on seniority and controls, restricted to the control arm, returns no significant coefficients (n=173). It is not, therefore, that more senior staff are systematically more able to select the more effective option. However, it may be that ability matters particularly when stress is greatest – i.e. when the Minister is involved. We can test this by adding an interaction between self-assessed ability in handling quantitative evidence and treatment to specification 3, the results of which are reported as specification 5 in table 3. This provides evidence that the seniority effect is partly driven by ability under stress: while the coefficients and significance of treatment 1 barely change, the interaction effect between seniority and treatment 1 falls slightly (from 0.22 to 0.18) and loses significance (from p=0.038 to p=0.13). However, the coefficient on the interaction between ability and treatment 1 is insignificant. Ability under pressure is part of the explanation, but perhaps not the primary mechanism.

A second possible reason we observe treatment effect varying with position in the hierarchy is that promotion has a causal effect on the ability of civil servants to challenge Ministers, or is awarded to those most willing to do so. The survey provides some useful data here. All respondents were asked to report their self-assessed comfort in challenging Ministers; for this explanation to have weight we would expect a clear correlation with position in the hierarchy.

Figure 4 demonstrates that this is exactly what we observe: a clear step-wise increase in willingness to challenge

Figure 4: Average comfort challenging Ministers by civil service grade (0=don't know, 4=very comfortable)



Ministers as we move up the civil service grades (confirmed by regression analysis with a full set of controls). We can also test whether this comfort in challenging Ministers is driving the results observed in specification 3 by including it as a control variable in the regression, which is reported as specification 4 in Table 3. It is omitted in the preferred specifications precisely because it was expected to be at least one of the mediators through which effects of Ministerial opinion would be observed on bureaucrat choice (i.e. respondents select sub-optimal options because they are unwilling to challenge the Minister).

There is only a small effect from controlling for self-reported comfort confronting Ministers. The coefficient on treatment 1 over the sample falls slightly, corresponding to a reduction in the effect of treatment of 1 percentage point or so; and the coefficient on the interaction with hierarchy also falls by around 0.8 percentage points. These effect sizes are small, though in the direction expected. One explanation for these relatively small effects might be that self-reported comfort is a weak measure of actual willingness to challenge Ministers. Alternatively, it may be that comfort challenging Ministers only matters when Ministerial opinions are salient to the decision. Specification 6 of table 3 includes an interaction term between a dummy variable capturing high comfort challenging Ministers and treatment. The interactions are insignificant and have almost no impact on the coefficients of interest, nor their significance. Self-reported comfort challenging Ministers does not appear to explain the results.

If ability and comfort challenging Ministers aren't the primary drivers of better performance of senior managers in this task, what is? Though untestable from the data available, it is tempting to return to the civil service code, discussed earlier. The code is most familiar and salient to the more senior civil servants who most routinely find themselves giving advice to Ministers, either in person or in writing. Figures 1 and 2 clearly demonstrate that there is something more difficult about challenging Ministers than other senior figures. It is striking therefore, that among the senior

decision makers surveyed it is not just those with high ability or least uncomfortable with challenging Ministers who take this step, but most of them. The idea that an ethos that privileges impartiality and 'doing the right thing' even at a personal cost may trump self-interest or discomfort isn't new: Oliver Williamson suggested that probity would be more important than performance incentives in 'sovereign transactions' such as foreign policy, to which international aid is a close relation (Williamson 1999). The idea that the impact of hierarchy on organisational performance depends on the extent of the corrupting (or otherwise) influence of power is also widespread in the literature: the review by Anderson and Brown finds that empirical study offers support that this is an important moderator between hierarchy and performance (Anderson and Brown 2010). More recently, economists have increasingly investigated morality, identity and norms as determinants of economic behaviour and choices (Akerlof and Kranton 2010; Bowles 2016). Use of a code to encourage pro-organisational behaviour is in line with these ideas, and, if this is indeed the mechanism at work, suggests that the explicit and tacit components of the public service bargain at work in this department may be aligned, empowering civil servants to challenge Ministers (Elston 2016). This matters: it has been suggested that this kind of challenge function can also be used to counter high-risk policies proposed by Ministers due to 'one-shot bias', driven by Ministers being temporarily in positions of influence, though in this case the objective of the Minister is the object of challenge, not the means of achieving it (Elston and Bevan 2021). The next section discusses the implications of these findings.

8 Discussion

Organisational performance is a product of not only the selection and incentivisation of individuals, but the management processes they use and the ways in which they are combined. In some sense, the selection and incentivisation of staff determines (some of) the inputs of an organisational production function (especially talent and effort), while the management of them, the structures through which they interact and the ways in which they work together help turn these inputs into organisational outputs.

While we know an increasing amount about how best to select staff and incentivise their efforts (Ashraf et al. 2020; Bandiera and Lee 2015; Banuri and Keefer 2015; Besley and Ghatak 2005). But in elite organisations, which already make substantial investments in selection and contracting, how likely is it that we can find meaningful margins on which to improve? It may be necessary to accept that there is no feasibly adopted contract or selection process that completely eliminates error. This approach underlies the recent research into the use of algorithms (Anderson et al. 2016) and process technologies for error minimisation (Gawande 2010; Kahneman et al. 2019; Sibony et al. 2017). These approaches do not eradicate error; they merely provide an additional layer of process to reduce it. At worst, some can magnify bias, though there are ways to guard against this (Mullainathan and Obermeyer 2017). Further, few of these approaches are capable of handling incentives to select sub-optimal decisions, which this paper argues can be significant.

While these approaches tend to focus on the individual decision or decision maker (indeed some aim to eliminate

them), moving to a genuinely organisational approach to addressing errors generates different kinds of solutions. These solutions are built in to organisational structure and complement rather substitute for the approaches Gawande, Kahneman, Mullainathan and others observers put forward. I relate some of these solutions to the results of my experiment now.

Firstly, a resilient organisational decision-making process may need to build in redundant capacity or cognitive function, an approach that goes against the efficiency-maximising ethos of management consultancy and some mainstream economic thinking on industrial organisation. A number of studies find improvements in cognition at the group level (Ahmed 2017; Charness and Sutter 2012; George and Chattopadhyay 2008; Kao and Couzin 2014). It may pay to use groups to make decisions even when an individual can process the information alone. This does, of course, raise other issues, not least the possibility of groupthink if groups are constructed from like-minded raw material (Janis 1982).

Another approach is to explicitly build redundancy into the decision-making process, overlapping individual or team remits in such a way that any cognitive process is performed by at least two individuals before contributing to a final decision. In Edwin Hutchins' classic anthropological study of organisational decision-making, Cognition in the Wild, this is exactly the structure adopted by the navigational team he studied (Hutchins 1995). He examined how cognition proceeded both in routine circumstances and under conditions of high stress. His study demonstrated that particularly under pressure, 'redundant' cognitive capacity played an important role, with errors being picked up and questioned before a final decision was taken. However, getting the design of such overlapping remits right is not trivial. In a separate study of bureaucratic decision-making I find that the use of ex ante peer review or proposals before final decisions are taken has limited observable impact on final performance, with evidence of agents gaming the system apparent (Dissanayake and Ritchie 2022).

Secondly, decisions made by technically skilled and competent staff may still benefit from passing through serious review and scrutiny by seniors in the hierarchy before finalisation as a way of identifying and mitigating errors—though this comes at a cost in terms of time and cognitive load, and needs to be designed with care, lest it exacerbate problems of information overload rather than improve performance. Though this kind of 'slow thinking' is often held up as a problem of bureaucracy (when used as a pejorative), it may pay cognitive dividends through three channels. Firstly, it provides another way of creating 'cognitive redundancy' of the type Hutchins identified. Secondly, it can improve accuracy if the senior member of staff is more cognitively able under pressure; this may be more efficient if checking decisions allows the more sophisticated staff member to influence more decisions than doing them from scratch. And thirdly when personal incentives may run counter to the best choice for the organisation and stakes are high, it can provide an organisational buffer to the effects of misaligned incentives.[19] While research has examined the overall value of CEOs (Lieberson and O'Connor 1972), their biases (Malmendier and Tate 2015), the role of hierarchy in coordination or convening power (Anderson and Brown 2010) or in determining the boundaries of the firm (Hart and Moore 2000), the role of hierarchy in organisational cognition has received less attention. In the economics literature, early work suggested that managerial skill is a multiplicative term in a firm's production function (Lucas 1978; Rosen 1982).[20] If seniors in the hierarchy were carefully assessing the work of their juniors or essentially re-litigating key

[19] Of course, it's also possible that in some cases the malign incentives will affect only the senior staff and not the more junior.

[20] Though, as James March has observed, the exact identity of these managers may not matter, simply that they are chosen from a pool

aspects of their work, they might be generating significant value by improving their decisions, at least judging by the results of this experiment. However, the literature on hierarchy in organizations suggests that this is not what is happening. Rather, managers are more likely to suffer from information overload than to be able to process and examine the decisions and information provided to them by those in their span of control. As Herbert Simon wrote as far back as 1973: "the scarce resource is... processing capacity... Attention is the chief bottleneck in organizational activity, and the bottleneck becomes narrower and narrower as we move to the tops of organizations..." (Simon 1973). Meanwhile, bounded rationality can limit the ability of managers to understand the subtle messages transmitted up the chain towards them (Geanakoplos and Milgrom 1991) and it may be optimal for managers to commit to lower monitoring of employees through increased span of control (Aghion and Tirole 1997). While hierarchy could theoretically provide a valuable form of cognitive redundancy, the evidence suggests that in practice they typically do not. How, and indeed if it is possible, to rectify this may be a useful line of future inquiry. One line of investigation might be to consider the extent to which public bureaucracies promote 'specialised problem solvers', skilled in dealing with complex problems (such as managing conflicts between evidence on effectiveness and Ministerial preferences) in line with the model proposed by Garicano (2000).

Ultimately, all organisations are likely to be imperfect in multiple ways. From the beginnings of the theory of the firm (Coase 1937; Williamson 1985) and theories of public administration and organisational behaviour (Gibbons 2013; Simon 1997), emphasis has been placed on how organisations arise in response to imperfections in markets, information and contracting, and in turn may be characterised by multiple imperfections themselves. The true challenge is to recognise these various imperfections and to adapt organisational structures and processes to mitigate them. This experiment finds negative effects from introducing information of Ministerial preferences on decision-making. However, the civil service promotes the most cognitively able individuals and has a culture of independence and challenge (perhaps through the civil service code) to try and limit the effects of Ministerial preferences on technical advice. This is an organisational adaptation to an insurmountable organisational imperfection. The civil service code, and its supporting infrastructure may, then, be playing an important role in cognition.

References

Acemoglu, D. (2010). Modeling inefficient institutions. *Advances in Economics and Econometrics: Theory and Applications, Ninth World Congress, Volume I*, pages 341–380.

Aghion, P. and Tirole, J. (1997). Formal and Real Authority in Organizations. *The Journal of Political Economy*, 105(1):1–29.

Ahmed, H. (2017). Group Decision-Making : Evidence from a framed field experiment with community organisations in Pakistan [U+F047].

of near-identically competent candidates. See: https://www.theatlantic.com/magazine/archive/2009/06/do-ceos-matter/307437/

Akerlof, G. A. and Kranton, R. E. (2010). *Identity economics : how our identities shape our work, wages, and well-being*. Princeton University Press.

Anderson, A., Kleinberg, J., and Mullainathan, S. (2016). Assessing Human Error Against a Benchmark of Perfection. *KDD*, pages 705–714.

Anderson, C. and Brown, C. E. (2010). The functions and dysfunctions of hierarchy. *Research in Organizational Behavior*, 30(C):55–89.

Ashraf, N., Bandiera, O., Davenport, E., and Lee, S. S. (2020). Losing prosociality in the quest for talent? sorting, selection, and productivity in the delivery of public services. *The American economic review*, 110(5):1355–1394.

Ashraf, N., Bandiera, O., and Jack, B. K. (2014). No margin, no mission? A field experiment on incentives for public service delivery. *Journal of Public Economics*, 120:1–17.

Bandiera, O. and Lee, S. S. (2015). Do-Gooders and Go-Getters : Career Incentives , Selection , and Performance.

Banuri, S., Dercon, S., and Gauri, V. (2019). Biased Policy Professionals. *The World Bank economic review*, 33(2):310–327.

Banuri, S. and Keefer, P. E. (2015). Was Weber right ? the effects of pay for ability and pay for performance on pro-social motivation, ability and effort in the public sector.

Besley, T. (2007). *Principled agents? : the political economy of good government*. Oxford University Press.

Besley, T. and Ghatak, M. (2005). Competition and incentives with motivated agents.

Bowles, S. (2016). *The moral economy : why good incentives are no substitute for good citizens*.

Buchanan, J. and Tullock, G. (1962). *The Calculus of Consent*. University of Michigan Press, Ann Arbor, MI.

Charness, G. and Sutter, M. (2012). Groups Make Better Self-Interested Decisions. *Journal of Economic Perspectives*, 26(3):157–176.

Coase, R. (1937). The Nature of the Firm. *Economica*, 4(16):386–405.

Dewatripont, M., Jewitt, I., and Tirole, J. (1999). The Economics of Career Concerns, Part II: Application to Missions and Accountability of Government Agencies. *The Review of Economic Studies*, 66(1):199–217.

DFID (2011). How to note: Reviewing and Scoring Projects. Technical Report November 2011.

DFID (2016). Smart Rules for Better Programme Delivery. Technical report, DfID.

DFID (2020). Smart Rules. Technical Report April.

Dissanayake, R. and Ritchie, E. (2022). A Higher Bar or an Obstacle Course ?

Elston, T. (2016). Conflict between Explicit and Tacit Public Service Bargains in U.K. Executive Agencies. *Governance (Oxford)*, 30(1):85–104.

Elston, T. and Bevan, G. (2021). Using opportunity costs to counter "one-shot bias" in policy innovation. In Sullivan, H., Dickinson, H., and Henderson, H., editors, *The Palgrave Handbook of the Public Servant*. Palgrave Macmillan.

Garicano, L. (2000). Hierarchies and the Organization of Knowledge in Production. *The Journal of political economy*, 108(5):874–904.

Gavas, M. and Calleja, R. (2020). DfID is a world leader in tackling poverty. Our international standing is weakened without it — Aid — The Guardian.

Gawande, A. (2010). *The checklist manifesto : how to get things right*. Profile.

Geanakoplos, J. and Milgrom, P. (1991). A theory of hierarchies based on limited managerial attention. *Journal of the Japanese and international economies*, 5(3):205–225.

George, E. and Chattopadhyay, P. (2008). Group Composition and Decision Making. *The Oxford handbook of organizational decision making*, (October):361–379.

Gibbons, R. (2013). Cyert and March (1963) at Fifty: A Perspective from Organization Economics. *National Bureau of Economic Research*, NBER Organ(1963):1–11.

Hart, O. and Moore, J. (2000). On the Design of Hierarchies: Coordination Versus Specialization. *Ssrn*.

Hutchins, E. (1995). *Cognition in the wild*. MIT Press.

Iyer, L. and Mani, A. (2012). Traveling agents: political change and bureaucratic turnover in India.(Author abstract)(Report). *Review of Economics and Statistics*, 94(3):723.

Janis, I. L. I. L. (1982). *Groupthink : psychological studies of policy decisions and fiascoes*. Houghton Mifflin.

Kahan, D. M., Peters, E., Dawson, E. C., and Slovic, P. (2017). Motivated numeracy and enlightened self-government. *Behavioural Public Policy*, 1(1):54–86.

Kahneman, D., Lovallo, D. P., and Sibony, O. (2019). A structured approach to strategic decisions. *MIT Sloan Management Review*, 60(1):1–12.

Kahneman, D., Rosenfield, A. M., Gandhi, L., and Blaser, T. (2016). Noise: How to overcome the high, hidden cost of inconsistent decision making. *Harvard Business Review*, 2016(October).

Kao, A. B. and Couzin, I. D. (2014). Decision accuracy in complex environments is often maximized by small group sizes. 281(1784):1–8.

Leaver, C. (2009). Bureaucratic Minimal Squawk Behavior : Theory and Evidence from Regulatory Agencies. *American Economic Review*, 99(3):572–607.

Lieberson, S. and O'Connor, J. F. (1972). Leadership and Organizational Performance: A Study of Large Corporations. *American Sociological Review*, 37(2):117–130.

Lucas, R. E. (1978). On the Size Distribution of Business Firms. 9(2):508–523.

Malmendier, U. and Tate, G. (2015). Behavioral CEOs: The Role of Managerial Overconfidence. *Journal of Economic Perspectives*, 29(4):37–60.

Mullainathan, S. and Obermeyer, Z. (2017). Does machine learning automate moral hazard and error? *American Economic Review*, 107(5):476–480.

Neumark, D. (2018). Experimental research on labor market discrimination. *Journal of Economic Literature*, 56(3):799–866.

Rasul, I. and Rogger, D. (2018). Management of Bureaucrats and Public Service Delivery: Evidence from the Nigerian Civil Service. *The Economic journal (London)*, 128(608):413–446.

Rasul, I., Rogger, D., and Williams, M. J. (2017). Management and Bureaucratic Effectiveness: a Scientific Replication.

Rogger, D. and Somani, R. (2018). Hierarchy and Information.

Rosen, S. (1982). Authority, Control, and the Distribution of Earnings. *The Bell Journal of Economics*, 13(2):311–323.

Sasse, T. and Norris, E. (2019). Moving On: The costs of high staff turnover in the civil service. *Institute for Government*.

Sibony, O., Lovallo, D., and Powell, T. C. (2017). Behavioral Strategy and the Strategic Decision Architecture of the Firm. *California Management Review*, 59(3):5–21.

Simon, H. A. (1973). Applying Information Technology to Organization Design. *Public Administration Review*, 33(3):268–278.

Simon, H. A. (1997). *Administrative Behavior: A study of decision- making processes in administrative organizations*.

Simon, H. A. H. A., Smithburg, D. W., and Thompson, V. A. (1991). *Public administration*. Transaction Publishers.

Thaler, R. H. (1987). Anomalies: The January Effect. *Journal of Economic Perspectives*, 1(1):197–201.

Williamson, O. E. (1985). The Economics of Organization: The Transaction Cost Approach. *American Journal of Sociology*, 87(3):548–577.

Williamson, O. E. (1999). Public and Private Bureaucracies : A Transaction Cost Economics Perspective. *Journal of Law, Economics, & Organization*, 15(1):306–342.

Appendix A Robustness tests: Alternative definitions of seniority

In the main specifications reported, seniority is defined as a dummy variable taking the value 1 for middle and senior managers and 0 for junior staff. As a robustness test, we run the same specifications using alternative ways of coding the seniority variable. First, we simply code seniority as a continuous variable taking the value 0 for the most junior grade in our sample (which, recall, is restricted to only those officials with some decision-making responsibility) and 6 for the most senior stratum (the 'Senior Civil Service'). Secondly, we create three groups: junior, middle and senior staff. Tables 4 and 5 provide the results.

Table 4: Seniority coded as a continuous variable

DV: Selected the more effective option	Specification (all models are LPM)					
Model	(1)	(2)	(3)	(4)	(5)	(6)
T1: Minister backs less effective option	0.011	0.004	-0.183*	-0.184*	-0.209*	-0.185*
	(0.05)	(0.045)	(0.102)	(0.102)	(0.103)	(0.103)
T2: Minister backs more effective option	0.026	0.010	-0.080	-0.082	-0.119	-0.079
	(0.047)	(0.046)	(0.111)	(0.111)	(0.111)	(0.111)
T1 x Seniority			0.057**	0.058**	0.046*	0.057**
			(0.025)	(0.025)	(0.027)	(0.026)
T2 x Seniority			0.027	0.028	0.012	0.028
			(0.028)	(0.028)	(0.029)	(0.028)
Seniority		0.042***	0.013	0.012	0.022	0.012
		(0.014)	(0.021)	(0.021)	(0.022)	(0.022)
Ability		0.057	0.057	0.051	-0.058	0.051
		(0.039)	(0.039)	(0.040)	(0.074)	(0.041)
Comfort challenging Ministers				0.059		0.065
				(0.060)		(0.093)
T1 x Comfort Challenging Ministers						0.060
						(0.142)
T2 x Comfort Challenging Ministers						-0.52
						(0.138)
T1 x Ability					0.137	
					(0.94)	
T2 x Ability					0.198**	
					(0.097)	
Tenure (>1 <5 years)		0.139**	0.143***	0.146***	0.150***	0.147***
		(0.056)	(0.056)	(0.055)	(0.056)	(0.056)
Tenure (>5<10 years)		0.046	0.050	0.055	0.053	0.054
		(0.060)	(0.059)	(0.059)	(0.060)	(0.060)
Tenure (>10 years)		0.038	0.040	0.043	0.034	0.041
		(0.059)	(0.059)	(0.059)	(0.057)	(0.059)
Constant	0.440***	0.579***	0.575***	0.575***	0.604***	0.573***
	(0.082)	(0.078)	(0.096)	(0.096)	(0.091)	(0.096)
Controls	N	Y	Y	Y	Y	Y
Comfort challenging Ministers controlled?	N	N	N	Y	N	Y
R-squared	0.0006	0.1108	0.1201	0.1214	0.1279	0.1222
N	536	536	536	536	536	536

Notes: Robust standard errors in parentheses * Statistically significant at the 10% level ** Statistically significant at the 5% level *** Statistically significant at the 1% level Basic controls: seniority coded as 0 to 6, increasing in seniority; a dummy variable coding whether or not the staff member was recruited in country as opposed to being put through the full recruitment process in the UK; a dummy variable for whether or not the respondent self-assesses as having high ability in interpreting quantitative evidence (levels 4 or 5 from the 5 point likert scale); a dummy variable for whether teacher incentives was the 'correct' choice; and gender, tenure and professional cadre.

Table 5: Seniority coded to junior, middle and senior management

DV: Selected the more effective option	Specification (all models are LPM)					
Model	(1)	(2)	(3)	(4)	(5)	(6)
T1: Minister backs less effective option	0.011	0.003	-0.158*	-0.156*	-0.181*	-0.157
	(0.05)	(0.045)	(0.095)	(0.095)	(0.095)	(0.095)
T2: Minister backs more effective option	0.026	0.008	-0.111	-0.112	-0.149	-0.109
	(0.047)	(0.046)	(0.103)	(0.103)	(0.104)	(0.104)
T1 x Middle			0.207*	0.208*	0.155	0.0204*
			(0.116)	(0.0911)	(0.122)	(0.116)
T1 x Senior			0.249**	0.252**	0.201	0.250**
			(0.121)	(0.121)	(0.129)	(0.122)
T2 x Middle			0.176	0.180	0.113	0.179
			(0.122)	(0.122)	(0.127)	(0.122)
T2 x Senior			0.135	0.135	0.067	0.139
			(0.132)	(0.132)	(0.139)	(0.133)
Middle manager		0.154***	0.023	0.0208	0.067	0.021
		(0.057)	(0.091)	(0.091)	(0.094)	(0.091)
Senior manager		0.183***	0.051	0.045	0.096	0.046
		(0.068)	(0.103)	(0.103)	(0.105)	(0.104)
Ability		0.057	0.059	0.053	-0.055	0.053
		(0.039)	(0.040)	(0.040)	(0.075)	(0.041)
Comfort challenging Ministers				0.059		0.069
				(0.060)		(0.141)
T1 x Comfort Challenging Ministers						0.042
						(0.141)
T2 x Comfort Challenging Ministers						-0.043
						(0.138)
T1 x Ability					0.137	
					(0.95)	
T2 x Ability					0.192**	
					(0.097)	
Tenure (>1 <5 years)		0.148***	0.155***	0.158***	0.160***	0.158***
		(0.056)	(0.056)	(0.056)	(0.056)	(0.056)
Tenure (>5<10 years)		0.058	0.065	0.071	0.067	0.069
		(0.061)	(0.061)	(0.061)	(0.061)	(0.061)
Tenure (>10 years)		0.057	0.059	0.063	0.053	0.061
		(0.061)	(0.060)	(0.060)	(0.061)	(0.061)
Constant	0.487***	0.579***	0.580***	0.574***	0.605***	0.573***
	(0.079)	(0.078)	(0.092)	(0.093)	(0.093)	(0.093)
Controls	N	Y	Y	Y	Y	Y
Comfort challenging Ministers controlled?	N	N	N	Y	N	Y
R-squared	0.0006	0.1065	0.1210	0.1225	0.1282	0.1229
N	536	536	536	536	536	536

Notes: Robust standard errors in parentheses * Statistically significant at the 10% level ** Statistically significant at the 5% level *** Statistically significant at the 1% level Basic controls: seniority to three levels: junior, middle and senior staff, with junior staff the reference level; a dummy variable coding whether or not the staff member was recruited in country as opposed to being put through the full recruitment process in the UK; a dummy variable for whether or not the respondent self-assesses as having high ability in interpreting quantitative evidence (levels 4 or 5 from the 5 point likert scale); a dummy variable for whether teacher incentives was the 'correct' choice; and gender, tenure and professional cadre.

Both sets of results preserve the direction, relative magnitude and significance of the main specifications. The sole exception is specification 6 when seniority is coded into three levels, where the coefficient on treatment 1 falls just short of statistical significance at the 10% level (p=0.101).

Appendix B Robustness tests: Logistic regression

As a robustness test, we run the main specifications using a logistic regression. Table 6 shows that the main results are all replicated using this approach.

Table 6: Effect of information on Ministers' preferences on selection of the most effective investment option

DV: Selected the more effective option	Specification (all models are logistic)					
Model	(1)	(2)	(3)	(4)	(5)	(6)
T1: Minister backs less effective option	1.06	1.00	0.46*	0.47*	0.40*	0.46*
	(0.253)	(0.262)	(0.205)	(0.206)	(0.177)	(0.206)
T2: Minister backs more effective option	1.15	1.04	0.58	0.57	0.45	0.59
	(0.283)	(0.281)	(0.274)	(0.271)	(0.220)	(0.282)
T1 x Seniority			3.32**	3.35**	2.44	3.31**
			(1.815)	(1.830)	(1.44)	(1.803)
T2 x Seniority			2.40	2.42	1.67	2.43
			(1.377)	(1.396)	(1.018)	(.404)
Seniority		2.24***	1.11	1.10	1.45	1.09
		(0.658)	(0.500)	(0.493)	(0.688)	(0.490)
Ability		1.43	1.45	1.40	0.714	1.39
		(0.340)	(0.350)	(0.345)	(0.306)	(347)
Comfort challenging Ministers				1.53		1.75
				(0.710)		(1.433)
T1 x Comfort Challenging Ministers						1.60
						(2.354)
T2 x Comfort Challenging Ministers						0.61
						(0.634)
T1 x Ability					2.46	
					(1.404)	
T2 x Ability					3.42**	
					(1.992)	
Tenure (>1 <5 years)		2.25**	2.35***	2.38***	2.46***	2.39***
		(0.709)	(0.751)	(0.758)	(0.785)	(0.757)
Tenure (>5<10 years)		1.40	1.45	1.49	1.47	1.46
		(0.447)	(0.460)	(0.472)	(0.475)	(0.468)
Tenure (>10 years)		1.46	1.47	1.48	1.40	1.46
		(0.437)	(0.444)	(0.448)	(0.429)	(0.447)
Constant	2.76***	0.814	1.28	1.25	1.49	1.23
	(0.476)	(0.321)	(0.587)	(0.577)	(0.700)	(0.569)
Controls	N	Y	Y	Y	Y	Y
Comfort challenging Ministers controlled?	N	N	N	Y	N	Y
Pseudo R-squared	0.0005	0.1005	0.1090	0.1105	0.1172	0.1116
N	536	536	536	536	536	536

Notes: Robust standard errors in parentheses * Statistically significant at the 10% level ** Statistically significant at the 5% level *** Statistically significant at the 1% level Basic controls: a dummy variable taking the value 1 if the respondent is a middle or senior manager and 0 if not; a dummy variable coding whether or not the staff member was recruited in country as opposed to being put through the full recruitment process in the UK; a dummy variable for whether or not the respondent self-assesses as having high ability in interpreting quantitative evidence (levels 4 or 5 from the 5 point likert scale); a dummy variable for whether teacher incentives was the 'correct' choice; and gender, tenure and professional cadre.

Computing the average marginal effect of each reported variable and comparing them with the coefficients obtained from the LPM model confirms that the effect sizes are very similar to those obtained using the main LPM specification.

3

A Higher Bar or an Obstacle Course? Peer Review and Organizational Decision Making in an International Development Bureaucracy

A Higher Bar or Obstacle Course? Peer Review and Organizational Decision Making in an International Development Bureaucracy

Ranil Dissanayake [*] Euan Ritchie [†]

May 2022

Abstract

Many public organizations employ mechanisms of scrutiny such as peer review or quality assurance to improve their performance and decision-making. Such mechanisms may affect performance and decision-making directly, through scrutiny, and indirectly, through behavioural responses by agents within the organization. We examine one such policy in a large public sector organization in the UK. By comparing the distribution of project sizes before and after the introduction of a system of assurance implemented through a simple decision-rule, we document substantial manipulation in the area around the threshold for review by agents designed to avoid scrutiny. Furthermore, there is suggestive evidence that the relative performance of projects qualifying for review improves after the implementation of the review system, as measured by fidelity to the planned completion date and annual review scores, though we cannot distinguish between causal effects of review and selection effects of review avoidance. Our results suggest that organisations considering such a system of scrutiny need to investigate both the naïve effect of the policy, and how agents will respond to its existence, setting a new organisational equilibrium.

Keywords: Bureaucracy, Public Organization, Organizational Behaviour

1 Introduction

The improvement of public sector decision-making and performance is of first-order importance to public welfare (Kaufman 2021). Three broad strategies for improving each have been studied: how they choose and articulate their objectives (for example, Rainey (1993)); how they are optimally organized to deliver them (e.g. Williamson (1999, 2002)); how they and their constituent agents pursue public and private objectives (e.g. Besley (2007); Buchanan and Tullock (1962)). At heart, this body of literature identifies three ways to improve public sector performance (taking the objective as given): hire better people; incentivize those hired better; and structure them and their work better. One commonly-adopted structure in the public sector is the use of ex ante peer review or quality assurance of proposed projects or activities. Such systems exist in South Korea (where the Public and Private Infrastructure Investment

[*]Center for Global Development and Blavatnik School of Government, University of Oxford
[†]Center for Global Development

Management Centre can be seen as a tool for quality assurance) [1]; across the UK Government (where certain project proposals are subject to the Treasury's Major Project Approval and Assurance system)[2]; and at international institutions such as the World Bank, where peer review of project proposals is commonplace. Such systems are rarely applied universally to all business undertaken by a Government or agency. Usually, some decision-rule governing whether a project must be subject to review or assurance systems exists. However, we know surprisingly little about the effect such systems have on performance and organizational decision-making. Organisations that adopt review and/or assurance processes usually assume that greater scrutiny leads to better performance, but this should not be taken as given. There is a risk that such scrutiny will signal distrust (Arnaud and Chandon 2013; Deci and Cascio 1972) especially in contexts in which staff feel personal affinity to the task or organization (Frey 1993), and thus decrease worker effort. This "crowding-out" of intrinsic motivation may outweigh benefits from closer supervision, especially in organisations in which intrinsic motivation is high to begin with (Bertelli 2006).

The net effect of a peer review and assurance system depends on both the direct effect of review (which might be positive, negative or negligible) and its indirect effect on agent behaviour. When decision-rules determining eligibility for review are transparent (in that agents within the organization know or are able to infer the rule), agents may change their behaviour to avoid review, thereby distorting organizational behaviour.[3] As such the net effect of review and assurance systems may be positive (if the direct benefits from review outweigh costs of distortion) or negative.

This paper investigates the implementation of a system of ex ante peer review and quality assurance in what was the UK's main foreign aid bureaucracy, the Department for International Development (DFID, which existed from 1997 till 2020). The system was implemented as a key part of the organisation's 'Better Delivery' infrastructure and played an important role in the decision-making process for project approval or rejection, and can be seen as part of a broader trend towards exercising bureaucratic control via audits and reviews (Hoggett 1996). The review system was applied using a clear decision rule, with only projects valued above £40 million (or deemed 'novel and contentious', in practice a very small number) subject to review. Using a novel dataset, obtained by scraping information from a public database of project documents from DFID, we use standard tests of manipulation around a discontinuity to document that the establishment of this system resulted in widespread avoidance of review and distorted the organizational spending profile as agents sought to avoid the review process. We further document suggestive evidence that the relative performance of projects that qualified for review compared with those that did not improved after the implementation of the review system, as measured by project review scores and timeliness of project completion, though we find no effect on over- or under-spending relative to budget, and note that these results cannot distinguish between the causal effect of review and the selection effect of review-avoidance, and are sensitive to our choice of estimation strategy. Finally, we discuss possible mechanisms driving these results. Our results are similar to those found in the context of decision-rules constraining bureaucrat discretion in public procurement in Italy, where similar manipulation is observed, with mixed effects on performance (Coviello et al. 2021). This study extends such findings

[1] A summary of the mandate and mode of operation of PIMAC is available here: https://www.kdi.re.kr/kdi_eng/kdicenter/pimac_main.jsp

[2] The guidelines for this process are set out here: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/179763/major_projects_approvals_assurance_guidance.PDF.pdf)

[3] There may also be different behaviour change when these rules are not transparent but are known to exist.

to a broader segment of public policy than procurement.

This work contributes to an empirical literature that investigates how organizational decision-making is structured and governed, and its effect on performance, by documenting distortionary unintended consequences of the system adopted, complementing Martinez et. al. (2015) who find that organizational culture is an important determinant of the impact of such systems, and Coviello, Spagnolo and Lotti (2021) who find similar results in a public procurement setting. It also contributes to a literature on the impact of evaluation, peer review and audit on organizational performance (for example: Crijns, Ottenhoff, and Ring 2021; Higgs and Gelman 2021; Kells 2011; Morin 2001; Soderberg et al. 2021) by documenting an ex ante review process in a public sector organisation and the effect of peer review in a non-academic setting. It also complements work documenting the gaming of audit systems in the UK (Elston and Zhang 2022).

The paper proceeds as follows: the next section briefly discusses the literature on how organizations can structure and govern the work of their agents to improve decision-making, and the role of peer review and quality assurance specifically, and how this study contributes to them. The following sections discuss the setting and specific organizational structures we study here; and the data available. The results section documents the effects of adopting the peer review and assurance system: its effect on the decision-making of bureaucrats, its effect on project performance and its effect on the structure of the organization's overall portfolio; the final section concludes with a discussion of the results.

2 Motivation and Contribution

The ways in which organizations can structure and govern decision-making to improve performance have been studied in a number of settings. In the public sector, important decisions are almost always made through human deliberation, often among senior political figures though they may be advised by technocrats and career civil servants. How this process is structured and governed, and the systems used to support decision-making is of substantial policy importance.

Organizational choices in structuring the relationship between different agents, in governing the processes which they follow or in making different kinds of input and support available to them may all play a role in supporting such human deliberation. Martinez et. al. (2015) study the use of an algorithm (a simple checklist) in a hospital setting to improve decision-making in diagnosis and find that the adoption of the system improved both diagnosis and health outcomes – but note that this outcome was driven in large part by a cultural change in the hospital studied, which made staff more receptive to the algorithm's use. Nicholas Bloom and co-authors argue that the management systems and practices adopted by an organization matter for its performance (Bloom et al. 2016), and investigate its application in the school sector in the UK, finding a positive association between the adoption of good management practices and school performance, as measured by teacher outcomes (Bloom et al. 2015). In a developing country setting, Dunsch et. al. (2021) finds that an intensive management intervention generates large short term effects, but that after one year, these effects had disappeared, suggesting that short term adoption of performance-enhancing improvement does not guarantee that it is sustained. In the private sector Sibony et. al. suggest that random variation in decision-making

outcomes within an organization can be mitigated a simple checklist for minimizing error – much like the algorithms studied by Martinez, and the checklists advocated by Atul Gawande (Gawande 2010; Kahneman et al. 2016, 2019; Sibony et al. 2017).

The role of evaluation and audit as a mechanism for improving performance in public organisations has also been extensively studied. Most of these studies consider ex post evaluation, usually undertaken by an independent evaluation office or supreme audit body (internal ex post performance audit is also common). Though initially developed in its modern form in the US, operational value-for-money auditing has deep roots (extending back at least to 17th Century Britain), and has become commonplace around the world, with different countries using different approaches (Flesher and Zarzeski 2002). When such audit takes place ex ante, it can be understood as a system of governance for decision-making, evaluating and assessing the expected return of a decision or the quality of a proposal. Such evaluation can take the form of peer review (as is common in academia) or quality assurance. These are related but distinct approaches, and not mutually exclusive. Quality assurance typically has some institutional basis and is vested with authority through either institutional relationships or hierarchy. Peer review, by contrast, tends to occur horizontally, when agents engaged in similar tasks or on similar work assess each others' work.

Such systems are costly. Peer review in academia is time-consuming for both reviewer and reviewee. One study found that the monetary value of the time US-based peer reviewers spent on reviews in 2020 was over $1.5 billion (Aczel et al. 2021). Another found that the peer review and revisions process in economics took around six to nine months in the 1970s, and by the early 2000s had stretched to several years (Ellison 2002). An attempt to speed this up in the Journal of Public Economics managed to shorten the time taken for initial review by several days, but it remains notable that most reviewers still missed their deadlines (Chetty et al. 2014). Assurance processes can involve a substantial financial outlay to staff and equip a body with some hierarchical or organizational standing to critique others. The net operating expenditure of the UK's National Audit Office, for example, was around £85 million in 2020/21.[4]

Though form varies, audit and peer review can have a number of functions. The most common justifications are instrumental (to improve performance and quality) or intrinsic to the activity (for the value of security and transparency for its own sake), with great emphasis often placed on the former (Kells 2011; Lonsdale 2000). However, assurance and peer review might have other purposes, too. They may be a form of outward-facing theatre, designed to project the image of a careful, technocratic and evidence-driven organizational, regardless of the reality. They may also be a piece of internal theatre, designed to inculcate a culture or identity of rigour, challenge and value-for-money, irrespective of the direct effects of peer review and assurance. In this study, we focus on the merits of the instrumental justification for the system of peer review studied, taking organizational statements of purpose at face value.

The instrumental case for peer review and assurance is typical made on the basis of its direct effects on performance. For ex ante review, as we study here, these may include making worse proposals or programmes (such as those motivated by pleasing political leaders rather than expected quality (Dissanayake 2022), or sub-standard proposals submitted

[4] As reported in its Annual Report and Accounts 2020-21, accessed on 30/12/21, at https://www.nao.org.uk/wp-content/uploads/2021/06/NAO-Annual-Report-and-Accounts-20-21.pdf

due to pressure to spend allocated funds[5]) less likely to be approved, as with peer review for academic journals. It could also increase prospective effort by agents in the knowledge that it will be quality assured (the 'chilling effect of audit' (Flesher and Zarzeski 2002), or through the effect of organizational monitoring (Gibbons 2016)). It may apply additional information or cognitive capacity to proposals, a form of cognitive redundancy (Hutchins 1995) which may be important in the context of bounded rationality (Simon 1997) or when mistakes are common (Kahneman et al. 2016). And it may ameliorate behavioural or cognitive biases that are widespread in the public sector (Banuri et al. 2017). To the extent that any of these operate, we would expect them to improve average portfolio quality at least among the proposals subject to review.

However, the existence of peer review and assurance may also have indirect effects, potentially changing the way in which behaviour is rewarded in unintended ways (Kerr 1995), and subsequently, how agents structure their work and proposals. Put simply, if review comes at a cost, the desire to minimize or avoid this cost may induce agents to manipulate their proposals to avoid peer review. Similar phenomena have been found elsewhere, in particular public procurement where it has been documented extensively. In Italy, Coviello, Spagnolo and Lotti (2021) find that agents manipulate the value of the contracts they issue so as to avoid rule-based procurement systems and to retain personal discretion. Palguta and Pertold (2017) find a similar result in the Czech Republic, and in a recent working paper, Tas (2019) finds that in EU jurisdictions, there is a high probability of bunching of contract values just below thresholds above which public procurement is subject to stricter rules. If there is a transparent decision-rule of this type that triggers additional review, assurance or oversight, similar behaviour may be observed.

The net effect of these channels of impact is uncertain. The cost of reviews or audits is an increasing function of its quality (Power 1999). Peer reviewing has a cost that doesn't necessarily scale with project size, and so the benefits of reviewing smaller projects may not exceed the costs. Molander (2014) found that the average cost of public procurement rules begins to exceed the average benefits at project sizes around €5,000 in the Swedish government procurement setting.

In addition, much of the literature on audit is devoted to ways in which it can fail (Kells 2011; Morin 2001), including by avoidance of audit altogether. While many peer review systems are universal (that is, all proposals must be reviewed and there is no possibility of avoidance), peer review suggestions are often incorporated only cosmetically, with a recent study finding that the majority of accepted suggestions - themselves only around one third of the reviewer suggestions made, were taken on in the title or abstract of a paper only (Crijns et al. 2021), raising the question of whether the substantial cost of the peer review process is worthwhile. Similarly, even when audit itself cannot be avoided, reporting against how the recommendations of audit have been actioned may be, perhaps limiting its effectiveness (Elston and Zhang 2022). Furthermore, peer review can achieve little in some cases, when scientists lie outright, or adapt their behaviour to the existence of peer review in their pursuit of academic kudos and rewards (Bright 2021). And even successfully navigated peer review offers little guarantee that published research findings replicate or are

[5] A pressure that was widely anticipated in DFID, given that its precursor organization. the Overseas Development Administration, housed in the UK's Foreign Office was put under pressure to mis-spend funds illegally in support of foreign policy aims, specifically on the Pergau Dam (Lankester 2013). Secondly by 2011, when the Quality Assurance Unit was set up, DFID's budget was increasing rapidly, and there were attempts at establishing a legally-mandated spending floor for the Department (Dissanayake 2021), as eventually came to pass. A common criticism of this legislation, before it was enacted, was that it would create a pressure to 'get money out the door' and hence a downward pressure on spending quality.

robust (Ioannidis 2019; Della Vigna and Linos 2020). Recent work has found that ex ante peer review, that is reviews of research designs rather than the completed research (analogous to the process we study, elaborated below, which consists of peer reviewing intervention proposals, rather than the intervention itself), performs at least as well as the normal, pre-publication but post-research completion review (Higgs and Gelman 2021; Soderberg et al. 2021).

This study contributes to each of these literatures. Our primary research question is whether the implementation of a new system governing decision-making resulted in unintended consequences for the organizational portfolio of activities, by generating an 'indirect channel' response by agents in the organization. Our secondary research question is whether there is any associated difference in the observable quality of projects subject to the new system and those that are not. In examining how a new governance structure for decision-making with the purpose of improving performance was implemented we document the existence of distortionary unintended consequences of the system, contributing to this literature and complementing the work by Martinez et. al. (2015), who find that organizational culture is of critical importance for the successful adoption of a new decision-making structures, and Coviello, Spagnolo and Lotti (2021) who document a similar effect in a more restricted domain of public sector decision-making. By considering metrics of project quality for reviewed and non-reviewed projects, we contribute to the literature on assurance (introducing an example conducted ex ante) and peer review (introducing an example conducted in a non-academic setting).

3 Setting

Our setting is what was the UK's primary aid bureaucracy during the period 2011 to 2020. The Department for International Development (DFID) was a highly respected aid bureaucracy with a strong global reputation, winning praise by other donors and foreign politicians (UK Parliament 2020), think tanks (Gavas and Calleja 2020), and independent institutions set up in the UK to govern UK public finances generally and development policy and spending specifically (Mitchell and Baker 2019). It was created in 1997 as a separate Government department; in 2020, it was merged with the Foreign and Commonwealth Office (FCO) to form the Foreign, Commonwealth and Development Office (FCDO), though for the first year after this merger, the FCO and DFID still operated separate systems to manage their spending (which had already been allocated to the departments individually that year). Despite its strong reputation the DFID enjoyed a difficult relationship with the UK print media, which often portrayed aid spending as wasteful.[6] This adversarial relationship may have heightened political desire for control-processes within the department (Carpenter and Krause 2012).

DFID staff tended to report high levels of intrinsic motivation (as evidenced by the Civil Service People Survey), had among the lowest staff turnover of any government departments, and were the least likely to report an intention to leave (Sasse and Norris 2019). This could be consistent with staff whose preferences were highly aligned with the department's and for whom the degree of managerial oversight may not be expected to decrease effort (Bertelli 2007; Brehm and Gates 1997).

The decision-making structures adopted by DFID had substantial public policy and global welfare consequences. In

[6] See, for example, this page in which the Government rebuts various claims of waste and mismanagement made in the media: https://www.gov.uk/government/news/media-reports-on-uk-aid-projects-setting-the-record-straight

2020, the year of the merger, DFID (as distinct from FCDO or FCO) disbursed £10 billion in Official Development Assistance, equivalent to 0.7 percent of Gross National Income, a level which it was legally required to reach each year since 2015. In the years covered by this paper, it disbursed on average £9.7 billion each year, and a total of £97 billion pounds. The UK was routinely the most generous G7 donor as a proportion of GNI, and in absolute terms one of the largest funders of both multilateral development institutions and bilateral development projects. The global welfare implications of DFID's choices are clear from the decision to cut UK ODA in 2021, in response to which a number of research institutions (Kennedy McDade and Mao 2021), think tanks (Hares and Rose 2021; Mitchell et al. 2021) and NGOs (Watts 2021) estimated substantial costs in terms of lives saved, people reached and research projects foregone. To the extent that aid projects have different expected values (Banerjee et al. 2020) and that donors have decision-making agency to select which projects to fund in which countries (Briggs 2017), the decision-making process DFID adopted had real-world welfare implications.

DFID's organizational objectives were clearly stated. The legislative basis on which DFID was able to spend Official Development Assistance was given by the 2002 International Development Act, which stated: "In this Act "development assistance" means assistance provided for the purpose of—

(a) furthering sustainable development in one or more countries outside the United Kingdom, or

(b) improving the welfare of the population of one or more such countries."[7]

Alongside this legislative requirement for its activities to contribute to sustainable development DFID made a series of public commitments to achieve value for money. The Independent Commission on Aid Impact remarked: "Against the background of the 0.7% aid spending commitment, the pledge to achieve 100 pence of value for every pound of aid spent has become central to the political case for the aid programme." (ICAI 2018). These objectives were operationalised through internal processes and guidance aimed at measuring success in achieving stated project objectives, and close monitoring of the value-for-money of DFID spending. Project outputs and objectives were assessed through the Annual Review process, of which internal guidance wrote "Regular and effective monitoring, reviewing and lesson learning are key to how DFID measures the Results of its projects and demonstrates Value for Money (VfM). How we review our projects, including our approach to project [Annual Review] scoring, is important since it allows us to establish progress against planned Outputs in an objective and transparent way... At the Annual Review (AR) achievement against the Outputs will be scored alongside an assessment (but not a score) of the Outcome."(DFID 2011b). Similarly, DFID's Approach to Value-for-Money (VfM), a document published in July 2011 noted "How we manage our aid programme is vital. We need to have strong programme management to make sure our programmes stay on track, achieve the intended results, and are delivered on time and within budget" (DFID 2011).

DFID was widely praised for its approach to generating and using evidence, and the high standard of scrutiny it subjected itself to. It was subject to scrutiny by a Parliamentary Committee (the International Development Committee), a statutory body set up by Act of Parliament to scrutinize the effectiveness of aid spending (the Independent Commission for Aid Impact) and the routine scrutiny of the UK's National Audit Office, which undertook a number

[7] https://www.legislation.gov.uk/ukpga/2002/1/section/1

of investigations of DFID during the period 2011-20.

These institutions provided ex-post scrutiny that aims to investigate historical spending and activity with the aim of generating recommendations for future policy. They were complemented by an extensive internal structure of ex-ante scrutiny and evidence assessment aimed at improving the value-for-money and delivery of DFID spending primarily through: better decision-making over what specific projects and activities to invest in, and how to govern the management and procurement decisions for them; and on-going evaluation and learning during the life of and on the completion of projects.

This scrutiny takes three forms. The first is paper documentation, in the form of extensive 'business cases' that made the case for a project or funding stream and are submitted to Ministers for approval, and annual and project completion reviews which examine the performance of projects against its objectives; the second is guidelines that govern how this paper documentation should be acted upon: for example, if annual review scores are sufficiently low, a Project Improvement Plan (PIP) should be produced. The third is a formal process of peer review and assurance which assesses the quality of (some) of this documentation, and delivers written reports on this to the decision-makers involved.

This paper focuses on the third of these forms of scrutiny: formal peer review, specifically of the business cases on which spending decisions are made. These business cases may be anything between five and more than one hundred pages long, and provide detailed information on the project or funding stream being proposed. Typically, they set out a 'strategic case', that argues that the funding addresses an important problem or challenge, that the UK has some competence in addressing. Then they set out an 'appraisal case' which provides a number of options for how to address the strategic problem set out, usually including a 'do nothing' counterfactual. These appraisal cases were typically conducted by economists and usually included some quantitative assessment of the expected return to the proposed funding under different scenarios. A 'commercial case' set out the management and procurement implications of the different options (with the emphasis on the preferred one). A 'financial case' set out the funding implications and how finances will be managed and risks mitigated and monitored. A 'management case' lays out the proposed governance arrangements of the project or funding proposal.

Depending on the size of the proposed funding, approval or rejection of a business case lies at different levels of the department. For spending of less £5 million, a Head of Department (a member of the Senior Civil Service) could make the decision. For projects spending more than this, but less than £70 million a 'Junior Minister' (that is a politician appointed to a Ministerial portfolio within the Department) approved or rejected the business case. And for projects or funding streams spending more than £70 million, the Secretary of State (that is, the most senior political figure associated with the Department) made the decision.

Formal peer review and assurance of business cases was conducted by the Quality Assurance Unit (QAU) of the Department, established in 2011. QAU was an independent team, not answerable to any civil servant with a spending mandate (thereby avoiding direct conflicts of interest in its reporting line). The team had an assurance function and a peer review function. The Head of QAU was a career civil servant drawn from the Government Economic Service reporting directly to the Chief Economist of the organization, an academic economist of high standing recruited from

outside of the civil service and outside the normal rotation of civil service roles. The staff who reported to the Head of QAU were subject to the usual 3 year rotation, and are drawn from a variety of backgrounds. All had experience of writing business cases and would go on to write further business cases at the end of their rotation. Thus, QAU provided both an assurance function, through the Chief Economist who sits largely outside the normal departmental lines of accountability, and a peer review function, provided by the staff who worked on each business case. It reported on its overall activities to DFID's Investment Committee, chaired by the Director-General for Finance and Corporate Performance.

QAU reviewed every business case in DFID that satisfied either of the following criteria:

- A new business case proposal valued above £40 million.
- A new business case (of any value) that is 'novel or contentious'

In practice, the vast majority of proposals reviewed fell under the former category. For cost extensions that bring the total value of a project to over £40 million, or were in themselves over £40 million, the Director (a senior civil servant) overseeing the business case could use their discretion over whether or not to submit the business case for quality assurance.

Based on their review deliberations, QAU issued a short report based on "an evidence based assessment of the vfm [value for money] of the BC [business case] and its spending proposal".[8] The report was led by QAU staff, but might draw on expert reports commissioned from other civil servants working across the department. The reviewing team had discretion over who is asked to undertake these expert reports, though they were voluntary. The proposing team were not able to propose reviewers. All reviews are co-signed by the Chief Economist. This short report QAU produced was accompanied by a 1-4 score, with the scoring system and implications as follows:

- 1: Limited recommendations
- 2: Broader recommendations
- 3: Resubmission to assess recommendations
- 4: Resubmission to assess major recommendations

The score and QAU report were submitted to the team proposing the spending, copied to the Director to whom the proposing team report. In the case of scores 3 or 4, the extent to which recommendations have been addressed is assessed, but no further score or resubmission is required; the proposal may be submitted for approval. The QAU report was required to be appended to the business case upon submission to relevant Minister for approval. The QAU itself did not approve or reject spending proposals. It simply reviewed the proposal, with its review part of the decision-making process by the responsible Minister. Given that in the period 2011-2020, only two DFID Secretary of States had direct experience of running a development intervention, the QAU report was a potentially important input into Ministerial decision-making.[9] Ministers were in turn accountable to Parliament through and the public

[8] This section draws on personal communication by email with the Head of the Quality Assurance Unit, dated 02/06/21.

[9] Andrew Mitchell, who established Project Umumbano midway through his spell as Secretary of State, and Rory Stewart, who had such experience prior to his appointment, but was Secretary of State for International Development for just under three months.

through the mechanisms outlined above. In the case of anything going wrong, they might be compelled to disclose the advice received from civil servants which underlay their decision to approve (or reject) a given proposal to the National Audit Office or the International Development Committee in Parliament. It is within the Minister's power to ignore the content of the QAU report, but given that this is clear evidence of official advice as to the quality of a spending project, and by ignoring it the Minister invites personal responsibility should things go wrong, it was highly likely that projects with very negative reviews would struggle for approval in the absence of documented remedial action.

The quality assurance process is inexpensive, but not costless. A back-of-the-envelope calculation suggests that each review costs around £8,000 in staff time alone.[10] Additionally, it imposed a delay on the process of seeking funding approvals. QAU reserved the right to take up to five weeks to review and report back to proposing teams, and where a score of 3 or 4 is issued, a further several-week delay could be expected to address recommendations and resubmit.

QAU was formally justified on instrumental grounds. The DFID Smart Rules described QAU as "a key part of the second line of defence" in the pursuit of a higher quality portfolio (DFID 2020). It constituted the bulk of the 'internal scrutiny' component of DFID's Approach to Value-for-Money (DFID 2011). It did not appear to be a piece of outward-facing theatre, designed to convince the public or Parliament of the rigour of the aid portfolio. This is borne out by the extreme paucity of public references to the workings of QAU. There are no mentions in reports by the Independent Commission for Aid Impact—including in a 2014 review of DFID's Smart Rules, or a series of review documents of DFID's Approach to Value for Money in Programme and Portfolio Management (ICAI 2018, 2019), in Hansard (the record of proceedings in the UK Parliament) or—to the best of our knowledge—in the UK press.[11] It may have had some function of internal theatre and signaling of values, but certainly, its stated objectives were firmly functional, focused on improving performance. In this way it can be seen as a form of "accountability as continuous improvement" (Aucoin and Heintzman 2000), in which negative feedback from the peer review process will create pressure to improve the quality of future business cases submitted. While there were no formal sanctions for continual low scores from QAU, they were likely to have career consequences, and therefore officials did face consequences.[12] This view is supported by the existence of an annual document produced by the QAU summarizing the main reasons for low scores, disseminated with the purpose of improving the overall quality of proposals.

However, to the extent that peer review and assurance benefits those projects that go through it (which is not a given), the net effect of the review process will depend on the extent of the 'indirect channel' through which agents change their behaviour in response to the existence of peer review. In this setting, doing so is straightforward: since projects under £40 million are highly unlikely to be subject to peer review and assurance, expending effort to revise proposals

[10] This calculation is based on the staff cost of one A1 adviser (the Grade of the Head of QAU), one A2 adviser and one B1 adviser (both more junior grades which compose the majority of QAU staff, each spending 50% of their time on a review for the full five weeks it takes to produce a QAU report, plus half a day of time from the Chief Economist, to consider the Business Case, the QAU report and to sign off. Staff costs are taken from a 2015 Freedom of Information Request (https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/581487/DFID-staff-pay-bands.pdf) and data from glassdoor.co.uk for the Chief Economist's salary (https://www.glassdoor.co.uk/Salaries/london-chief-economist-salary-SRCH_IL.0,6_IM1035_KO7,22.htm), both accessed on 17/05/22

[11] We searched Google News for mentions of 'Quality Assurance Unit' and 'DFID' in in all UK news sources. There were just three matches, of which two come from the UK Government website, GOV.uk. These each mention QAU a total of three times between them, once in the context of the Chief Economist's job description. The remaining mention was an unrelated keyword match. We searched Hansard for mentions of 'Quality Assurance Unit' or 'QAU' and no relevant matches were returned.

[12] The QAU process therefore meets the criteria set out in Bovens (2007) to be classed as form of 'narrow accountability'.

to fit under the £40 million cut off dramatically reduces the risk of review. This could be done by splitting a project in half (turning a £60 million project into two £30 million projects), trimming project size (so reducing the scope and ambition of a £45 million point project until it is smaller than £40 million) or by department or unit managers planning a portfolio made up of many smaller projects rather than a few large ones. The next section sets out the data we use to investigate this.

4 Data

We generate a novel dataset on DFID's activities before and after the establishment of peer review using publicly available documentation. Since making a policy decision to adhere to the strictures of the International Aid Transparency Initiative (IATI) in 2011, DFID began uploading information and documentation relating to virtually every project it approved or active from that point onwards to a database called DevTracker. The 2011 start date means that we have information on projects approved since the establishment of QAU, and information on every project that was still being implemented in 2011 but approved before the establishment of QAU.

The documentation uploaded includes Business Cases, Annual Reviews, Project Completion Reviews and various addenda (including applications for no-cost and cost-extensions of each project). These documents provide the value of the project as set out in the proposal that was vetted through the organizational decision-making process, risk and achievement scores for each year of the project's life from annual reviews and a final assessment from project completion reviews (where available), and—sometimes benefit-cost ratios and other ex-ante assessments of expected value, when included in the Business Case. Additionally, DevTracker includes a unique project code, the spending to date of every project (including the evolution of spending over time) and the date of its planned and actual commencement and completion. All of this information and documentation is stored in the online IATI database, facilitating their extraction.

From this raw material, we scrape information using a hierarchy of methods to generate two dataframes: one at the project level, and one at the project-year level. The methods we use are:

1. In the first instance, we use a RegEx (regular expression) command to collect the proposed project value from the Business Case.

2. If, for whatever reason, the Business Case is present, but the RegEx fails to pick up the proposed value (for example, if it is listed outside of the usual cover sheet table), we manually extract the proposed value.

3. If the business case is missing, we use a RegEx applied to the most recent Annual Review uploaded to extract the original project value. If both are available, the business case value is always preferred as it is this that determines whether the proposal is subject to assurance and peer review.

4. We use a RegEx to extract the risk scores and project performance scores for each year of the project from its most recent Annual Review, which includes a table of previous scores.

Table 1: Data and completeness of data scraped from DevTracker

	obs	*% of relevant total*	*Min*	*Max*	*Mean*
Project-years	8541				
Unique projects	5034				
Budgeted programme size, £m	1943	38.6	0.01	6,035.86	44.1
Actual spend to date, £m	4942	98.2	-0.82	4,899.65	22.0
Actual spend at project completion, £m	4942	98.2	-0.82	4,899.65	22.9
Variance from planned spend, £m	1909	98.3	-5,430.84	3,333.49	-4.2
Planned start date	5010	99.5	1/4/1987	1/4/2024	23/2/2011
Planned completion date	5006	99.4	18/8/2009	31/3/2045	28/5/2015
Actual completion date	1400	27.8	2/11/2015	26/11/2021	26/7/2018
Variance from planned completion date (days)	1400	27.8	-8382	2450	79.0
Benefit-cost ratio	409	8.1	0.48	400.06	7.6
Annual review score (1-C to 5-A++) scale)	4946	57.9	1 (C)	5 (A++)	3.0
Risk score (1-Low to 4-Severe)	4946	57.9	1 (Low)	4 (Severe)	2.2

5. We use a RegEx, supplemented by a manual search to identify benefit-cost ratios that are reported in Business Cases.

6. We use a RegEx to extract data on project spend to date for all projects, as well as the planned and actual start and completion dates, as well as a unique project identifier.

This yields 8541 project-years over 5034 projects. Almost all missing data occurs when projects are very small (less than £100,000) and last for less than one year, as business cases are sometimes not uploaded for these, and no annual review exists if the project is completed within a year. Such programmes are within the delegated limits for civil servant clearance and tend to be very small procurements or payments to contractors. In a very small number of cases, business cases and project details are withheld for security reasons, but these are rare. Some missing data also arises when incorrect documents are uploaded, or where information has simply not been entered into the system, though these, again, are rare.

Table 1 summarises the data available.

The vast majority of missing actual completion dates represent projects that are ongoing: only 0.6% of observations are genuinely missing. Annual review and risk scores are recoded to numeric values. Annual reviews are coded from 1 (lowest performance) to 5 (highest performance) and risk from 1 (lowest risk) to 4 (highest risk). Annual review scores are collected to assess the extent to which projects achieved their stated outputs. Care should be taken in interpreting these scores, however, as these are ordinal scales, and there is no clear sense that the jump between each score is the same, or that an annual review score of 4 is twice as good as one of 2. Variance from planned spend and variance from planned completion date are used to assess how closely projects adhered to DFID's own guidance on achieving value-for-money, quoted above. Annual review scores and fidelity to spending and completion plans were cited by DFID internally as markers of how well the organization kept to its value of money pledges, and hence the political case for the aid programme.

We winsorize the following variables before analysis: variance from planned spending, variance from planned completion date, and benefit-cost ratio. Winsorizing these variables recodes extreme values (those from the 0th to the 5th percentile and those above the 95th percentile) to the 5th and 95th percentile values respectively. This reduces the influence of

extreme values on the results. We do not winsorize programme values or annual review or risk scores.

5 Method and Empirical Strategy

The two hypotheses we will test are that implementation of the review and assurance system resulted in unintended consequences, specifically manipulation of project size to avoid the review process by agents in the organization (with a null hypothesis of no manipulation); and the second is that review is associated with better project outcomes (with a null hypothesis of no difference in project outcomes).

To test these hypotheses, we would ideally observe all projects proposed in a world without peer review and their proposed value; the projects that were eventually selected in this counterfactual world, and the metrics of project quality collected; and to compare these projects to the ones proposed under the peer review system. Such a comparison would enable us to observe whether the very threat of peer review changes the projects proposed; whether peer review changes which projects are selected, and whether their performance is different to the universe of selected projects in our counterfactual world. Sadly, our counterfactual world is unobservable. Instead, we can observe only the projects eventually selected, before and after the institution of peer review. While this means we cannot ask how peer review changed what was proposed compared to our unobserved counterfactual, we can nevertheless observe the actual distribution of selected projects by their value and if this distribution suggests a distortion in response to peer review, and whether performance metrics are suggestive of an effect on project quality.

We will initially test the first hypothesis visually, using a histogram of project sizes, with £1m bins, comparing the distribution of projects planned to commence before and after 2011, the year in which peer review and assurance was instituted. Manipulation would be indicated by a 'notch' just above £40 million, and 'bunching' just below, suggesting projects were taken from above the discontinuity and reallocated below.

We will confirm visual evidence using two tests of manipulation around a discontinuity. We will first use the test proposed by McCrary (2008), which tests for a discontinuity in the density function of the running variable at the cut-off level for eligibility for treatment (in this case, the running variable is project size, the treatment is peer review and the cut off is £40 million). The test uses a Wald test of the null hypothesis of no discontinuity in the density function of the running variable at the cut off, based on a finely-binned histogram of the running variable and two local linear regressions, on either side of the cut-off. The test requires that manipulation is monotonic, in the sense that manipulation is expected in one direction only. In our case, this is reasonable: we would expect teams to manipulate project size to avoid review, rather than to select into it. We implement this test using the DCdensity package in the statistical programme R.

The second test of manipulation we will implement is the Cattaneo, Jansson and Ma (Cattaneo et al. (2018), henceforth CJM) test for manipulation, which operates on a similar basis to the McCrary test, but based on local polynomial techniques and requiring no pre-binning or other transformation of the data; and implements a test of the null hypothesis of no manipulation using robust bias correction. We implement the CJM using the rddensity package in R. The choice of bandwidth (that is, the observations near to the cut off used for estimation) is estimated using the

available data, and not specified by the researchers.

In each case, if the test statistics returned are sufficiently large, they will reject the null hypothesis of no manipulation. We test the second hypothesis, that review and assurance is associated with better programme performance using three indicators of project quality: annual review scores, fidelity to the planned project completion date and fidelity to the original planned budget. Annual Review scores measure the extent to which projects have met their expected outputs (and as such are somewhat open to manipulation, if project designers set easy targets to guarantee good review scores). Planned completion dates and project spending relative to budget are more objective measures of a particular type of implementation quality. We will compare these three variables above and below the cut off after the implementation of peer review, using t-tests of equality of sample means for each variable; and then compare all projects above and below the review threshold before the implementation of the review system to establish whether the pattern of performance was substantially different before the establishment of the review system (in this second analysis we use all projects due to the smaller pre-2011 sample size). If review is associated with better project outcomes, we would expect to see a relative improvement in annual review scores, time over-runs and variance from the planned budget in projects above the review threshold compared to those below the threshold in the post-2011 sample compared to the pre-2011 sample.

To supplement these results, we use a regression discontinuity design to test for a discontinuity in each of these outcome variables around the £40 million cut off for review, controlling for sector, recipient country and start year of the project, following Briggs (2020) and Honig (2018). We estimate the following equation:

$$Y_i = \beta_0 + \beta_1 QAU_i + \beta_2 ProgVal_i + \beta_3 Sector_i + \beta_4 Recipient_i + \beta_5 StartYear_i + \epsilon_i \quad (1)$$

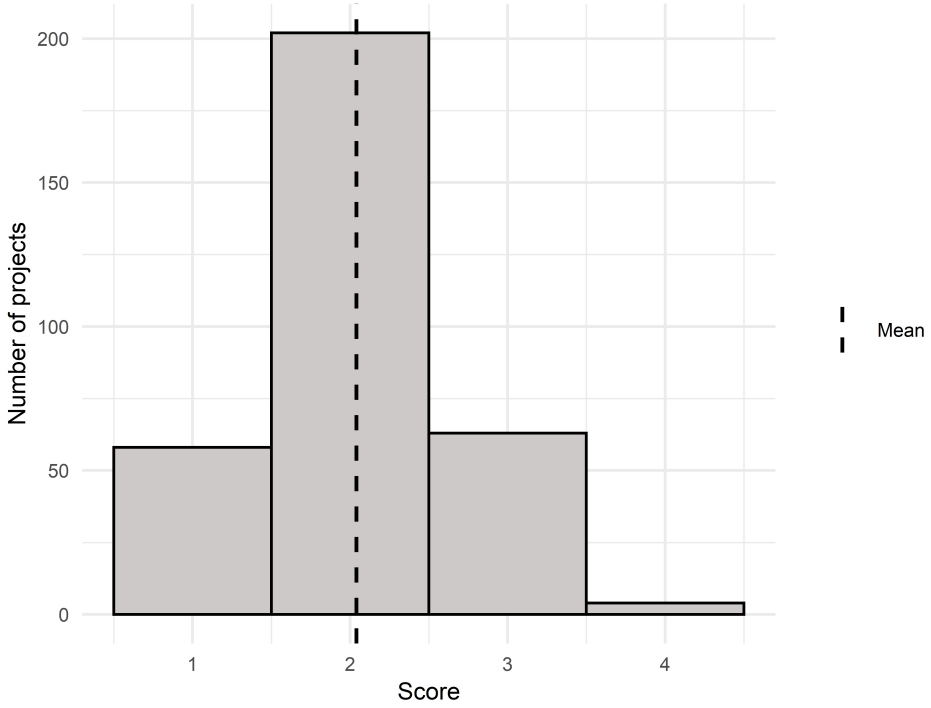
Where QAU is a dummy variable taking the value 0 for a project above the quality assurance threshold, ProgVal is the running variable (programme value of the project), Sector, Recipient and StartYear refer to dummy variables controlling for the location, primary objective and year of actual commencement of each project and ϵ is an error term. The Y variables we investigate are annual review scores, variance from the planned project completion date and variance from the planned project budget. These variables are selected because they correspond to DFID's own guidance to staff on the assessment of value for money in DFID's operations, as outlined in the Setting and Data sections. As such, imperfect though they may be as objective measures of project outcome, they map directly on to the guidance DFID staff received about meeting organizational objectives, and their monitoring would reasonably have been expected by staff. In other words, to the extent that DFID staff felt that programme performance would be monitored, they would likely have expected it to be monitored according to these dimensions. We implement this regression using the rdrobust package in R, and report the conventional, bias-corrected and robust coefficients and standard errors, with the preferred coefficient estimate the conventional and the preferred standard errors the robust, as suggested by Calonico, Cattaneo, and Farrell (2021). We run these regressions for both the pre- and post-QAU samples, to investigate if any differences between projects above and below the future threshold that was observed before the creation of QAU was reversed by the establishment of peer review. This RDD estimation is supplemented in

Appendix C by a classic 2x2 difference-in-difference analysis, with the data structured as repeated cross-sections, and a difference-in-difference analysis adjusting for the same control variables used in the RDD. The estimating equations, results and a discussion are included in the Appendix. The RDD is our preferred specification since it restricts the comparison of outcomes to projects just above and just below the threshold for review (over which range the impact of avoidance is most likely to be concentrated), and because of complications in interpreting results for a difference-in-difference analysis with controls, discussed further in Appendix C

Three points are important to note at the outset. If manipulation around the review cut off is observed, we cannot infer causality using the RDD estimates (the same concern applies to the difference-in-difference analyses). If some projects are being manipulated to come in under the cut off, and we find that project quality metrics are higher for those projects just to the right of the cut off, this could reflect either negative selection or the causal effect of peer review, with no way to disentangle the two effects. Nevertheless it is informative to know if either effect is observed. Secondly, the data available allow us to test for quality on only a few possible dimensions. It may be that project quality is affected in ways that we do not have the data to test. And thirdly, given the small cost of review set out earlier, relatively small effects on project quality would be organizationally meaningful—improvements of even a fraction of a percentage point over many projects would easily pay for a cost of £8000 per review. We are unlikely to be powered to detect such small, but meaningful effects. As such positive evidence of higher project quality above the cut off (assuming no manipulation) will be highly suggestive of a positive effect of peer review; however, the absence of a clear effect, given sample sizes and the multi-dimensionality of quality does not equally imply a positive finding of no effect. Fourthly, an inherent limitation of the data used in this analysis is that we only observe projects that were approved and implemented. If the establishment and functioning of the Quality Assurance system caused a change in the likelihood of being approved among those projects reviewed, then the sample of projects larger than £40 million contains a survivorship bias, a difference in the likelihood of being approved compared to the pre-policy period.

The effect of survivorship is worth considering further. If the quality assurance process is able to identify and discourage the progression of projects likely to perform poorly (despite the already-noted fact that QAU itself does not reject or approve projects), then we would expect projects above the £40m threshold to be of higher quality. To get an idea of how significant the problem of survivorship might be, we requested and were granted data summarising the scores granted to projects reviewed by QAU in the period covered by this study. Recall that QAU awards four scores, from 'Limited Recommendations' to 'Resubmission to Assess Major Recommendations'. Of the four scores, only score 4, which indicates severe problems with a project suggests that a project might struggle to obtain approval even after reworking to address concerns. Figure 1, below, provides a histogram of QAU scores over the period.

Figure 1: Distribution of Quality Assurance Scores, 2011-2020



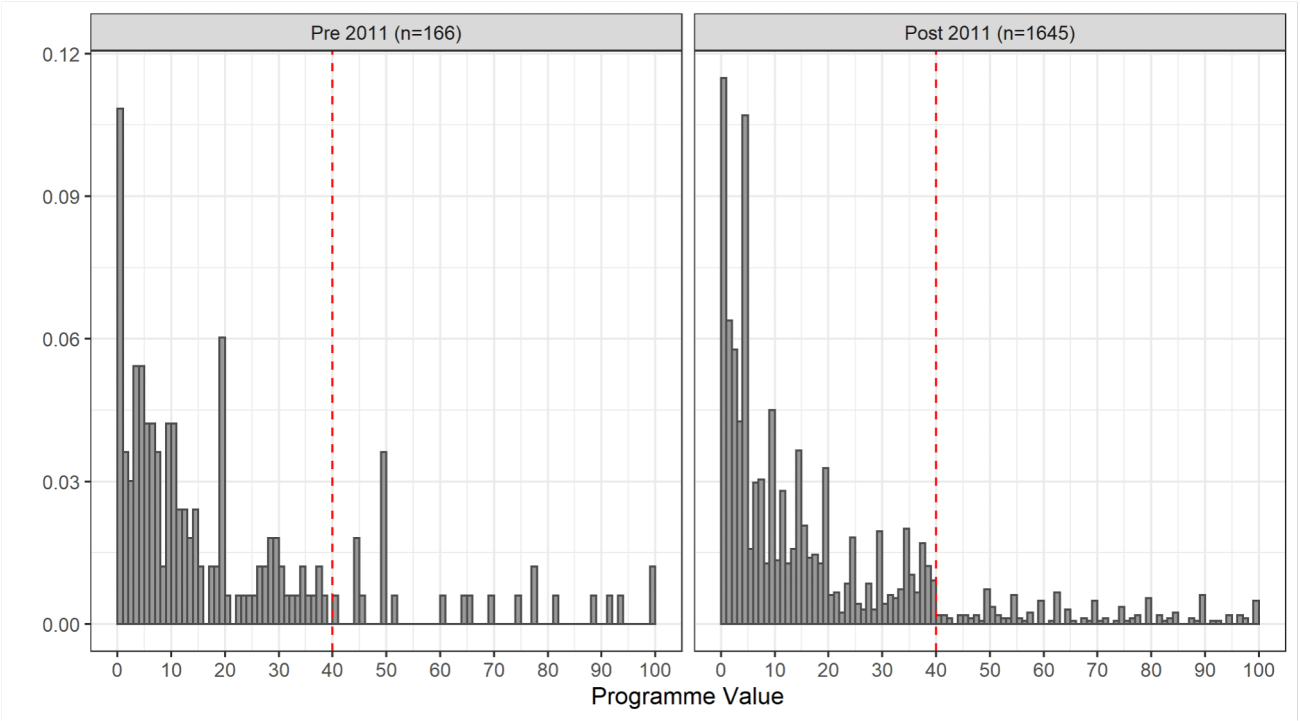
Source: Data provided by DFID in response to author enquiries.

Only a very small minority of projects received the lowest score over the period studied. This suggests that survivorship is likely to play a relatively small role in the data observed. We may additionally be concerned that some projects scoring '3' would also be rejected, even after revision. Reassuringly, the number of '3' scores are still small relative to the number of 1s and 2s; taken together with the observed distribution of projects after the establishment of QAU (see Figure 2), which shows a pattern which cannot be explained by survivorship, this suggests that survivorship is not driving our results (there is a fuller discussion of this on page 19). Nevertheless, to the extent that any effect is observed at all, if QAU function well, and identify the worst proposals, we should expect projects that have been through the process to be, on average, slightly better performing than those which did not need to go through the process. However, if QAU added no value to the project design and approval process (that is, scores are distributed randomly with respect to true project quality), then survivorship should create no systematic bias in project performance metrics in the period after the institution of the quality assurance system. We discuss this further below.

6 Results

A visual inspection of the distribution of proposal values before and after 2011 shows signs of manipulation of proposal sizes around the £40 million threshold for peer review and assurance. Figure 2 shows this clearly.

Figure 2: Distribution of project sizes before and after the institution of Quality Assurance



Note: Bin width is £1 million

The histogram on the right, showing the distribution of project sizes after the establishment of peer review and assurance shows a clear notch above the £40 million threshold, coupled with bunching just below the threshold, a telltale sign of agents trimming projects that would otherwise fall just above the threshold to avoid peer review. Such bunching is absent in the pre-review and assurance sample, though number of observations is smaller here.

The visual inspection is confirmed by both the McCrary (2008) and CJM (2018) tests of manipulation around a discontinuity. Table 2 summarises the results of McCrary tests for manipulation around the discontinuity for both post- and pre-peer review and assurance samples, while Figure 3 presents the plot visualizing the results of the McCrary test. Table 3 and Figure 4 do the same for the CJM tests. Recall that both tests evaluate the null hypothesis of no manipulation, using local linear (McCrary) and polynomial (CJM) techniques. A large test statistic (and small p-value) suggests rejection of the null hypothesis that the density function of the running variable (project size) is smooth around the treatment threshold at £40 million.

Table 2: Results of McCrary Density Test

McCrary (2008) Test Results

	theta	*se*	*z*	*bin size*	*bandwidth*	*p-value*
Before peer review instituted	-0.69	0.82	-0.84	3.24	19.32	0.40
After peer review instituted	-1.65	0.27	-6.03	0.98	22.88	0.00***

Note: *** p<0.01, ** p<0.05, * p<0.1.

Figure 3: Plot of McCrary Density test outputs

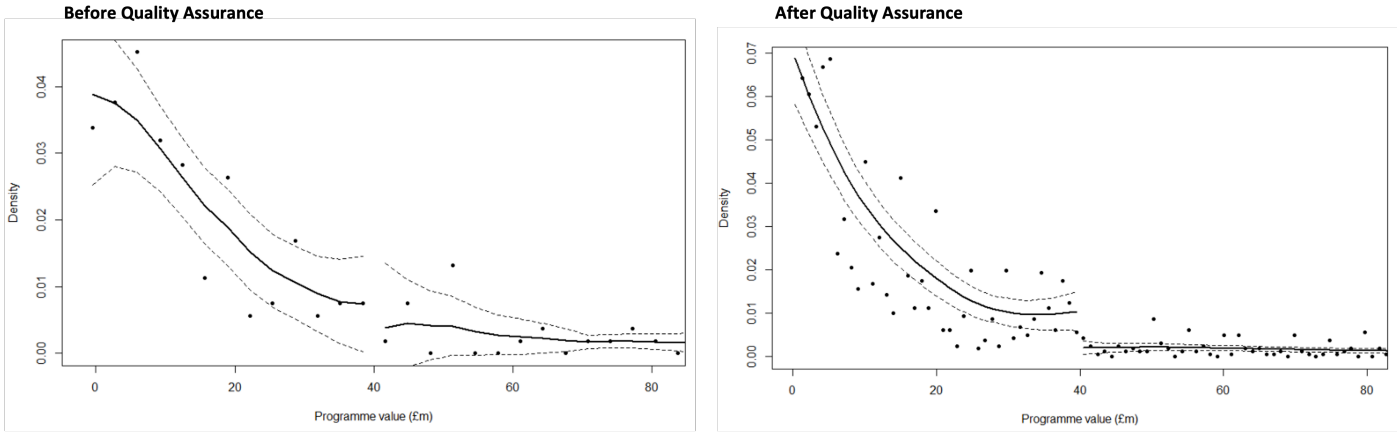
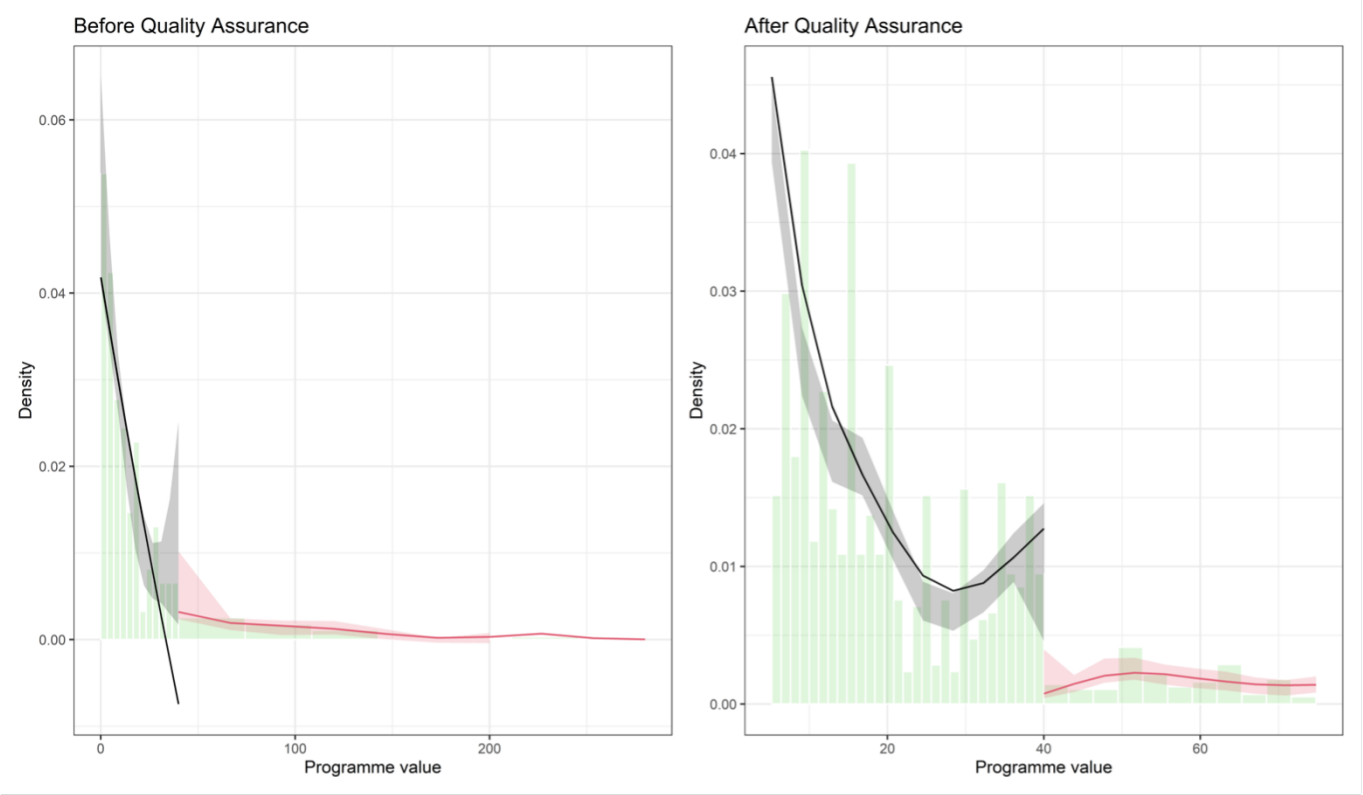


Table 3: Cattaneo, Jansson & Ma (2018) test results

	Before Quality Assurance		**After Quality Assurance**	
Number of observations	184		1759	
Model	unrestricted		unrestricted	
Kernel	triangular		triangular	
BW method	estimated		estimated	
VCE method	jacknife		jacknife	
Cutoff	40		40	
	Left of cutoff	*Right of cutoff*	*Left of cutoff*	*Right of cutoff*
Number of observations	141	43	1474	285
Effective number of observations	141	32	192	46
Order est (p)	2	2	2	2
Order bias (q)	3	3	3	3
BW est. (h)	80	80	11.6	11.6
Method	Robust		Robust	
T	-1.1002		-2.3477	
p-value	0.2712		0.0189**	

Note: *** p<0.01, ** p<0.05, * p<0.1.

Figure 4: Plot of CJM manipulation test results



Both tests strongly reject the null hypothesis of no manipulation around the discontinuity at £40 million for the post-2011 sample, but fail to reject this hypothesis for the pre-2011 sample, strongly suggesting that the introduction of peer review and assurance with a clear decision-rule for eligibility at £40 million pounds induced a substantial response via the 'indirect channel', with agents reorganizing their proposed activities to avoid peer review, with a clear effect on the overall structure of the organisations portfolio.

As a robustness test, we ran both the McCrary and CJM test to investigate the presence of similar discontinuities at other visually-striking thresholds. This pattern of results does not replicate at £10 million and £20 million for both the pre- and post-2011 samples. The CJM test for a discontinuity at £10 million in the pre-QAU sample is significant at the 10 percent level, as is the McCrary test for a discontinuity at £20 million in the pre-QAU sample. All other estimates are insignificant, including all in the post-QAU data. Full results are presented in Appendix B.

These results are consistent with widespread manipulation of project sizes beginning once the new system of peer review and assurance of proposals is introduced. It is not consistent with peer review and assurance picking up suboptimal projects around the threshold and dramatically increasing their chances of rejection-survivorship, in other words. Such a phenomenon cannot explain the clearly observed "bunching" of projects just below the threshold (instead we, would expect a smooth decline in density with respect to project size up to the discontinuity, followed by a sudden reduction), nor can it explain the 'recovery' of the distribution observed around £50 million pounds, since there is no reason to expect that poor quality would be observed and flagged by quality assurance only in those projects just above the threshold. Rather, this recovery suggests that the margin over which project size can be manipulated is somewhat narrow, or that the cost of such manipulation (smaller projects, perhaps leaving economies of scale unexploited or

team budgets unused) is only worth the reward (avoiding scrutiny) up to a point. This is consistent with the data provided in Figure 1, which suggests that, judging by the distribution QAU scores, the effect of outright rejection due to poor quality assurance scores is small.

Turning to our second hypothesis, we investigate project quality either side of this threshold. The observed manipulation suggests we should expect lower quality below the threshold because of negative selection, project rejection or because quality assurance and peer review improve projects. We first investigate whether project review scores are different just above and just below the manipulation point; and whether objective measures of performance (degree of over- or under-spending, or over- and-under running compared to expected project lifetime) vary around this point, after the implementation of peer review. The objective measures are preferred: annual review scores simply reflect the extent to which a project achieves expected outputs, and are thus endogenous to project quality (since worse projects can simply report more conservative expected outputs. Equally there is an incentive for all projects, good or bad, to 'lowball' their expected outputs and so make achieving high scores in annual reviews easier) Table 4 summarises the results.

Given the evidence of manipulation of project sizes to avoid peer review found, we might expect to see higher

Table 4: Project performance around the quality assurance threshold, 2011-2020

Project implementation quality indicators, QAU period			
	Projects within £10 million of cutoff		t-test
	Below cutoff (obs)	*Above cutoff (obs)*	*null: equal means*
Annual review score (1-C -5-A++)	2.93	2.99	0.496
	(532)	*(73)*	
Difference between planned and actual completion (days)	95.71	76.91	0.3905
	(98)	*(16)*	
Difference between planned and actual spend (£m)	-4.73	-0.08	0.1833
	(175)	*(27)*	

Note: This table presents the mean (number of observations in parentheses) of Annual Review scores (higher numbers indicate better performance relative to expectation); the difference between the planned and actual project completion date in days; and the difference between planned and actual project spend in millions of pounds for projects whose original budget within £10 million pounds of the review threshold. The last column prevents the p value of a t-test of equality of means, with the null hypothesis that the difference between means is 0. The null fails to be rejected for each variable tested.
** p<0.01, ** p<0.05, * p<0.1.

performance among projects being peer reviewed, even if the effect is purely driven by selection into (or rather, out of) treatment. However, comparing projects just above and just below the peer review threshold in the years since peer review and quality assurance was implemented yields no clear evidence of performance benefits from peer review. Annual review scores (which, as discussed, are also open to manipulation), implementation overruns and overspending (which are more objective measures) are all similar among projects that are just above and just below the review threshold. Extending these tests to all projects above and below the threshold does not change the story. Nor does limiting analysis of annual review scores to the years just after approval, when the effect of quality at inception of the project may be expected to be highest.

It is possible that this finding of no significant difference itself reflects progress: that before quality assurance was implemented smaller projects performed systematically better than bigger projects. Table 5 investigates this possibility, using all projects above and below the (future) review threshold for the pre-2011 sample (due to fewer observations).

It does not clearly support this possibility. There is no significant difference in Annual Review scores or implementation over-runs above and below the threshold before the review system was implemented, though larger projects were more likely to over-run, which was not the case after the establishment of the Quality Assurance system. There is a significant difference in actual spending relative to planned spending, with large projects likely underspend by around £10 million on average, and smaller projects to overspend by around £4 million on average (as a percentage of project size, this translates to around 10 percent and 40 percent of original project size, respectively).

Table 5: Project performance around the future quality assurance threshold, pre-2011

Project implementation quality indicators, pre-QAU period			
	Projects within £10 million of cutoff		t-test
	Below cutoff (obs)	*Above cutoff (obs)*	*null: equal means*
Annual review score (1-C -5-A++)	3.08	3.09	0.9245
	(416)	*(196)*	
Difference between planned and actual completion (days)	91.88	139.03	0.1584
	(73)	*(27)*	
Difference between planned and actual spend (£m)	3.73	-9.80	0.0005***
	(137)	*(43)*	

Note: This table presents the mean (number of observations in parentheses) of Annual Review scores (higher numbers indicate better performance relative to expectation); the difference between the planned and actual project completion date in days; and the difference between planned and actual project spend in millions of pounds. The last column prevents the p value of a t-test of equality of means, with the null hypothesis that the difference between means is 0.
** p<0.01, ** p<0.05, * p<0.1.

However, the regression discontinuity analysis suggests some improvement among projects large enough to qualify for review relative to those below this threshold after the establishment of the Quality Assurance Unit in 2011. These results indicate relative improvements in Annual Report scores and time overruns among projects qualifying for review that may reflect an impact of peer review directly, of project rejection or of negative selection below the threshold. Table 6, below, reports the results of this analysis for the pre and post-QAU periods.

In the pre-QAU period, annual review scores were slightly lower for projects above what would become the cut-off for peer review, but after the implementation of review there is no significant difference between projects just above and just below the threshold. This suggests a relative improvement in annual review scores among projects large enough to be reviewed compared to those falling below the review threshold after the establishment of QAU. And in the years during which peer review was applied, reviewed projects had significantly smaller deviations from their planned completion date, whereas before peer review was implemented there was no significant difference around the future cut-off for review. Each of these changes suggest slightly better performance among those projects eligible for review in the period after review is implemented, though the magnitudes involved are quite small. Comparing point estimates on Annual Review scores suggests an improvement of around half a grade; for project over-runs, around 60 days. As discussed earlier, we cannot distinguish between the causal effect of peer review (that is, the existence of review improves project quality, either through the review process itself or by changing the quality of proposals made in the first instance) and selection effects, which may be driven by negative selection below the threshold (that is, by projects with lower annual review scores and longer overruns being shifted to just below the £40 million cut off) or by project rejection (that is, by projects with worse expected performance being caught in the peer review process and

Table 6: Test for a discontinuity in project performance around the £40 million threshold

	Before Quality Assurance			**After Quality Assurance**		
	Time overruns	Spend vs. budget	AR scores	Time overruns	Spend vs. budget	AR scores
Conventional	-12.7	-9.9	-0.6**	-74.0**	3.4	0.1
	(223.4)	(11.0)	(0.2)	(34.4)	(5.9)	(0.1)
Bias-Corrected	-40.1	-10.1	-0.6**	-86.8**	3.8	0.1
	(223.4)	(11.0)	(0.2)	(34.4)	(5.9)	(0.1)
Robust	-40.1	-10.1	-0.6**	-86.8**	3.8	0.1
	(258.0)	(12.6)	(0.3)	(37.6)	(6.7)	(0.1)
nobs.left	73.0	137.0	406.0	869.0	1448.0	3373.0
nobs.right	27.0	43.0	196.0	118.0	281.0	864.0
nobs.effective.left	18.0	9.0	89.0	110.0	243.0	775.0
nobs.effective.right	9.0	5.0	55.0	25.0	60.0	196.0
cutoff	40.0	40.0	40.0	40.0	40.0	40.0
order.regression	2.0	2.0	2.0	2.0	2.0	2.0
order.bias	2.0	2.0	2.0	2.0	2.0	2.0
kernel	Triangular	Triangular	Triangular	Triangular	Triangular	Triangular
bwselect	mserd	mserd	mserd	mserd	mserd	mserd

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Note: The 'Before Quality Assurance' columns report results for Equation (1) run on data in the pre-quality assurance period, with the dummy variable QAU set to 1 for projects that would have been peer reviewd after the establishment of the Quality Assurance Unit (that is, those equal to or larger than £40 million). Time overruns are in days. Spending relative to budget is in millions of pounds. Annual Review scores run from 1 (C) to 5 (A++). Coefficients give the difference in the outcome variable among observations just above and just below the cut off in the units specified above, Calonico, Cattaneo and Farrell (2020) recommend using conventional coefficients with robust SEs.nobs.left and nobs.right provide the number of observations on each side of the discontinuity investigated. The MSE-Optimal bandwidth was used in each of these regressions.

being rejected at a higher rate).

Furthermore, neither a difference-in-difference regression following the classic 2x2 structure (with data structured as repeated cross sections) nor one including the same control variables included in the RDD analysis suggest that there are any significant differences across the three performance metrics used. These results are presented in Table 11 in Appendix C. The RDDs are preferred since they restrict comparisons to projects just above and just below the threshold, where the impact of avoidance of review is likely to be concentrated, and because of inferential challenges of using a difference-in-difference design with control variables.[13] Given the small magnitudes (and the sensitivity of these results to alternative specifications), these results should be treated as suggestive rather than definitive evidence of a small effect from the institution of the quality assurance regime, though not on all measurable dimensions of project quality, and one which my reflect selection (either through manipulation of project size or through project rejection) as well as or instead of the direct effect of peer review on project quality.

7 Discussion

The foregoing analysis leads us to conclude that the introduction of a review system with a clear decision-rule for eligibility led to substantial manipulation of project sizes to avoid review. However, we find only weak evidence of an associated effect on project quality, either through selection effects or the causal effect of review. There is some evidence that annual review scores, previously slightly lower among projects just above the £40 million mark, were

[13] For a careful and up-to-date discussion of these issues, see Roth et. al. (2022).

no longer significantly different from between reviewed and not reviewed projects after the institution of peer review; and some evidence that time overruns in projects was signifcantly smaller among projects just above the reviewing threshold after the institution of review. These results are not, however, confirmed by a difference-in-difference estimation.

The finding of clear manipulation suggests that organisations considering the implementation of new systems of review and assurance must consider not only the "naïve" effect of the system but also the impact its adoption has on how agent behaviour within the organization, which may be substantial and costly.

The project quality estimates suggest that even the naïve effect should not be taken for granted. The small size and fragility of the effect of review on quality metrics, even in the presence of documented manipulation around the threshold is, on the face of it, a puzzle. We propose four possible explanations here, each of which is consistent with the observed results.

First, it may be that peer review adds little value in this context. This could be the case, for example, if the reviewers suffer from incentive problems, for example, withholding some criticisms of potential future colleagues or managers—though, as discussed, the most senior figures associated with the review team are outside of the usual civil service churn in this case. It may also arise if reviewers suffer from the same cognitive and informational limitations as the reviewed agents and additional scrutiny adds little value to proposals (unlike the effect of 'cognitive redundancy' identified in Hutchins 1995). If this were the case, project rejections would not result in large quality differences above and below the review threshold as QAU would be able to distinguish between better and worse projects only imperfectly and rejection would not strongly enforce any quality differential. A third possibility is that there was no problem to resolve, and proposal quality was already as good as could be achieved, given other systems in place. The lack of clear evidence for a problem among large projects in the pre-2011 sample suggests this may be part of the explanation. It is also possible that peer review improves project proposal quality, but that proposal quality is only weakly related to actual project implementation quality.

A second possibility is that while the manipulation around the cutoff observed is indeed negative selection of worse projects, this is small relative to the effect of peer review and the average quality of proposals in the organization. For a given level of negative selection, the smaller any positive causal effect on those that are not negatively selected, and the higher the average quality of proposal (and the smaller gap between better and worse proposals), the less likely we are to detect any performance advantage among the reviewed proposals. A related possibility is that the existence of review had spillover effects (for example if agents who have one proposal reviewed subsequently improve the quality of all other proposals, including those that are not reviewed) and improve projects both above and below the review threshold.[14]

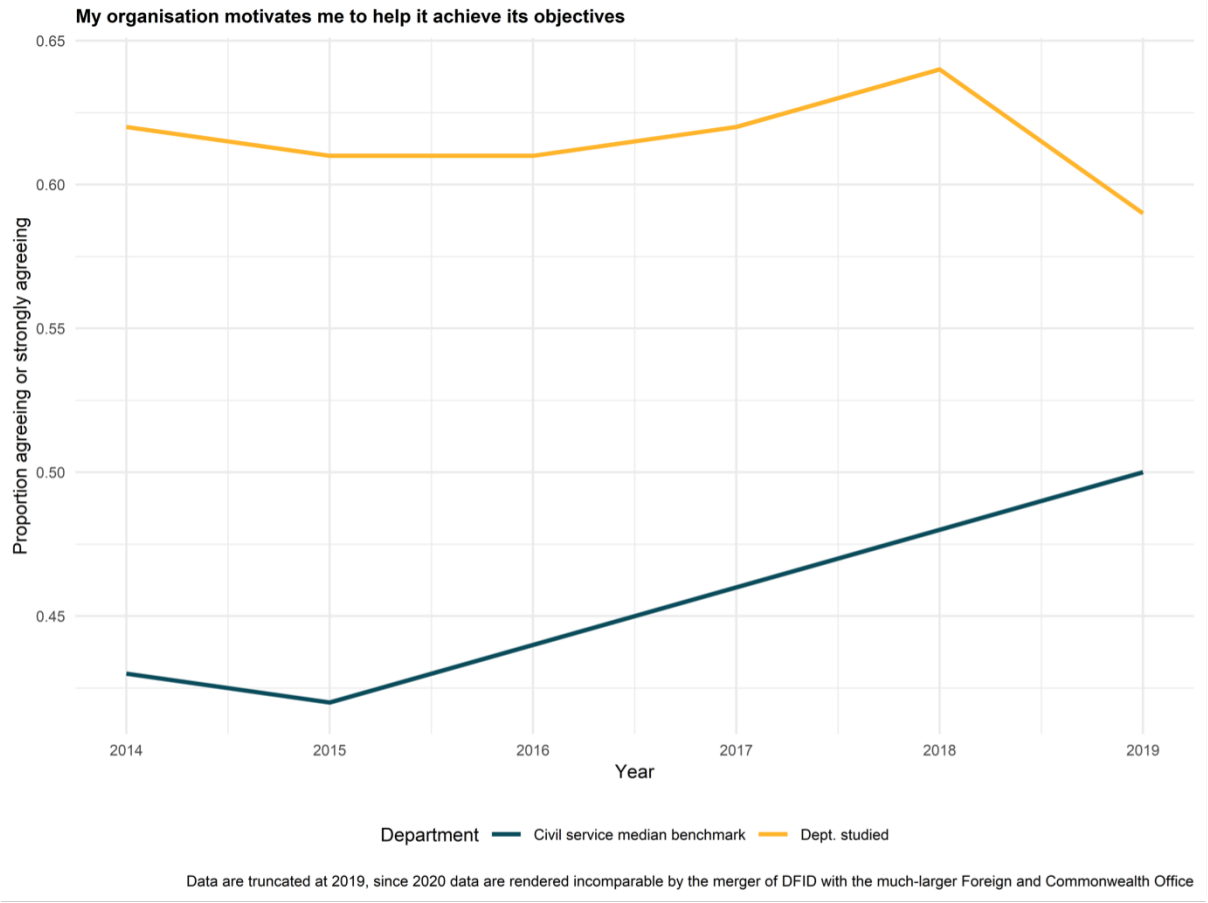
A third possibility is that the metrics tracked are simply unable to adequately assess project quality. Annual Review scores may be endogenous to proposal quality (though it should be noted that they have been used to demonstrate differences between more and less successful programmes, for example in Honig (2018). Project over-runs could reflect

[14]The difference between actual and planned completion dates and actual and planned spending is significantly (in both the statistical and practical senses) smaller for the entire universe of projects during the review era compared to those implemented before the review system was established. However, other factors, including increasing professionalism, other systems of quality control and greater external scrutiny could also explain this.

projects being extended for good performance rather than poor implementation. Under- or over-spending could reflect economy or cost extensions due to high performance respectively. It may be that a much more complex metric of project quality is required to fully assess the effect of review, though, as has been noted already, these metrics were highlighted and in internal guidance as they key ways in which DFID's commitment to value for money would be assessed. If they are insufficiently informative to provide an insight into project performance, despite programme managers being provided with guidance that stressed their importance, this is itself a missed opportunity for transparency and accountability.

A fourth possibility is that the observed manipulation reflects positive, rather than negative selection. This may be the case if the mechanism underlying avoidance behaviour is 'control aversion' (Bowles 2016; Falk and Kosfeld 2006; Ziegelmeyer et al. 2012). Highly intrinsically motivated agents may bridle against constraints on their action designed to change the behaviour of agents with low intrinsic motivation. In such a case, avoidance behaviour may not always reduce performance at the project level, since these intrinsically motivated agents are likely to still exert effort and design programmes well-aligned to the organizational mandate. Instead it would lead to projects being sub-optimally small for their level of effectiveness. We cannot test this directly, but evidence from representative surveys conducted across the Civil Service (the so-called 'People Survey') are consistent with DFID having a high proportion of highly intrinsically motivated agents.

Figure 5: Intrinsic motivation in DFID, 2014-19



Most organisations aspire to a more structured decision-making process than the near-anarchic decision making or the gradual process of muddling through to better decisions proposed by Cohen, March, and Olsen (1972) and Lindblom (1959), respectively. However, the findings of this study suggest that the case for even outwardly plausible mechanisms for improving organizational function need greater scrutiny. Such mechanisms may fail on design grounds or because agents within the organization adjust to their presence. The example examined here illustrates both possible problems.

References

Aczel, B., Szaszi, B., and Holcombe, A. O. (2021). A billion-dollar donation: estimating the cost of researchers' time spent on peer review. *Research integrity and peer review*, 6(1):1–14.

Arnaud, S. and Chandon, J.-L. (2013). Will monitoring systems kill intrinsic motivation? An empirical study. *Revue de Gestion des Ressources Humaines*, N° 90(90):35–53.

Aucoin, P. and Heintzman, R. (2000). The Dialectics of Accountability for Performance in Public Management Reform. *International review of administrative sciences*, 66(1):45–55.

Banerjee, A., Andrabi, T., Grantham-McGregor, S., Yoshikawa, H., Saavedra, J., Banerji, R., Akyeampong, K., Dynarski, S., Glennerster, R., Muralidharan, K., Schmelkes, S., and Piper, B. (2020). Cost-effective approaches to improve global learning: What does recent evidence tell us are "Smart Buys" for improving learning in low- and middle-income countries? Technical report.

Banuri, S., Dercon, S., and Gauri, V. (2017). Biased Policy Professionals. *World Bank Policy Research Working Paper*, 8113(June).

Bertelli, A. M. (2006). Motivation Crowding and the Federal Civil Servant: Evidence from the U.S. Internal Revenue Service. *International public management journal*, 9(1):3–23.

Bertelli, A. M. (2007). Determinants of Bureaucratic Turnover Intention: Evidence from the Department of the Treasury. *Journal of public administration research and theory*, 17(2):235–258.

Besley, T. (2007). *Principled agents?: The Political Economy of Good Government*. Oxford University Press.

Bloom, N., Lemos, R., Sadun, R., and Van Reenen, J. (2015). Does Management Matter in schools? *Economic Journal*, 125(584):647–674.

Bloom, N., Sadun, R., and Van Reenen, J. (2016). Management as a Technology?

Bovens, M. (2007). Analysing and Assessing Accountability: A Conceptual Framework. *European law journal : review of European law in context*, 13(4):447–468.

Bowles, S. (2016). *The moral economy : why good incentives are no substitute for good citizens*. Yale University Press, New Haven, CT, US.

Brehm, J. and Gates, S. (1997). *Working, Shirking, and Sabotage: Bureaucratic Response to a Democratic Public*. University of Michigan Press, Michigan.

Briggs, R. C. (2017). Does Foreign Aid Target the Poorest? *International Organization*, 71(1):187–206.

Briggs, R. C. (2020). Results from single-donor analyses of project aid success seem to generalize pretty well across donors. *Review of International Organizations*, 15(4):947–963.

Bright, L. K. (2021). Why Do Scientists Lie? *Royal Institute of Philosophy Supplement*, 89(May):117–129.

Buchanan, J. and Tullock, G. (1962). *The Calculus of Consent*. University of Michigan Press, Ann Arbor, MI.

Calonico, S., Cattaneo, M. D., and Farrell, M. H. (2021). Optimal bandwidth choice for robust bias-corrected inference in regression discontinuity designs. *The Econometrics Journal*, 23(2):192–210.

Carpenter, D. P. and Krause, G. A. (2012). Reputation and Public Administration. *Public Administration Review*, 72(1):26–32.

Cattaneo, M. D., Arbor, A., Jansson, M., and Ma, X. (2018). Manipulation testing based on density discontinuity. *The Stata Journal*, 18(1):234–261.

Chetty, R., Saez, E., and Sándor, L. (2014). What Policies Increase Prosocial Behavior? An Experiment with Referees at the Journal of Public Economics †. *Journal of Economic Perspectives*, 28(3):169–188.

Cohen, M. D., March, J. G., and Olsen, J. P. (1972). A Garbage Can Model of Organizational Choice. *Administrative Science Quarterly*, 17(1):1–25.

Coviello, D., Spagnolo, G., and Lotti, C. (2021). Rules, bunching and discretion in emergency procurement: Evidence from an earthquake. In Bandiera, O., Bosio, E., and Spagnolo, G., editors, *Procurement in Focus: Rules, Discretion, and Emergencies*, pages 13–22. CEPR Press, London.

Crijns, T. J., Ottenhoff, J. S. E., and Ring, D. (2021). The effect of peer review on the improvement of rejected manuscripts. *Accountability in research*, ahead-of-p(ahead-of-print):1.

Deci, E. L. and Cascio, W. F. (1972). Changes in Intrinsic Motivation as a Function of Negatrive Feedback and Threats.

Della Vigna, S. and Linos, E. (2020). *RCTs to Scale: Comprehensive Evidence from Two Nudge Units*. Working paper series (National Bureau of Economic Research : Online) ; working paper no.27594. National Bureau of Economic Research, Cambridge, Mass.

DFID (2011). DFID's Approach to Value for Money (VfM). Technical Report July.

DFID (2020). Smart Rules. Technical Report April.

Dissanayake, R. (2021). Hitting 0.7 For 0.7's Sake: The Perils of the Global Aid Funding Target — Center For Global Development.

Dissanayake, R. (2022). The Importance of Being Earnest: Noise, Incentives and Hierarchy in Public Sector Decision-Making.

Dunsch, F., Evans, D., Eze-Ajoku, E., and Macis, M. (2021). Management, Supervision, and Health Care: A Field Experiment. *NBER Working Paper Series*, page 23749.

Ellison, G. (2002). The Slowdown of the Economics Publishing Process. *Journal of Political Economy*, 110(5):947–993.

Elston, T. and Zhang, Y. (2022). Implementing Public Accounts Committee Recommendations: Evidence from the UK Government's 'Progress Reports' since 2012. *Parliamentary Affairs*.

Falk, A. and Kosfeld, M. (2006). The Hidden Costs of Control. *American Economic Review*, 96(5):1611–1630.

Flesher, D. L. and Zarzeski, M. T. (2002). The roots of operational (value-for-money) auditing in English-speaking nations. *Accounting and business research*, 32(2):93–104.

Frey, B. S. (1993). Does monitoring increase work effort? The rivalry with trust and loyalty. *Economic inquiry*, 31(4):663–670.

Gavas, M. and Calleja, R. (2020). DfID is a world leader in tackling poverty. Our international standing is weakened without it — Aid — The Guardian.

Gawande, A. (2010). *The checklist manifesto : how to get things right*. Profile Books, London, England.

Gibbons, R. (2016). Incentives in Organizations. *Journal of Economic Perspectives*, 12(4):115–132.

Hares, S. and Rose, P. (2021). As it Assumes Leadership of the Global Education Agenda, the UK Slashes Its Own Aid to Education — Center For Global Development.

Higgs, M. and Gelman, A. (2021). Research on registered report research. *Nature Human Behaviour*.

Hoggett, P. (1996). New modes of control in the public sector. *Public administration (London)*, 74(1):9–32.

Honig, D. (2018). *Navigation by judgment : why and when top-down management of foreign aid doesn't work*. Oxford University Press, New York.

Hutchins, E. (1995). *Cognition in the wild*. MIT Press.

ICAI (2018). DFID's approach to value for money in programme and portfolio management A performance review. Technical Report February.

ICAI (2019). ICAI follow-up of: DFID's approach to value for money in programme and portfolio management A summary of ICAI's full follow-up. Technical Report July 2019.

Ioannidis, J. P. (2019). Why Most Published Research Findings Are False. *Chance (New York)*, 32(1):4–13.

Kahneman, D., Lovallo, D. P., and Sibony, O. (2019). A structured approach to strategic decisions. *MIT Sloan Management Review*, 60(1):1–12.

Kahneman, D., Rosenfield, A. M., Gandhi, L., and Blaser, T. (2016). Noise: How to overcome the high, hidden cost of inconsistent decision making. *Harvard Business Review*, 2016(October).

Kaufman, D. (2021). It's complicated: Lessons from 25 years of measuring governance. *Brookings Future Development*.

Kells, S. (2011). The Seven Deadly Sins of Performance Auditing: Implications for Monitoring Public Audit Institutions. *Australian Accounting Review*, 21(4):383–396.

Kennedy McDade, K. and Mao, W. (2021). UK aid cuts will put global health systems at risk - The BMJ.

Kerr, S. (1995). On the folly of rewarding A, while hoping for B. *Academy of Management perspectives*, 9(1):7–14.

Lankester, T. (2013). *The politics and economics of Britain's foreign aid : the Pergau Dam affair*. Routledge.

Lindblom, C. E. (1959). The Science of " Muddling Through ". *Public Administration Review*, 19(2):79–88.

Lonsdale, J. (2000). Developments in value-for-money audit methods: impacts and implications. *International Review Of Administrative Sciences*, 66(1):73–89.

Martinez, E. A., Beaulieu, N., Gibbons, R., Pronovost, P., and Wang, T. (2015). Organizational Culture And Performance. *American Economic Review: Papers & Proceedings*, 3(4):512–527.

McCrary, J. (2008). Manipulation of the running variable in the regression discontinuity design: A density test. *Journal of econometrics*, 142(2):698–714.

Mitchell, I. and Baker, A. (2019). How Effective Is UK Aid? Assessing the Last 8 Years of Spending — Center For Global Development.

Mitchell, I., Hughes, S., and Ritchie, E. (2021). An Overview of the Impact of Proposed Cuts to UK Aid — Center For Global Development.

Molander, P. (2014). Public procurement in the European Union: The case for national threshold values. *Journal of public procurement*, 14(2):181–214.

Morin, D. (2001). Influence of Value for Money Audit on Public Administrations: Looking Beyond Appearances. *Financial Accountability & Management*, 17(2):99–117.

Palguta, J. and Pertold, F. (2017). Manipulation of Procurement Contracts: Evidence from the Introduction of Discretionary Thresholds. *American economic journal. Economic policy*, 9(2):293–315.

Power, M. (1999). *The audit society: rituals of verification.* Oxford scholarship online. Oxford University Press, Oxford.

Rainey, H. G. (1993). A theory of goal ambiguity in public organizations. In *Research in public administration.*, volume 2, pages 121–166.

Roth, J., Sant'Anna, P. H., Bilinski, A., and Poe, J. (2023). What's trending in difference-in-differences? A synthesis of the recent econometrics literature. *Journal of Econometrics*.

Sasse, T. and Norris, E. (2019). Moving On The costs of high staff turnover in the civil service. Technical report, Institute for Government.

Sibony, O., Lovallo, D., and Powell, T. C. (2017). Behavioral Strategy and the Strategic Decision Architecture of the Firm. *California Management Review*, 59(3):5–21.

Simon, H. A. (1997). *Administrative Behavior: A study of decision- making processes in administrative organizations.*

Soderberg, C. K., Errington, T. M., Schiavone, S. R., Bottesini, J., Thorn, F., Vazire, S., Esterling, K., and Nosek, B. (2021). Initial evidence of research quality of registered reports compared with the standard publishing model. *Nature Human Behaviour*.

Tas, B. K. O. (2019). Bunching Below Thresholds to Manipulate Public Procurement. RSCAS Working Papers 2019/17, European University Institute.

UK Parliament (2020). Effectiveness of UK aid: potential impact of FCO/DFID merger - International Development Committee - House of Commons. Technical report, UK Parliament.

Watts, R. (2021). The latest view of UK aid: death by a thousand cuts — Save the Children UK.

Williamson, O. E. (1999). Public and Private Bureaucracies : A Transaction Cost Economics Perspective. *Journal of Law, Economics, & Organization*, 15(1):306–342.

Williamson, O. E. (2002). The theory of the firm as governance structure: from choice to contract. *Journal of economic perspectives*, 16(3):171–195.

Ziegelmeyer, A., Schmelz, K., and Ploner, M. (2012). Hidden costs of control: four repetitions and an extension. *Experimental Economics*, 15(2):323–340.

Appendix A Statement from co-author

To whom it may concern:

I confirm that Ranil Dissanayake was the primary author of our article, A Higher Bar or Obstacle Course: Peer Review and Organizational Decision Making in an International Development Bureaucracy. Ranil generated the idea behind the article, and identified our variables of interest, as well as being primary author of the article itself.

Euan Ritchie

Senior Policy Adviser

Development Initiatives

Euan.Ritchie@devinit.org

Appendix B Robustness tests for manipulation

Table 7: Results of McCrary Density Test at £10 million

McCrary (2008) Test Results - £10 million						
	theta	*se*	*z*	*bin size*	*bandwidth*	*p-value*
Before peer review instituted	0.18	0.36	0.5	1.00	13.67	0.61
After peer review instituted	0.01	0.11	0.12	0.98	15.55	0.90

Note: *** p<0.01, ** p<0.05, * p<0.1.

Table 8: Results of McCrary Density Test at £20 million

McCrary (2008) Test Results - £20 million						
	theta	*se*	*z*	*bin size*	*bandwidth*	*p-value*
Before peer review instituted	1.03	0.61	1.69	3.24	17.87	0.09*
After peer review instituted	0.28	0.17	1.62	0.98	20.25	0.11

Note: *** p<0.01, ** p<0.05, * p<0.1.

Table 9: Cattaneo, Jansson & Ma (2018) test results at £10 million

	Before Quality Assurance		**After Quality Assurance**	
Number of observations	184		1759	
Model	unrestricted		unrestricted	
Kernel	triangular		triangular	
BW method	estimated		estimated	
VCE method	jacknife		jacknife	
Cutoff	10		10	
	Left of cutoff	*Right of cutoff*	*Left of cutoff*	*Right of cutoff*
Number of observations	72	112	807	952
Effective number of observations	54	31	327	238
Order est (p)	2	2	2	2
Order bias (q)	3	3	3	3
BW est. (h)	8.84	8.84	5.634	5.637
Method	Robust		Robust	
T	1.6918		-0.3893	
p-value	0.0907*		0.697	

Note: *** p<0.01, ** p<0.05, * p<0.1.

Table 10: Cattaneo, Jansson & Ma (2018) test results at £20 million

	Before Quality Assurance		**After Quality Assurance**	
Number of observations	184		1759	
Model	unrestricted		unrestricted	
Kernel	triangular		triangular	
BW method	estimated		estimated	
VCE method	jacknife		jacknife	
Cutoff	20		20	
	Left of cutoff	*Right of cutoff*	*Left of cutoff*	*Right of cutoff*
Number of observations	106	78	1151	608
Effective number of observations	78	32	470	199
Order est (p)	2	2	2	2
Order bias (q)	3	3	3	3
BW est. (h)	17.1	17.1	13.5	13.5
Method	Robust		Robust	
T	0.0804		-1.2899	
p-value	0.936		0.1971	

Note: *** p<0.01, ** p<0.05, * p<0.1.

Appendix C Difference-in-Difference Estimation

We supplement our main regression results with a difference-in-difference (DID) estimation. The DID estimates compare the difference between projects smaller and larger than £40 million in the pre- and post-peer review periods, testing the null hypothesis that the performance gap between these groups did not change after the introduction of peer review.

We intially use the classic DID set-up with two groups: treated (that is, large enough to be subject to peer review) and untreated (too small to be subject to peer review) and two time periods (before and after the introduction of peer review). Our estimating equation is:

$$Y_i = \beta_0 + \gamma_1 QAU_i + \gamma_2 post_i + \gamma_3 QAU * post_i + \epsilon_i \tag{2}$$

Where Y is each of our three performance metrics in turn (time overruns, spend vs. budget, and Annual review scores); QAU is a dummy variable taking the value 1 when a project is larger than £40 million, *post* is a dummy variable taking the value 1 for the period after peer review is instituted, and *QAU*post* is an interaction term, taking the value 1 when a project is both larger than £40 million and implemented after the introduction of peer review. The γ_3 coefficient gives the Average Treatment Effect on the Treated. ϵ is a error term. We also implement a version of this regression including the same control variables we used in our preferred RDD specification:

$$Y_i = \beta_0 + \gamma_1 QAU_i + \gamma_2 post_i + \gamma_3 QAU * post_i + \alpha_1 Sector_i + \alpha_2 Recipient_i + \alpha_3 StartYear + \epsilon_i \tag{3}$$

Again, γ_3 is the coefficient of interest. Standard errors are White-corrected for heteroscedasticity. The results are presented in Table 11.

Table 11: Difference-in-Difference Regressions

	Basic			With Controls		
	Time overruns	Spend vs. budget	AR scores	Time overruns	Spend vs. budget	AR scores
(Intercept)	154.3***	3.7***	3.1***	3.1***	527.0***	-6.8
	(15.4)	(0.8)	(0.0)	(0.0)	(124.5)	(14.4)
QAU ($\geq$ £40m)	-15.3	-13.5***	0.0	0.0	1.4	-11.8***
	(34.9)	(3.6)	(0.1)	(0.1)	(28.5)	(2.5)
Post (2011)	-55.7***	-4.2***	-0.1**	-0.1**	-422.8***	-17.1
	(15.8)	(0.9)	(0.0)	(0.0)	(110.3)	(19.4)
QAU x Post	19.1	0.3	0.0	0.0	-9.0	-0.8
	(37.0)	(4.0)	(0.1)	(0.1)	(31.2)	(2.6)
Num.Obs.	1400	1909	4946	4946	1394	1909
R2	0.016	0.116	0.002	0.002	0.214	0.340
R2 Adj.	0.014	0.115	0.001	0.001	0.085	0.256
AIC	17628.5	15490.0	9427.8	9427.8	17621.2	15358.4
BIC	17654.7	15517.8	9460.3	9460.3	18658.7	16563.7
Log.Lik.	-8809.256	-7740.000	-4708.893	-4708.893	-8612.616	-7462.222
F	4.600	32.373	3.538	3.538		
RMSE	130.76	13.95	0.63	0.63	116.68	12.06
Std.Errors	HC3	HC3	HC3	HC3	HC3	HC3

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Note: Time overruns are in days. Spending relative to budget is in millions of pounds. Annual Review scores run from 1 (C) to 5 (A++). QAU x Post is the difference-in-difference estimator. The basic model is a classic 2x2 difference-in-difference design. The "With Controls" columns include the same controls used in the RDD estimates in the main paper—specifically sector, recipient, and start year (all included as dummy variables).

The RDD estimates presented in the main paper are preferred: they compare projects just above and below the QAU threshold, which are most likely to be similar, and over which bandwidth manipulation is most likely to occur (so effects of negative selection are most likely to be observed). Further, the use of controls in difference-in-difference estimation can lead to difficulties of inference, and these estimates should be treated with some caution. For a careful and up-to-date discussion of these issues, see Roth et. al. (2023).

4

What motivates public sector workers to form peer networks? Evidence from a survey experiment

What motivates public sector workers to form peer networks? Evidence from a survey experiment

Ranil Dissanayake* Sarah Thompson† Dan Honig‡ Elizabeth Linos§

September 2022

Abstract

Peer networks are common in the public sector, but little is known about what motivates public sector workers to engage in them. We use a pre-registered survey experiment with a diverse sample of 1354 public servants to investigate whether framings focusing on career benefits, benefits to the end-users of public services or the personal happiness and emotional satisfaction of the public servants themselves increase self-reported likelihood of participating in a peer network relative to a neutral control. None of these framings have significant average treatment effects on likelihood of participation. We leverage pre-existing heterogeneity among respondents to investigate if there is differential response to some framings according to respondent motivation and engagement. We find some evidence of this, though only respondents with high levels of work engagement and low extrinsic motivation have a significant overall response to any treatment, reporting lower likelihood of participation when gains are framed in terms of emotional satisfaction. These results suggest that some public servants do respond differentially to framings that appeal to their motivation or engagement levels, even when, on average, there is no overall effect.

Keywords: Bureaucracy, Public Organization, Organizational Behaviour

1 Introduction

Peer networks are common across the public sector, both organized by public sector entities themselves (such at the Government Leadership College in the UK), and by third parties with an interest in public sector performance (such as those organized by the University of Oxford's Government Outcomes Lab). These networks are usually organized with the aim of supporting public sector workers or improving their skills and knowledge (Jackson and Bruegmann 2009; Linos et al. 2021; Andrews and Manning 2015b; Roberts et al. 2018), but they are also costly and may not achieve their objectives. And many public sector workers report being substantially overworked (Diehl et al. 2021; Phillips 2020; Wood 2019). Better understanding why public sector workers join peer networks, and why public sector

*Center for Global Development and Blavatnik School of Government, University of Oxford
†Evidence Action
‡University College London
§Harvard Kennedy School

workers take costly action in their working life more generally is of policy and research importance.

This study investigates why public sector workers take voluntary, bottom-up action to engage in peer networks, conditional on already having entered into (or been selected into) the public sector and how pre-existing differences in their characteristics explain differences in this using a survey experiment implemented with 1,354 public sector workers active on the online survey platform Prolific. Our respondents are a diverse sample of public sector workers from 28 countries, and a wide range of sectors and jobs within the public sector. We investigate whether framings priming extrinsic motivation, intrinsic motivation or personal happiness and emotional satisfaction at work, relative to a neutral control, are associated with a higher expressed likelihood of participating in a peer network, supplementing this with pre-registered analysis of heterogeneity according to respondent characteristics.

None of the framings we investigate have any overall (average) treatment effect: they do not cause higher self-reported likelihood of participating in a peer-to-peer network across the whole sample. Leveraging pre-existing heterogeneity among respondents we investigate if some framings elicit a stronger response among some public servants. In line with the Expectancy Theory of Motivation (Vroom 1964), we might expect that workers will work harder towards ends they particularly desire, and be more willing to take costly action in pursuit of them. To the extent that the ends they desire are heterogeneous across workers, we may then expect that different workers will respond differently according to which ends are emphasised as an expected outcome of the network. There is some evidence of this: we find some differential response to treatment framings according to the expected respondent characteristics but significant treatment effects are found only for the personal happiness and emotional satisfaction framing, among those respondents who already have a high level of work engagement and relatively low extrinsic motivation, who respond negatively. Two pre-registered robustness tests (using a binary dependent variable and using ordered probit estimation) suggest that these results are sensitive to choice of dependent variable and functional form, though the main, pre-registered, dependent variable and functional form are preferred for conceptual and statistical reasons. [1]

These results suggest that different ways of framing at least some organizational initiatives (here, peer networks) may not have any overall effect on how most public servants respond to them, though there is some heterogeneity in responses, and a subset of public servants respond to some framings. [2] If interventions are aimed at specific sub-groups of public servants, better understanding if and how they differ in their response to ways of framing organizational initiatives is useful.Indeed, public servants may respond quite differently to the same motivations. While those with higher extrinsic motivation responded more positively to framings around personal happiness and emotional satisfaction than those with lower extrinsic motivation, those with higher work engagement respond more negatively than those with lower work engagement.

This experiment contributes to literatures on public sector worker motivation, selection, retention and behaviour, as well as a nascent literature focusing specifically on peer learning and support networks in the public sector (Andrews and Manning 2015a,b; Linos et al. 2021). The role of intrinsic and extrinsic motivation has been investigated for at

[1] Specifically: the dummy dependent variable provides information only on 'crossing' a specific threshold of enthusiasm; as our PAP set out, we are interested in changes in enthusiasm for peer networks across the entire scale. Use of the probit specification is contra-indicated by a Brant test; we report it nevertheless, since it was pre-specified.

[2] It is also possible, of course, that the treatment is too weak to elicit a strong response and a more concerted attempt to frame benefits would have larger effects.

least half a century (Lepper et al. 1973; Deci 1975; Deci and Ryan 1985) and plays a central role in both the theoretical contracting literature (Acemoglu et al. 2007; Bénabou and Tirole 2003; Besley and Ghatak 2005) and in empirical studies (Ali et al. 2021; Ashraf et al. 2014, 2020; Belle and Cantarelli 2015; Georgellis et al. 2011; Park and Word 2012). Much of this literature looks at selection of public servants depending on the form of motivation invoked in recruitment, and their performance thereafter. Our study sheds light on motivation once already selected.

The possibility that emotional satisfaction is an important motivating factor in organizations has equally long pedigree. In Kenneth Arrow's The Limits of Organization, he wrote that collective enterprise may be driven by "those who seek deeper emotional satisfaction" (Arrow 1974, p. 16) but then sets this aside in order to consider only the 'rational spirit' of organization. Charles Perrow similarly devotes several pages of Complex Organizations to a discussion of 'norms and sentiments' of workers in large organizations and their importance for how the organization function, but concludes that the empirical work to date on these issues was inconclusive, with observational studies finding no relationship to productivity but a positive relationship with absenteeism and turnover (Perrow 1979, p. 90-99). This has broadly remained true as more work on the topic has been conducted (Reizer et al. 2019; Hosie et al. 2012). More recent, experimental work, has complicated the first part of this picture at least, with personal happiness in both the lab and field found to cause—in at least some settings—increased productivity (Oswald et al. 2015). For the most part, however, modern experimental work with public bureaucracies has tended to neglect the importance of the emotional satisfaction of workers. Yet, with staff retention and motivation serious and long-standing problems in both developed and developing countries (Eldor 2018; Linos et al. 2021; Willis-Shattuck et al. 2008), this may be missing an important aspect of the long-term performance of public bureaucracies. Our experiment provides new, experimental evidence in this domain.

Our experiment and empirical strategy was pre-registered in the EGAP Registry (Registration no: 20220806AB). The analysis in this paper follows the pre-analysis plan (PAP) we registered.

This paper proceeds as follows: the next section sets out our motivation. Section 3 describes our specific research hypotheses. Section 4 describes our survey experiment and survey tool. Section 5 describes the data we collected and coding choices. Section 6 sets out the empirical strategy we follow (pre-specified in our PAP). Section 7 describes our sample and recruitment procedure. Section 8 presents our results and Section 9 concludes with a discussion and implications for policy.

2 Motivation

The improvement of public sector service delivery and functioning might plausibly come from recruitment of "better" people or motivating them with "better" systems of reward and management (however 'better' might be defined), or from incorporating these workers in a better-functioning organizational structure. These are difficult problems to solve in a static, one-period set up (how do we identify and hire the best, most motivated people?), but become even more difficult viewed dynamically: what if the way in which people are recruited or incentivised have implications for their retention and future performance, for example?

While studies have effectively demonstrated that different ways of recruiting, contracting or managing public sector workers have implications for short- and medium-term performance, we also know that many public servants report overly-burdensome workloads, emotional exhaustion and low engagement in their work, including a low sense of personal achievement. In the UK civil service, for example, for the more than a decade, roughly 40% of staff have reported that their workload is not reasonable; only around half report that their organization motivates them to achieve its objectives; and fewer than a third believe that their pay is reasonable compared to others working in similar organizations.[3] This is not unusual: similar results have recently been found in the US and New Zealand, for example (Liss-Levinson 2022; Plimmer and Cantal 2016). Public sector workforce retention, recruitment and motivation are issues of enormous popular public salience, in the news in (to name just a few): the US[4] (Gottlieb and Bauer 2022), the UK[5], Australia and New Zealand[6], Japan[7] and South Africa[8]—even including places where public sector jobs are relatively highly paid and secure.

As part of their attempts to address these problems, public organizations may provide or encourage the use of peer networks in order to support innovation and performance (Jackson and Bruegmann 2009), happiness at work and retention (Linos et al. 2021), the transfer of 'tacit' knowledge (Andrews and Manning 2015b) and team-working, technical skills and leadership (Roberts et al. 2018). However, peer networks may not always be beneficial, either from an organizational or individual perspective. There are time costs to participation, and networks may not achieve their intended goals. In this context, it is valuable to understand what motivates over-worked and burnt-out public sector workers to engage in peer networks. Learning more about this is valuable in its own right, but also provides insight on what aspects of their working life people who have already selected into (and been selected into) public service are willing to voluntarily invest in to improve or enhance—that is, what they feel is missing, at the margin, from their working life. Our experiment investigates both of these questions.

Most studies investigating peer-to-peer networks, however, have looked at existing—often compulsory—networks (Roberts et al. 2018) or networks or peer support which staff are randomized into by researchers and management (Linos et al. 2021). We know little about what motivates over-worked and burnt-out public sector workers from engaging in a peer network themselves, what would encourage voluntary engagement in these networks. Learning more about this is valuable in its own right, as if they do successfully support public sector employees, we still need to be able to induce them to use such networks (Andrews and Manning devotes and extended section to discussing the need to motivate participation in peer networks in their 2016 how-to guide).

We lean on a rich seam of theory and empirical work to motivate our framing experiment investigating why public sector workers sign up for peer networks. Much of the literature on public sector workers focuses on what motivates them to seek these jobs, and to exert effort once in them. In this literature, one source of motivation is their 'mission

[3] All figures taken from the statistical release of the 2021 People Survey, available here: https://www.gov.uk/government/publications/civil-service-people-survey-2021-results, specifically the benchmark scores.

[4] https://www.careersingovernment.com/tools/gov-talk/about-gov/education/the-impact-of-workplace-culture-on-public-sector-employee-retention/

[5] https://www.theguardian.com/society/2022/feb/26/stressed-nhs-staff-quit-at-record-rate-of-400-a-week-fuelling-fears-over-care-quality

[6] https://www.afr.com/politics/federal/four-out-of-10-public-sector-workers-are-considering-changing-jobs-20220314-p5a4ej

[7] https://mainichi.jp/english/articles/20200620/p2a/00m/0na/004000c

[8] https://allafrica.com/stories/202109210207.html

drive' and the extent to which they are motivated by concerns intrinsic to the work.

Both in theory and in practice there is support for the view that at least to some extent the reward for which public sector workers exert effort is that of doing the job well, because they care about the effects of doing the job well for end users or society as a whole, or because they take direct utility from the work. These workers select in to specific jobs at least in part due to the match between organizational mission and what values or outcomes motivate them intrinsically (Besley and Ghatak 2005; Dewatripont et al. 1999), and once selected they exert effort for the reward of this specific job done well (Grant and Hofmann 2011; Kamenica 2012; Banuri et al. 2018).

Similarly, there are both theoretical and empirical foundations to the view that public sector workers seek rewards unrelated to the specific mission of the organization: salary payment and increases, bonuses, promotion and the 'glow of recognition' of a job well done, as well as better future prospects from being a recognized high performer (Gertler and Vermeersch 2012; Muralidharan and Sundararaman 2011; Karachiwalla and Park 2017; Ashraf et al. 2014; Leaver 2009; Iyer and Mani 2012). In the economics literature, at least, extrinsic motivation tends to be narrowly defined as either pecuniary reward (and very often, specifically piece-rates or pay-for-performance) or promotion or similar career prospects (such as the prospect of a new job in another institution).

But there is another motivating force for many workers in organizations (public sector and otherwise): their personal happiness and sense of belonging. As already noted, the wellbeing of workers and the role of organizations as social settings from which happiness and connectedness are generated by workers has long been recognised by organizational theorists (Arrow 1974; Perrow 1979). And indeed, conditional on already having 'matched' by mission and selected into a contract with an organization, such matters may be absolutely of first order importance to workers. In recent research using the European Working Conditions Survey (not limited to public sector workers), Nikolova and Cnossen (2020) show that the single most important factor in work meaningfulness for employees is 'relatedness'; that is a sense of belonging, or important interpersonal relationships. In non-experimental work, happiness was correlated with quit rates among Head Start teachers (Wells 2015) and creativity at work in the private sector (Amabile et al. 2005). It is—in at least some settings—causally related with productivity both in the lab and in the field (Oswald et al. 2015). And it appears to be relatedness which drives the positive effects of peer connection on burnout in the aforementioned Linos, Ruffini and Wilcox study (Linos et al. 2021). In short, research from a range of settings demonstrates the importance of emotional satisfaction, relatedness and happiness to worker (and employer) outcomes. Personal happiness and emotional satisfaction at work is usually cast as something separate to either extrinsic or intrinsic motivation, though in popular coverage of the workforce issues (including but not limited to the public sector) in a range of countries it usually provides a central role.[9]

We expect public servants to have different motivational make-ups. We might expect this heterogeneity to lead them to respond to different framings, in line with foundational theories of workplace motivation. In the expectancy theory of motivation (Vroom 1964), workers work harder for ends they in particular desire and think it possible to achieve.

[9]For example, see this CNBC coverage across sectors in the US https://www.cnbc.com/2022/08/12/job-unhappiness-is-at-a-staggering-all-time-high-according-to-gallup.html; this piece about the salary UK workers would forgo for a happier workplace https://www.theguardian.com/money/2022/jul/07/uk-workers-would-take-pay-cut-above-average-happiness-study; and this article associating staff turnover in the UK's National Health Service with stress in the workplace: https://www.theguardian.com/society/2022/feb/26/stressed-nhs-staff-quit-at-record-rate-of-400-a-week-fuelling-fears-over-care-quality

In this theory, motivation governs choices over voluntary action—which well-describes what we study in our survey experiment. Thus, we could expect more extrinsically motivated individuals will work harder to receive promotions than those who are less extrinsically motivated; those who care more about fulfilling the mission of the organization will be more swayed by a framing regarding beneficiaries than those less intrinsically motivated; and those who desire more engagement with their job will respond more to a framing regarding emotional satisfaction at work. In other words, within our sample, heterogeneity of motivational makeup and state might also suggest that there should be heterogeneity in how workers respond to the same stimulus, and hence their willingness to exert effort in response to it.

3 Hypotheses

We test three main hypotheses, and for heterogeneity of treatment effects within them. We describe the concepts and scales used for heterogeneity analysis below. Specifically, we hypothesize:

1. Framing the gains of participating in a peer network in terms of career benefits causes higher reported likelihood of participating in a peer network relative to a neutral control. This is our extrinsic motivation (EM) treatment.
   (a) Making extrinsic motivation more salient has a larger effect on more extrinsically motivated respondents. That is, the effect of the EM treatment varies according to the pre-existing level of extrinsic motivation of the respondent.
   (b) The EM treatment is not responsive to pre-existing levels of intrinsic motivation or work engagement.
2. Framing the gains of participation in a peer network in terms of benefits to end users of public services causes higher reported likelihood of participating in a peer network relative to a neutral control. This is our intrinsic motivation (IM) treatment.
   (a) Making intrinsic motivation more salient has a larger effect on more intrinsically motivated respondents. That is, the effect of the IM treatment varies according to the pre-existing level of intrinsic motivation of the respondent.
   (b) The IM treatment is not responsive to pre-existing levels of extrinsic motivation or work engagement.
3. Framing the gains of participation in a peer network in terms of the emotional satisfaction and personal happiness of respondents causes higher reported likelihood of participating in a peer network relative to a neutral control. This is our emotional satisfaction (ES) treatment.
   (a) To the extent that work engagement is a consequence of happiness and satisfaction at work, we may expect the ES framing to be responsive to pre-existing levels of work engagement.
   (b) The ES framing may be positively associated with either intrinsic or extrinsic motivation.

Our three main hypotheses test whether framing the benefits of peer network participation in terms of career value, benefits to end users or personal happiness and satisfaction elicit a higher expressed likelihood of participation among respondents. As set out in the previous section, there is an extensive literature which investigates the importance of intrinsic and extrinsic motivation in the public sector; Treatments 1 (EM) and 2 (IM) look at the extent to which framing gains from peer networks in these terms leads to higher reported likelihood of participation. The importance of emotional satisfaction and personal happiness at work is also deeply-seated in the literature (Arrow 1974; Perrow 1979), and survey data suggest it matters greately to workers (Nikolova and Cnossen 2020), and Treatment 3 (ES) allows us to investigate its value as a framing device for organizational initiatives that require voluntary worker effort. We test for heterogeneity in response by respondent characteristics for each. We draw directly on expectancy theory (Vroom 1964) for hypotheses 1(a), 1(b), 2(a) and 2(b). Specifically, we might expect that those respondents who are most extrinsically motivated will respond most to the framing based on career value, as there is a clear match between what they desire (extrinsic or intrinsic benefits) and how we frame peer networks in terms of what they can be expected to help them achieve. Heterogeneity in response to the emotional satisfaction framing is something more of a black box. To the extent that personal happiness and satisfaction are, in a work context, related to engagement at work (and given that the measure we use measures vigor and enthusiasm, which are plausibly related to emotional satsifaction in the work context at least), we might expect that those who desire greater engagement with work may be more responsive to this treatment. However, it is also worth investigating the effect of the emotional satisfaction framing with respect to extrinsic and intrinsic motivation.

We discuss the conceptual and empirical basis of these scales along which we measure heterogeneity of respondents, and how they relate to each other, clarifying how they are conceptually and empirically distinct, in the methods section, below.

The next section details exactly how our experiment is set up.

4 Method

We collected our data entirely through a survey instrument, implemented using Qualtrics. The full survey instrument is included as Annex B. The survey comprised four components (mostly interleaved, with some separate):

- The randomized survey vignette experiment
- a series of questions about the respondent (personal characteristics, characteristics of their job, intrinsic and extrinsic motivation, happiness in their job, the extent and function of their existing professional network);
- a series of questions about the organization for which they work (sector, size, trajectory, and so on); and
- questions about the characteristics the respondent would value in a peer-to-peer network.

The survey was piloted twice to check that all the questions worked correctly (that is, none had been miscoded or were impossible to answer), that there was sufficient variation in responses, and to remove questions which provided

no useful additional information. These pilots took place in late 2021 and early 2022 respectively.

The central experiment was implemented through a randomized survey vignette. It asks respondents about their willingness to participate in a peer-to-peer support network the researchers are designing, and suggests what kinds of benefit this network might bring. The benefits suggested to the respondent were experimentally varied. The vignette is structured as follows:

> *(VIGNETTE PAGE 1)*
>
> *We are designing a new initiative to support public sector workers through peer-to-peer engagement, organized in small groups. The goal of these groups is to support you as a person and as a public sector employee.*
>
> *Engaging in these groups would require a small but regular commitment of time and effort. By joining one of these groups, you will be supported personally and professionally by the other members of the group.*
>
> *(THE NEXT SENTENCE IS EXPERIMENTALLY VARIED)*
>
> ***Control group**: Based on past experience and feedback we expect these groups to have a number of benefits for participants.*
>
> ***Extrinsic motivation treatment**: Based on past experience and feedback we expect these groups to bring career benefits to participants, by sharing information and advice that help you advance in your career.*
>
> ***Intrinsic motivation treatment**: Based on past experience and feedback, we expect these groups to help participants better support the people your agency serves, by sharing information and advice that helps your organization improve the way it supports these end-users.*
>
> ***Emotional satisfaction treatment**: Based on past experience and feedback we expect these groups to provide emotional support and help you become a happier, more balanced, more satisfied person.*
>
> *(EXPERIMENTAL VARIATION ENDS)*
>
> *If offered the opportunity to join a group like this, how likely would you be to join?*
>
> 1. *Very Unlikely (1)*
> 2. *Unlikely (2)*
> 3. *Moderately Likely (3)*
> 4. *Likely (4)*
> 5. *Very Likely (5)*
>
> *(VIGNETTE PAGE 2)*
>
> *We call these peer groups practitioners circles, and they may have a number of benefits:*
>
> - *They may help you advance in your career.*
> - *They may help you support the end users of your organization better*
> - *They may help you become a happier, more balanced, more satisfied person.*

For the rest of the survey, when we refer to Practitioner's Circles, these small peer-to-peer networks are what we mean.

This vignette allowed us to experimentally manipulate the framing of the peer network, making intrinsic or extrinsic motivation, or personal emotional satisfaction the most salient outcome of such a network. By investigating how this experimental manipulation affected the self-reported likelihood of respondents participating in such a peer network we can learn something about what workers themselves want in their professional lives. Page 2 of the vignette 'resets' expectations and framing for the peer network by exposing all respondents to the same potential benefits before we go on to ask additional questions.

To investigate heterogeneity of treatment effects, we also collect information on intrinsic and extrinsic motivation and work engagement. Though related, these are conceptually and empirically distinct characteristics. Conceptually, public servants may be more or less extrinsically motivated; at the same time, they may be more or less intrinsically motivated (and we make no assumptions about whether these motivations are in conflict, mutually reinforcing or entirely independent). These are 'types' of public servant. Depending on their personal and workplace circumstances they may also be more or less engaged in their work—that is, they may draw more or less interest and satisfaction from it. This is a 'state' any public servant may be in at a given time. In our framework (and as measured by the metrics used, elaborated below), we expect intrinsic motivation and work engagement to measure different things. Though work engagement is a component of intrinsic motivation in self-determination theory, intrinsic motivation measures general preferences for work, while work engagement measures a specific level of engagement with a specific job at a specific moment in time, and thus can vary for the same individual both across jobs and across time within a job. Further, the component measures of these indicators we use are empirically distinct.

For the intrinsic and extrinsic motivation, we use a shortened version of the Work Preferences Inventory. The original was a thirty-item scale divided into two 'motivational orientations' (intrinsic and extrinsic), each further divided into two sub-categories (Amabile et al. 1994). The shortened version consists of only ten questions, a more parsimonious approach that loses little in validity or reliability (Robinson et al. 2014). It is explicitly designed so that intrinsict motivation and extrinsic motivation may be present in the same individual since intrinsic and extrinsic motivation are not mutually exclusive within an individual. The Work Preference Inventory investigates extrinsic motivation through three questions that measure the extent to which respondents value how others see and relate to their performance at work (that is, the recognition they earn from others, the extent to which others learn about their competence, and the extent to which they agree that success means 'doing better than other people) and two questions about the extent to which compensation (income and promotion) motivate respondents. The intrinsic motivation scale is measured through three questions that focus on the extent to which motivation derives from the challenge posed by work (tackling new problems, solving problems and the difficulty of the work) and two questions about how much enjoyment of work itself motivates respondents.

For work engagement we use the ultra-short version of the Utrecht Work Engagement Scale-3 (Schaufeli et al. 2019), again a much more parsimonious version of the original 17-item scale (Schaufeli et al. 2002) without loss of reliability or validity. The UWES-3 scale measures the vigor, dedication, and absorption of respondents. The UWES-3 is correlated

with, but distinct from, burnout. Table 1 sets out the questions used to construct the extrinsic motivation, intrinsic motivation and engagement scales we use. Each measures quite different things, and it is clear that intrinsic motivation captures the underlying preferences of the respondent, while engagement captures the specifics of their current job.

Table 1: Motivation and Engagement Index Construction

Work Preference Inventory-Extrinsic	**Work Preference Inventory-Intrinsic**	**UWES-3 Work Engagement**
I am strongly motivated by the recognition I can earn from other people.	I enjoy tackling problems that are completely new to me.	At my work, I feel bursting with energy
I want other people to find out how good I really can be at my work.	I enjoy trying to solve complex problems.	I am enthusiastic about my job
To me, success means doing better than other people.	The more difficult the problem, the more I enjoy trying to solve it.	I am immersed in my work
I am keenly aware of the promotion goals I have for myself.	What matters most to me is enjoying what I do.	
I am keenly aware of the income goals I have for myself.	It is important for me to be able to do what I most enjoy.	

The Work Preferences Inventory intrinsic motivation score measures the kind of problem a respondent likes to work on; and their preference for enjoyment of their task at work. The UWES-3 Work Engagement Score measures day-to-day vigour at work, enthusiasm at work and immersion in work. Comparing the five intrinsic motivation questions and the three work engagement questions side-by-side illustrates both key points (both measure agreement with the statements listed). On the left hand side, the WPI-Intrinsic statements are about general preferences for work (the 'type' of public servant one is, and not specifically related to the job at hand). On the right hand side, the Work Engagement Score relates to the current job conditions (a 'state' in which the public servant is in). They also measure different things. Intrinsic motivation includes engagement in the specific tasks which a job entails. Work engagement captures the broader emotions associated with work ('bursting with energy', 'enthusiasm' and absorption, as measured by immersion in the task). Since intrinsically motivated workers may be more or less successful in finding work that activates their intrinsic motivation, we might expect variation in engagement at any level of intrinsic motivation, as measured by these metrics. However, as we will see in the results section, in practice these scales are positively correlated. While there remains variation, as expected, it is the case that more engaged respondents tend also to report higher intrinsic motivation.

The use of a survey experiment to investigate this question has some drawbacks. Two in particular are worth drawing attention to. Firstly, some (but not all) respondents will have prior experience of peer networks and this experience will inevitably colour their response to the survey vignette. With a sufficiently large sample the hope is that these types will be randomly distributed across treatment and control arms. We collect data on the extent of the networks already engaged in (detailed in the data section), and control for this in our main specifications, but do not directly measure the extent and quality of previous exposure to peer networks specifically. Secondly, in the context of a survey experiment, indicating likely engagement in a peer network is essentially costless (though we attempt to signal the existence of costs, using the language 'a small but regular commitment of time and effort' in the experimental vignette).

In real life, of course, participation in a peer network is costly. The results of this experiment should properly be conceived of as an estimate of the enthusiasm for the idea of joining a peer network, not as an estimate of actual likelihood of participation. This will inevitably be lower than indicated in results here, due to the time and effort costs of participation in real life.

We describe the additional data we collected, and how the data were coded, in the next section.

5 Data and coding

The dependent variable for our primary analysis will be the 1-5 score of likelihood of participating in a peer-to-peer network, as described in the previous section.

Our primary variables of interest are dummies for which framing treatment the respondent was exposed to. We code assignment to treatment as a series of dummy variables taking the value of 0 for any respondent not assigned to a given treatment and 1 for any respondent assigned to that treatment, with respondents assigned to the control group left as the reference group.

The interaction variables for heterogeneity analysis we use are:

- Intrinsic motivation (WPI intrinsic motivation score: 1-4, with 4 being the highest)
- Extrinsic motivation (WPI extrinsic motivation score: 1-4, with 4 being the highest)
- Worker engagement (UWES-3 Work Engagement Scale: 1-7 with 7 being the highest)

These scales are constructed according to standard practice as set out in their original formulations (Robinson et al. 2014; Schaufeli et al. 2019).

We also collected a number of additional variables, for use as control variables and in exploratory analysis:

- Respondent gender (binary variable where 1 = female)
- Country (dummy variables taking the value 1 for a specific country, base group Australia).
- Sector of organization (dummy variables taking the value 1 for each of 16 sectors, base group Agriculture)
- Size of organization (dummy variables taking the value 1 for the following size ranges: 50-99 people, 100-199 people, 200-499 people and over 500 people, base group is 0-49 people).
- Government orientation: the extent to which respondents agree that Government is the place for those who want to do public service (numerical variable on a five point scale, with five being stronger agreement)
- Years experience in public sector (numerical)
- Management responsibility (dummy variables taking the value 1 for 'some management responsibilty', 'leader of department' or 'leader of organization', base group no management responsibility)

- Trajectory of organization (answer to the question: "On the whole, do you think things are getting better or getting worse at your organization?", 1-5 scale, with 5 being better)
- Personal agency in organizational trajectory (answer to the question: "Does your work have the potential to impact the organization's trajectory?", 1-5 scale, with 5 being better)
- Direct service delivery (dummy variable where 1 = job involves direct service delivery)
- Job function (dummy variables for 9 different job types, base group 'analysts')
- Extent of existing network (a dummy variable for 'highly networked', summarising data on the extent of the respondents work and personal networks)
- Work challenges identified (summarized to number of challenges identified out of a possible 16)

In addition to the vignette, the survey includes a novel question investigating the extent to which respondents feel their organization relies on 'Route X' (that is, monitoring and the use of incentives to control staff to achieve organizational objectives) compared to 'Route Y', characterized by empowering employees to use their judgement to pursue organizational objectives (Honig 2022; McGregor 1966). This exact question we ask is:

> "In your opinion, what is the balance in your organization between providing freedom for staff to pursue organizational goals (however they are defined) as they see fit on the one hand, and tightly controlling staff through targets, monitoring, oversight and incentive schemes? Use a 1-100 scale where 1 is full freedom for staff and 100 is complete control of staff action."

We include this question to engage with organisational culture, and how it might affect the returns to peer networking. We also ask one question that was not pre-specified in our PAP: "How enthusiastic would you be about joining a practioner's circle?", answered on a scale of 0 (total unenthusiastic) to 100 (fully enthusiastic). We do not include this in our main empirical analyses, but do use it to demonstrate the high level of internal consistency of answers from our respondents.

6 Empirical Strategy

Our analysis follows that which we set out in our pre-analysis plan.

We investigate hypotheses 1, 2 and 3 (that self-reported likelihood of participating in a peer network is affected by whether the framing of gains emphasizes career benefits, benefits to end-users of public service or emotional satisfaction and personal happiness of respondents) by using a basic Ordinary Least Squares specification to identify treatment effects. Specification 1 is our basic model, and only controls for differences between countries, by including country-fixed effects:

$$Y_{ic} = \boldsymbol{\beta Treatment_{ic}} + \alpha_c + \epsilon_{ic} \tag{1}$$

Where i is the respondent and c is the country in which the respondent works. Where Y_{ic} is the likelihood of participating on a 1-5 scale and **Treatment** is a vector of treatment dummies.

The $\boldsymbol{\beta'}$ coefficients are interpreted as the causal impact of altering the framing used to describe the purpose and impact of peer support mechanisms to make intrinsic motivation, extrinsic motivation or emotional satisfaction at work more salient to the respondent.

Our extended specification includes an additional set of control variables.

$$Y_{ic} = \boldsymbol{\beta Treatment_{ic}} + \boldsymbol{\gamma' X_{ic}} + \alpha_c + \epsilon_{ic} \tag{2}$$

Here X is a vector of control variables, specifically:

- Intrinsic motivation (WPI intrinsic motivation score: 1-4, with 4 being the highest)
- Extrinsic motivation (WPI extrinsic motivation score: 1-4, with 4 being the highest)
- Worker engagement (UWES-3 Work Engagement Scale: 1-7 with 7 being the highest)[10]
- Respondent gender (binary variable where 1 = female)
- Sector of organization (dummy variables taking the value 1 for each of 16 sectors, base group Agriculture)
- Size of organization (dummy variables taking the value 1 for the following size ranges: 50-99 people, 100-199 people, 200-499 people and over 500 people, base group is 0-49 people).
- Government orientation: The extent to which respondents agree that Government is the place for those who want to do public service (numerical variable on a five point scale, with five being stronger agreement)
- Job function (dummy variables for 9 different job types, base group 'analysts')
- Years experience in public sector (numerical)
- Management responsibility (dummy variables taking the value 1 for 'some management responsibilty', 'leader of department' or 'leader of organization', base group no management responsibility)[11]
- Direct service delivery (dummy variable where 1 = job involves direct service delivery)
- Extent of existing network (a dummy variable for 'highly networked', summarising data on the extent of the respondents work and personal networks)
- Work challenges identified (summarized to number of challenges identified out of a possible 16)
- Route X orientation of the organization (0-100 variable, with a higher value indicating tighter control)[12]

[10] These first three variables were included in specification 3 in our PAP but are included in both specification 2 and 4 here, with a new specification 3 including interaction effects without controls. They are included in response to comments that they were important control variables for specification 2.

[11] This is a minor deviation from our PAP, in which we specified a dummy variable; the change was taken so as to use all of the information collected in the survey.

[12] The inclusion of this variable as a control was not specified in our PAP, but was included after substantial variation in responses was found. Excluding it has no effect on our central findings.

To test hypotheses 1(a), 1(b), 2(a), 2(b), 3(a), and 3(b), that treatment effects are heterogeneous according to respondent characteristics (specifically, their level of work engagement, intrinsic motivation and extrinsic motivation, measured according to well-established scales used in the public administration literature), we run additional specifications in which we interact the treatment variable with a series of moderators, first without and then with controls.

$$Y_{ic} = \boldsymbol{\beta Treatment_{ic}} + \boldsymbol{\delta Treatment * M_{ic}} + \boldsymbol{\lambda M_{ic}} + \alpha_c + \epsilon_{ic} \tag{3}$$

$$Y_{ic} = \boldsymbol{\beta Treatment_{ic}} + \boldsymbol{\delta Treatment * M_{ic}} + \boldsymbol{\lambda M_{ic}} + \boldsymbol{\gamma' X_{ic}} + \alpha_c + \epsilon_{ic} \tag{4}$$

Where M is, in turn:

- Intrinsic motivation (WPI intrinsic motivation score: 1-4, with 4 being the highest)
- Extrinsic motivation (WPI extrinsic motivation score: 1-4, with 4 being the highest)
- Worker engagement (UWES-3 Work Engagement Scale: 1-7 with 7 being the highest)

In this specification each M variable is also included as a control, separately to the interaction effect. We investigate if any of these interactions significantly alters the strength or direction of the treatment effect, and whether the match of moderator and treatment matters. Our preferred specifications are (2) and (4), including a full set of controls.
As pre-specified, since we evaluate our hypotheses over 12 coefficients, we apply the Benjamini-Hochberg False Discovery Rate (FDR) correction to our results.
In line with our pre-analysis plan, we also undertake two robustness tests. First, we replace our dependent variable with a dummy variable which takes the value of 1 when the respondent reports being likely or very likely to join a peer network and 0 otherwise. This dependent variable evaluates a restricted set of information compared to the main dependent variable, which exploits all of the variation in the data. Secondly, we use an ordered probit regression in place of the linear probability model used in the main specification, testing for appropriateness using the Brant (1990) test.

7 Sample recruitment

Our sample was drawn from the online, paid survey platform Prolific. Prolific users include public servants from a range of countries in Europe, North America, Asia, Africa and Australasia. Survey participants are managed directly by Prolific, allowing for quality control and fair payment for their time. The platform is attractive for its geographic diversity, the generally high quality of responses collected on it, and because it allows pre-screening of respondents. We restricted our sample to those over the age of 18, currently working full-time in the public sector, and who speak English, without any geographic or nationality-based restrictions. At the time the survey was implemented (08/22), this amounted to 2,153 potential respondents. We screened out 200 respondents who responded to one of our two pilots, leaving 1,953 respondents. This is our sampling frame.
Once the survey was made live on Prolific, we set a target of 1,500 responses and allowed responses to run for three

weeks. Respondents were paid £9.95 per hour. Our respondents were simply those who completed the survey without withdrawing consent or 'timing out' (opening the survey but not completing it). Within this three-week time frame, we collected 1447 responses, 96 percent of our target sample size, and 74 percent of the eligible respondents on Prolific. Within this sample, assignment to treatment was randomized, with randomization implemented through the Qualtrics automatic randomization option. The survey sample was not stratified in any way prior to randomization. After completion of the survey we excluded any respondents who failed our standard attention check question (a question which, within in the body of the question text, instructs respondents to select option 2), and any respondents who completed the survey in less than one-third of the estimated survey completion time calculated by Qualtrics (which was 18 minutes, meaning we excluded any respondents completing in fewer than 6 minutes). These exclusions were both pre-registered in our Pre-Analysis Plan. This left us with 1,354 usable responses.

Calculating the statistical power of our experiment was to some extent guesswork, *a priori*. In our pre-analysis plan, we suggested that with 1500 subjects, using alpha=0.05, the study would have 80% power to detect an effect size of 0.5 points (on a five point scale) with a standard deviation as high as 2.5 when comparing the control and any one of the treatments, assuming no attrition from the sample (and without accounting for the use of control variables to reduce standard errors). We fell very slightly short of this number of respondents, but these calculations were likely to have been too conservative, given that they did not account for the effect of any control variables in reducing standard errors.

The next section describes our respondents and presents the results of our experiment.

8 Results

We drew a diverse sample, from a wide range of countries, with respondents reporting varied levels of seniority, sector of work, and job characteristics. Table 2 summarises the basic descriptive characteristics of our sample (excluding some factor variables for space reasons).

A few points about our sample are worth pointing out. They tend to be relatively experienced public sector workers, with a median of 12 years experience, but variation is wide: it includes some new joiners yet to complete their first year and at least one respondent with half a century of service. They tend to be relatively junior: though organizational level is not reported in this table, more than 50% of respondents report that they have no management responsibility, and the median number of staff managed is 0, though again, there is wide variation. Almost 60% of the sample is female, and a similar proportion report that they work in service delivery roles. The vast majority work for very large public sector entities, with 77% of respondents reporting that more than 500 people work for their organization. And they work in a wide range of countries: 28 are reported.

Digging deeper, we present jitter plots to show how different respondent characteristics correlate across our sample. Jitter plots are scatter plots with random noise added to the variables for clarity. This helps visualize data, especially where one or both variables is an ordered factor rather than continuous. In cases where one of the variables is continuous, we also fit a local trend line. Figure 1 shows that, among respondents, there is generally a positive

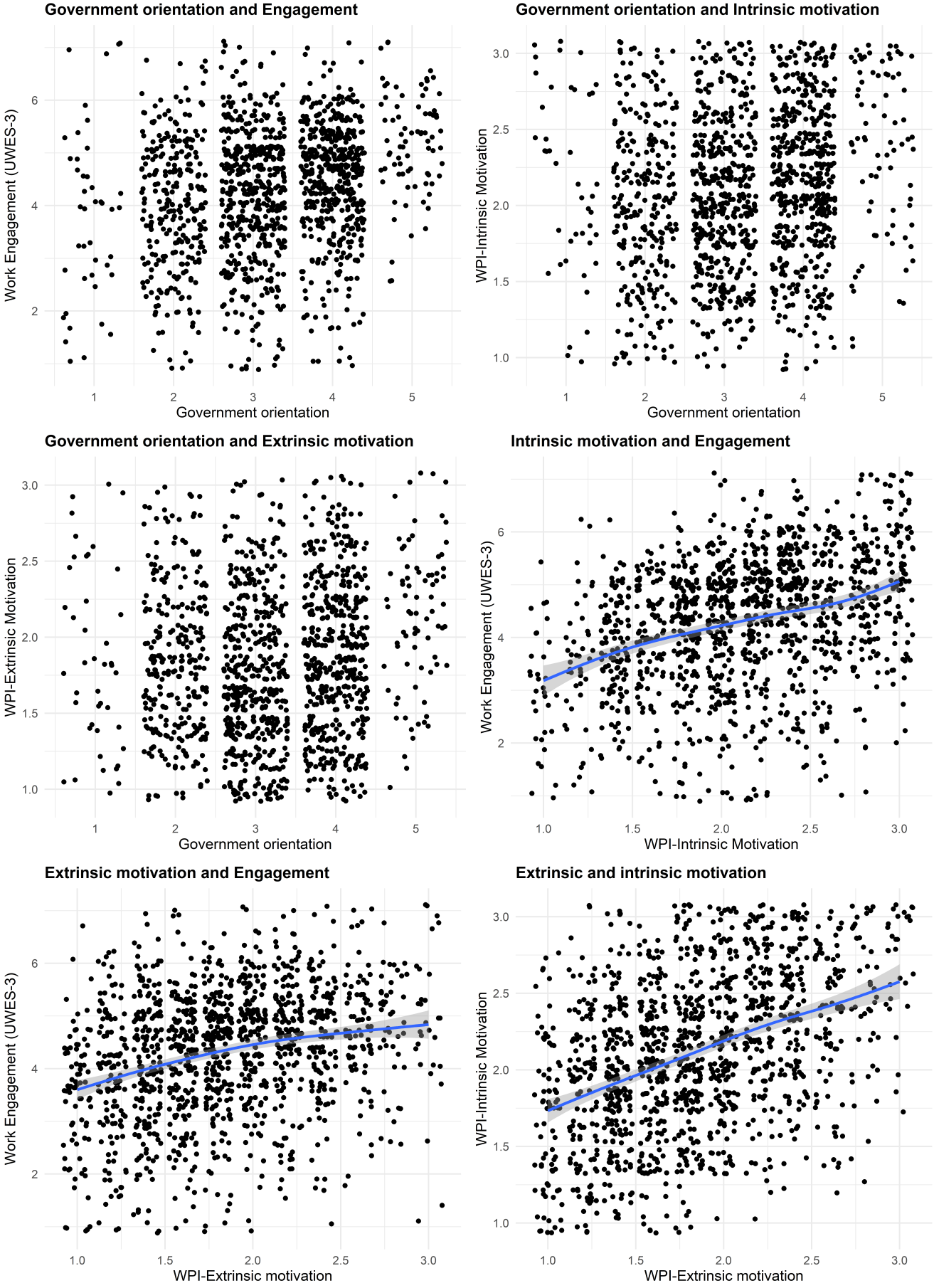
Table 2: Summary of Data

Variable	Unique	Missing (%)	Mean	SD	Min	Median	Max
Likelihood of joining peer network	5	0	3.14	1.08	1	3.00	5
Time taken	795	0	923.48	681.96	360	754.00	13813
Route X orientation (org)	98	0	56.36	24.04	0	60.00	100
Government orientation	5	0	3.20	0.94	1	3.00	5
Staff managed	37	0	3.32	11.05	0	0.00	250
Years experience	56	0	14.88	10.91	0.00	12.00	50.00
Challenges identified	17	0	5.37	3.38	0	5.00	16
Work Engagement	19	0	4.28	1.21	1.00	4.33	7.00
Extrinsic motivation score	11	0	1.84	0.50	1.00	1.80	3.00
Intrinsic motivation score	11	0	2.11	0.51	1.00	2.20	3.00
Highly Networked	2	0	0.22	0.42	0	0.00	1
Female	2	0	0.58	0.49	0	1.00	1
Service Delivery	2	0	0.57	0.50	0	1.00	1

Note: This table omits some variables (country, sector of organization, role in organisation, and organization size) for clarity. All three are included in Table 2, below, examining balance across treatment arms. Route X orientation

correlation between characteristics relating to their motivation, engagement and commitment to the public sector. It shows generally positive associations with government orientation and extrinsic and intrinsic motivation, as well as work engagement; and similarly positive associations between combinations of extrinsic and intrinsic motivation and engagement. However, there is also substantial variation: a number of respondents report high intrinsic motivation but low engagement, for example; or high intrinsic motivation and low extrinsic motivation. As such, we might expect different results depending on the measure used. The same is true for extrinsic motivation and engagement: though there is a positive correlation, there is also wide dispersion, so at each level of extrinsic motivation, there is a wide range of values of work engagement in the data. The graphs suggests that more motivated public servants are typically also the more engaged ones, but that there is nevertheless significant variation in these associations. Of particular note is that more extrinsically motivated public servants are generally also more intrinsically motivated but there is a moderately large concentration of responses in the top left quadrant of that plot: respondents who are highly intrinsically motivated, but not very extrinsically motivated. There are relatively fewer who are highly extrinsically motivated by only weakly intrinsically motivated. This descriptive result is in line with recent findings that recruitment that uses extrinsic motivation to attract talent does not sacrifice intrinsic motivation (Ashraf et al. 2020).

Figure 1: Jitter plots of respondent characteristics



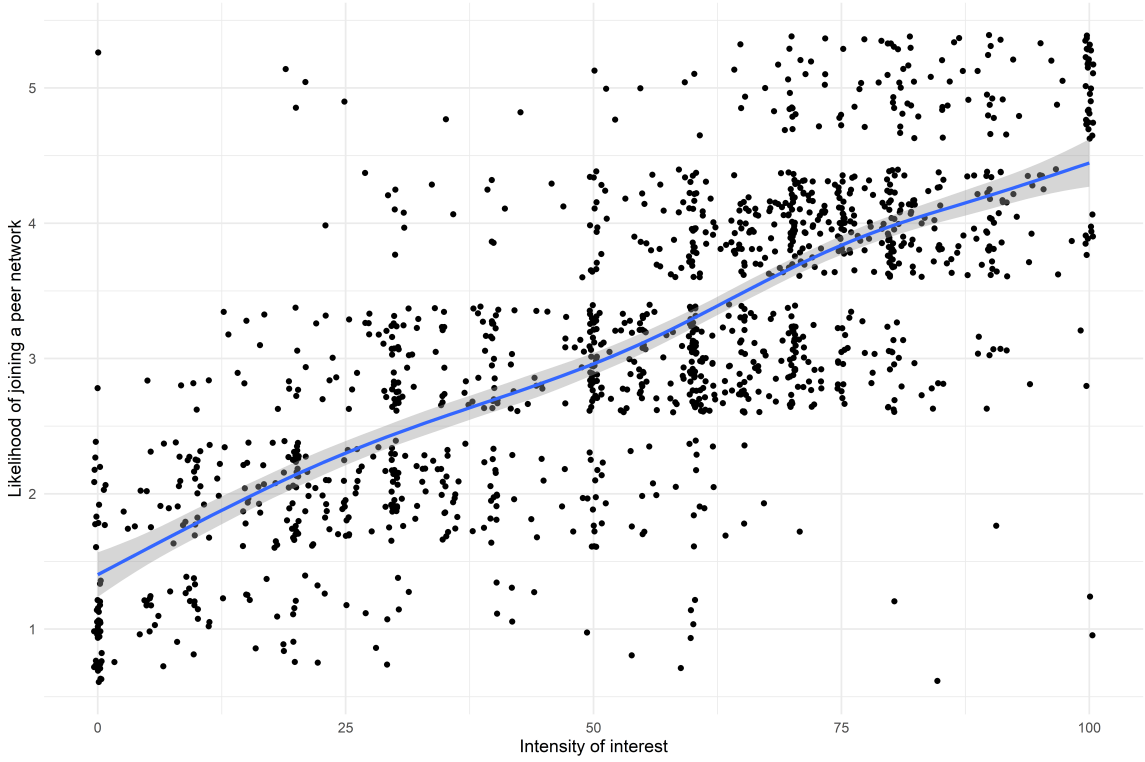
Government orientation is increasing in strength of agreement with the statement: 'Government is the best place for those who want to do public service'. The UWES-3 score is increasing in work engagement. The WPI-Intrinsic score is increasing in intrinsic motivation. The WPI-Extrinsic score is increasing in extrinsic motivation.

Looking at responses by country, we find that there is substantial variation in respondent characteristics within country, more so than there is across countries. Figures 3 to 5 in Appendix C present histograms of intrinsic motivation, extrinsic motivation, work engagement and Route X/Route Y orientation of the institutions respondents work for, for all countries for which we have more than 5 observations. Using a fixed effects estimation controls for the (smaller) variation across countries and uses the larger variation within countries to estimate the effects of our treatments.

Across the sample, peer networks are generally popular. The mean score on the likelihood scale of participation in a peer network in the control group (that is, without complication by any experimental framing) is 3.13 out of 5; 38% of respondents report being likely or very likely to participate.[13]

Lastly before moving on to the results of our main experiment, to allay concerns that respondents were responding at random in order to complete the survey quickly and receive payment, as well as excluding respondents who failed the attention check or completed the survey too quickly, we also checked for internal consistency of results by plotting the correlation between our 1-5 outcome variable and a question about how enthusiastic respondents are about joining a peer network, on a scale of 0-100, placed after an interregnum of several questions. Though they measure slightly different things (it is possible to be highly likely to join something that you suspect may help your career, but to be deeply unenthusiastic about the prospect of doing so), one might reasonably expect a strong correlation between the two measures if respondents are doing something more considered than 'button mashing' to complete a survey. This is exactly what we find.

Figure 2: Internal consistency of responses



Intensity of interest is a variable asking the question 'How enthusiastic would you be about joining a Practitioner Circle? - 0 (not enthusiastic) to 100 (extremely enthusiastic)'; the likelihood of joining a peer network varies from Very unlikely (1) to very likely (5)

[13] Throughout the text, we will use 'likelihood' as a shorthand for the score out of 5 on the likelihood of participating in a peer network scale.

While there are a few outliers, there is a strong correlation between the two measures, and a Pearson's product-moment correlation test strongly rejects the null of no association (estimated correlation is 0.75, p=0.000).

Our experiment randomly assigned this diverse group of public servants into one of three treatments or a control group. Table 3 presents the balance of key variables across our control and three treatment groups.

Table 3: Balance across Treatment Arms

		0 (N=355)		1 (N=336)		2 (N=347)		3 (N=316)	
		Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
Likelihood of joining peer network		3.1	1.1	3.2	1.0	3.2	1.1	3.1	1.1
Time taken		925.2	632.3	989.8	954.7	865.4	499.7	914.8	544.3
Route X orientation (org)		55.5	24.2	54.9	23.8	56.9	23.9	58.3	24.3
Government orientation		3.3	0.9	3.2	0.9	3.2	1.0	3.2	0.9
Staff managed		2.7	5.5	2.6	5.6	3.9	15.3	4.1	14.1
Years experience		14.7	11.2	14.1	10.9	15.4	11.1	15.5	10.3
Challenges identified		5.3	3.4	5.2	3.3	5.6	3.4	5.3	3.4
Work Engagement		4.4	1.2	4.3	1.3	4.3	1.2	4.3	1.2
Extrinsic motivation score		1.9	0.5	1.8	0.5	1.8	0.5	1.9	0.5
Intrinsic motivation score		2.1	0.5	2.2	0.5	2.1	0.5	2.1	0.5
Service delivery role		0.6	0.5	0.6	0.5	0.6	0.5	0.6	0.5
Highly connected		0.2	0.4	0.2	0.4	0.2	0.4	0.2	0.4
Female		0.6	0.5	0.6	0.5	0.6	0.5	0.6	0.5
		N	Pct.	N	Pct.	N	Pct.	N	Pct.
Organization size	0-49	20	5.6	15	4.5	18	5.2	12	3.8
	50-99	9	2.5	12	3.6	11	3.2	10	3.2
	100-199	18	5.1	19	5.7	16	4.6	20	6.3
	200-499	43	12.1	35	10.4	26	7.5	28	8.9
	$\geq$ 500	265	74.6	255	75.9	276	79.5	246	77.8

Note: Route X orientation is a 0-100 scale on which a higher number indicates that the organization the respondent works for is more inclined to use targets, incentives, and rules to manage staff than to allow them freedom to pursue their objectives. Government orientation is the response on a 1-5 scale to the question "How much do you agree that the government is the place for those who want to do public service?" Work Engagement is the score on the Utrecht Ultra-Short Work Engagement Score. Extrinsic and Intrinsic motivation scores are from the 10-question Work Preferences Index. Service delivery role is a dummy variable for an individual in a service delivery role. Highly connected is a dummy variable for whether the number of connections reported is higher than the 75th percentile of scores. Country, sector of organization, and type of role are omitted for brevity.

Again, some variables are omitted for brevity's sake. This table suggest randomization was successful; the characteristics of the control and each treatment group are strikingly similar, with no significant differences between control and any of the treatment arms.

Table 4, presents our results. Recall that our preferred specifications were (2) and (4), both including a full set of controls. The former tests hypotheses 1, 2, and 3, that any of the three alternative framings (designed to appeal to intrinsic motivation, extrinsic motivation or personal happiness and satisfaction) are associated with a significantly higher reported likelihood of participating in a peer network. The latter tests hypotheses 1(a), 1(b), 2(a), 2(b), 3(a), and 3(b), that the effect of these three framings depend on pre-existing respondent characteristics, specifically their level of work engagement, intrinsic motivation and extrinsic motivation. The p-values reported are unadjusted, but we also perform the Benjamini-Hochberg FDR adjustment.[14] We report the adjusted p-values in the text.

[14]We apply the adjustment to all the coefficients to which we attach a causal interpretation in our two preferred specifications, 2 and 4.

Table 4: Regression results (DV = Likelihood of participating in a peer network, 1-5)

	Basic		With interactions	
	(1)	(2)	(3)	(4)
T1: Career value	0.080	0.077*	0.339	0.234
	(0.053)	(0.040)	(0.277)	(0.308)
T2: Public benefit	0.070	0.059	-0.113	-0.032
	(0.057)	(0.057)	(0.253)	(0.295)
T3: Emotional Satisfaction	-0.062**	-0.045	-0.179	-0.184
	(0.024)	(0.038)	(0.302)	(0.406)
Work Engagement		0.139***	0.127***	0.197***
		(0.024)	(0.027)	(0.026)
Intrinsic Motivation		0.202***	0.189*	0.115
		(0.042)	(0.095)	(0.092)
Extrinsic Motivation		0.078	0.163	0.031
		(0.066)	(0.101)	(0.081)
Work Engagement x T1:CV			-0.042	-0.075
			(0.045)	(0.046)
Intrinsic motivation x T1:CV			-0.030	0.044
			(0.107)	(0.136)
Extrinsic motivation x T1:CV			-0.003	0.040
			(0.082)	(0.097)
Work Engagement x T2:PB			0.017	-0.030
			(0.064)	(0.047)
Intrinsic motivation x T2:PB			0.233***	0.183**
			(0.080)	(0.076)
Extrinsic motivation x T2:PB			-0.198**	-0.089
			(0.094)	(0.082)
Work Engagement x T3:ES			-0.068**	-0.118***
			(0.028)	(0.029)
Intrinsic motivation x T3:ES			0.076	0.125
			(0.138)	(0.161)
Extrinsic motivation x T3:ES			0.143**	0.206***
			(0.067)	(0.052)
Num.Obs.	1354	1354	1354	1354
R2	0.066	0.180	0.123	0.185
R2 Adj.	0.044	0.133	0.094	0.131
R2 Within	0.003	0.125	0.064	0.130
R2 Within Adj.	0.001	0.094	0.054	0.092
RMSE	1.04	0.97	1.01	0.97
Std.Errors	by: country	by: country	by: country	by: country
Controls	N	Y	N	Y
FE: country	X	X	X	X

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Standard errors are clustered by country and in parentheses. CV = T1, career value; PB = T2, public benefit; ES = T3, Emotional satisfaction. Controls not reported are: Number of organizational challenges identified by the respondent, dummy variable for female respondents, years of experience in the public sector, role type, size of organization, sector of organization, level of seniority, a dummy for highly networked respondents, a dummy for respondents working in service delivery, Route X orientation of the organization, belief that Government is the place for those who want to do public service and the number of staff managed by the respondent.

Hypothesis 1 was that framing the gains to peer network participation in terms of career value to participants (treatment 1, extrinsic motivation) would be sufficient to increase reported likelihood of participation relative to a neutral control. In specification 2 we find a small positive effect of this treatment at the 10% level (p=0.068), but this does not survive adjustment for multiple hypothesis testing (adjusted p=0.255). Hypotheses 1(a) tested the heterogeneity of this treatment effect according to extrinsic motivation level: perhaps those with higher extrinsic motivation were more responsive to this framing? This does not appear to be the case: we fail to reject the null that the relationship between extrinsic motivation and reported likelihood of participating in a peer network is the same under treatment 1 as under a neutral control. This could be because the framing was insufficiently strong to adequately engage the extrinsic motivation of respondents, though it is also striking that there is no effect of extrinsic motivation on reported likelihood of participation even in the control group (since the coefficient on extrinsic motivation is not significant under any specification). Hypothesis 1(b) was that Treatment 1 would not be responsive to pre-existing variation in intrinsic motivation or work engagement, and this is indeed what we observe: no difference in the relationship between these characteristics and reported likelihood of participation under treatment 1 compared to the neutral control.

Hypothesis 2 was that framing the gains to peer network participation in terms of the benefit to end users (treatment 2, our intrinsic motivation treatment) would increase reported likelihood of participation compared to a neutral control. Neither specification 2 nor 4 provide support for this hypothesis. Hypothesis 2(a) was that such an effect might be stronger for those with higher intrinsic motivation. There is some evidence to this effect in specification 4, with the interaction between treatment 2 and pre-existing intrinsic motivation of respondents positive and significant, (p=0.023), while the uninteracted intrinsic motivation coefficient loses signficance. This suggests that reported likelihood of participating in a peer network is increasing in intrinsic motivation under treatment 2, but there is no significant relationship under the neutral control framing. However, adjusting this p-value for multiple hypotheses renders this relationship just insignificant at the 10% level (adjusted p=0.114). Further, testing whether the joint coefficient on the public benefit treatment dummy and the interaction between this treatment and intrinsic motivation scores is significantly different to zero shows that even at the upper ranges of this scale, there is no effect significantly different to that in the neutral control. So, though more intrinsically motivated respondents respond more to the intrinsic motivation framing, they are still not responsive to treatment overall. Again, this might suggest that the treatment is weak. Hypothesis 2(b) was that treatment 2 would not be responsive to variation in pre-existing levels of extrinsic motivation or work engagement; again, this is what we observe, though it is interesting to note that the coefficient on extrinsic motivation interacted with treatment 2 is negative, and significant in our specification without additional controls, suggesting that more extrinsically motivated respondents are, if anything, reporting lower likelihood of participation in peer networks than less extrinsically motivated respondents, relative to a neutral control.

Hypothesis 3 was that framing the benefits of participation in a peer network in terms of emotional satisfaction and personal happiness (treatment 3) would be enough to affect reported likelihood of participation relative to a neutral control. We find no such effect in any of our specifications. However, we find some evidence in support of hypotheses 3(a) and 3(b). Specifically, while the uninteracted Utrecht Work Engagement Score is positively associated

with expressed likelihood of joining a peer support network in all specifications, the interaction between the UWES score and treatment 3 is negative and significant (p=0.000, adjusted p=0.004). This indicates that the relationship between the UWES score and reported likelihood of participating in a peer network is different under this treatment than under the neutral framing. At the same time, the interaction between extrinsic motivation and and treatment 3 is positive and significant (p=0.000, adjusted p=0.004), so respondents with higher extrinsic motivation report higher likelihood of participation than those with lower extrinsic motivation under this framing than under the neutral control. Evaluating the overall effect of treatment requires interpreting the uninteracted treatment effect together with the coefficients on these interactions at different levels. We find that for respondents who have both high work engagement and low extrinsic motivation, there is an overall negative treatment effect (that is, these respondents are report lower likelihood of participation than under a neutral control).[15] For respondents with low work engagement and high extrinsic motivation, there is a positive point estimate for treatment 3, but it is not significant (p=0.157).[16] Because engagement and extrinsic motivation are generally positively correlated, these cells (high engagement and low extrinsic motivation, or low engagement and high extrinsic motivation) are relatively small in the data, which suggests that using different framings to make specific forms of motivation more salient has a significant effect on a relatively small number of public servants.

Overall, these results suggest that none of the framings have a strong universal effect on public servants' likelihood of participating in peer networks, at least in our sample. And while there is evidence that making some motivations more salient elicits a different response along certain dimensions of pre-existing heterogeneity, the overall effect of treatment is signficiant for a minority of respondents.

We ran two pre-specified robustness tests, the results of which are both reported in the Appendix. In the first, we replace our standard outcome variable (the self-reported likelihood of joining a peer network on a scale of 1-5) with a dummy variable for reporting that they were likely or very likely to participate in a peer network. This robustness check is a subset of our primary specification that uses less of the information collected by our survey, since it collapses a 5 point scale into a binary variable (scores 1-3 are coded as 0; scores 4 and 5 are coded as 1), and estimates effects only around the threshold thereby created. It looks at the extent to which treatment causes respondents to move between the 1-3 and 4-5 groups, ignoring any movements within either group.

Using this restricted variable, our results (table 5 in the Appendix) confirm the effect of the interaction between intrinsic motivation and the 'public benefit' framing, but finds no significant relationship in either the interaction between extrinsic motivation and the personal happiness framing or work engagement and the personal happiness framing. This suggests that the effect of this interaction is concentrated within the 1-3 or 4-5 groups.

In our pre-analysis plan, we also specified the use of an ordered probit regression as a robustness test. This may be an appropriate specification if we care specifically about moving from, say 4 to 5, on our scale compared to the average treatment effect. On the other hand it is much more complicated to implement and draw inferences from, and

[15] The estimates for those with work engagement levels of 6 or 7 and extrinsic motivation levels at the lower end of the scale are -0.49 and -0.61, respectively, with p values of 0.10 and 0.035 respectively. We tested whether the linear restriction implied by these values of the scales was significantly different from zero.

[16] The estimate for those with work engagement levels of 1 and extrinsic motivation levels at the top end of the scale found in the data is 0.515, with a p value of 0.157.

simulations suggest that using a simple OLS performs similarly while making fewer assumptions about data structure and data generating processes.[17] If you care primarily about average treatment effects, as we do here, OLS may be preferred. Ordered probit approaches takes a latent variable approach and assume the ordinal scale used as the dependent variable expresses an underlying variable (in this case some continuous 'likelihood of engaging' in the peer network; the use of the latent variable is what complicates interpretation of coefficients). It also assumes that effects of treatment are the same between any two levels of the latent variable. We can test the appropriateness of this assumption using a Brant test (Brant 1990); in our case, we strongly reject the proportional odds assumption for our preferred specification, including interaction effects, which suggests the ordered probit estimation procedure is *not* appropriate in our case (see table 6). Nevertheless, we implement it, with results reported as table 7. It confirms the direction of the coefficients from our core specification, but generates larger standard errors so some of our results lose significance. Given the rejection of the Brant test, and our general interest in the average treatment effect, we put little weight on this result.

The next section provides a discussion of our results.

9 Discussion

Our results have implications on what motivates public servants to engage in peer networks; and about how heterogeneity of public servant motivations might affect how they respond to the same stimulus. Before discussing further, it is worth reiterating the limitation of using a survey experiment to investigate these questions: while our results can say something about the relative attractiveness of different framings to different public sector workers, the responses we collect will be coloured by the prior experience of peer networks some of our sample will have had. While we can use randomisation and controls to limit this effect, it should be borne in mind. And secondly, in practice peer networks are costly to join and participate in, while indicating enthusiasm in our survey is costless. As such, while the relative attractiveness of different framings to different public servants may still be informative, we should be cautious in interpreting enthusiasm to mean guaranteed or even likely participation.

Many organizations provide some form of peer networking as part of their career development or social infrastructure: leadership 'away days', conferences and training courses all entail at least some component of peer networking.[18] These networks, whether or not they achieve positive outcomes for either public servants themselves or the organizations they work for, are costly. Yet most are voluntary, or organized with the peer-network aspect incidental or secondary to some other purpose. We know very little about why public servants engage in peer networking, and which public servants join them, or show interest in doing so. Whether or not they serve their intended purposes, understanding why public servants are motivated to join them matters.

The results of our experiment suggest that framing the gains from participation in terms of benefits to end-users is heightens the contrast in likely participation between more and less intrinsically motivated public servants, though

[17] See, for example: https://declaredesign.org/blog/estimating-average-treatment-effects-with-ordered-probit-is-it-worth-it.html

[18] See, for example, the UK's Civil Service Leadership Academy, which explicitly includes peer networking as one of its selling points: https://www.gov.uk/guidance/training-for-leaders-in-the-civil-service-and-public-sector

the treatment effect of this framing is insignificant for all respondents. This may seem circular, if intrinsic motivation is about alignment to the mission of an organization, then we should expect that respondents with high intrinsic motivation are more likely to respond to a framing that focuses on the end users of the organization's work than those with low intrinsic motivation. However, our measure of intrinsic motivation, from the Work Preference Inventory, does not use any questions about organizational mission or mission match. Rather, it measures some preference for enjoyment and challenge in the work. The knowledge that public servants whose motivation comes from the job itself are more likely to respond to framings around effectiveness in meeting the organizational mission than those with low intrinsic motivation, is, while unsurprising, useful to confirm.

However, the more striking result is that the personal happiness and emotional satisfaction framing, though having no average effect overall, elicit differential responses depending on the level of work engagement and extrinsic motivation of respondents. Respondents with high levels of engagement and low levels of extrinsic motivation report lower likelihood of participation under this treatment. The estimate of the effect on those with the opposite characteristics (that is, low engagement and high extrinsic motivation) is positive, though statistically insignificant (p=0.157). This suggests that this framing elicits a differential response depending on the type and state of the public servants exposed to it. As we discussed in the introduction, since happiness and satisfaction at work are relatively less explored as a motivating factor for public servants, this opens up the possibility that effortful action in some domains might be motivated by a focus on employee well-being and satisfaction, at least among those workers with higher extrinsic motivation. At the same time, it may have the opposite effect on those who are already highly engaged in their work. The same approach may have very different effects on different public servants.

Though the ways in which peer networks are framed do not have any overall effect on likelihood of engagement, our finding suggest that some framings can be more or less attractive to specific subsets of public servants. In the UK, the Peer Learning groups coordinated by the University of Oxford's Government Outcomes Lab are mainly framed around intrinsic motivation and public benefit, focusing on knowledge sharing, achieving better value in public service and so on. Similarly, the Involve Public Sector Peer Learning network is framed around knowledge sharing and intrinsic motivation. The Andrews and Manning Peer Learning Guide also focuses on knowledge sharing and performance at work for the most part (Andrews and Manning 2016), while the UK Government's Leadership College (explicitly framed in part as a peer network) is framed very clearly around promotion and career prospects. None of these framings had significant overall treatment effects in our experiment relative to a neutral control. None of the major public sector peer learning networks we uncovered is framed explicitly around well-being, happiness or satisfaction at work, which reduced expressed likelihood of participation among the most engaged respondents in our experiment. The framings used may heighten the differential likelihood of participation of public servants of different types, or who are currently more or less engaged.

More extrinsically motivated respondents are more likely to respond positively to a framing that focuses on personal happiness and emotional satisfaction than less extrinsically motivated public servants, but are not responsive to a framing focused on career benefits, even compared to less extrinsically motivated respondents. One possible explanation here is simply that the peer networks they are being asked about is not judged to be suitable for supporting career

benefits—though such a perception would be at odds with received wisdom.[19] It may also be that that emotional satisfaction and personal happiness is a potentially important motivating factor for more extrinsically motivated public servants. If this is the case, it provides a margin on which budget-constrained public services can motivate workers without recourse to already overstretched budgets.

References

Acemoglu, D., Kremer, M., and Mian, A. (2007). Incentives in Markets, Firms, and Governments. *Journal of law, economics, & organization*, 24(2):273–306.

Ali, A. J., Fuenzalida, J., Gómez, M., and Williams, M. J. (2021). Four lenses on people management in the public sector: An evidence review and synthesis. *Oxford review of economic policy*, 37(2):335–366.

Amabile, T. M., Barsade, S. G., Mueller, J. S., and Staw, B. M. (2005). Affect and Creativity at Work. *Administrative science quarterly*, 50(3):367–403.

Amabile, T. M., Hill, K. G., Hennessey, B. A., and Tighe, E. M. (1994). The Work Preference Inventory. *Journal of personality and social psychology*, 66(5):950–967.

Andrews, M. and Manning, N. (2015a). A Study of Peer Learning in the Public Sector: Experience, experiments and ideas to guide future practice. Technical report, Effective Institutions Platform.

Andrews, M. and Manning, N. (2015b). Mapping Peer Learning Initiatives in Public Sector Reforms in Development — Harvard Kennedy School.

Andrews, M. and Manning, N. (2016). A Guide to Peer-to-Peer Learning: how to make peer-to-peer support and learning effective in the public sector? Technical report, Effective Instituitions Platform.

Arrow, K. J. (1974). *The limits of organization*. Fels lectures on public policy analysis (New York, N.Y.). Norton, New York ; London.

Ashraf, N., Bandiera, O., Davenport, E., and Lee, S. S. (2020). Losing prosociality in the quest for talent? sorting, selection, and productivity in the delivery of public services. *The American economic review*, 110(5):1355–1394.

Ashraf, N., Bandiera, O., and Jack, B. K. (2014). No margin, no mission? A field experiment on incentives for public service delivery. *Journal of Public Economics*, 120:1–17.

Banuri, S., Keefer, P., and de Walque, D. (2018). Love the Job... or the Patient? : Task vs. Mission-Based Motivations in Health Care.

Belle, N. and Cantarelli, P. (2015). Monetary Incentives, Motivation, and Job Effort in the Public Sector. *Review of public personnel administration*, 35(2):99–123.

[19] See, for example, this piece from Harvard Business Review: https://hbr.org/2016/05/learn-to-love-networking

Bénabou, R. and Tirole, J. (2003). Intrinsic and Extrinsic Motivation. *The Review of Economic Studies*, 70(3):489–520.

Besley, T. and Ghatak, M. (2005). Competition and incentives with motivated agents.

Brant, R. (1990). Assessing Proportionality in the Proportional Odds Model for Ordinal Logistic Regression. *Biometrics*, 46(4):1171–1178.

Deci, E. L. (1975). *Intrinsic motivation.* Perspectives in social psychology ; v 1. Plenum Press, New York.

Deci, E. L. and Ryan, R. M. (1985). *Intrinsic motivation and self-determination in human behavior.* Perspectives in social psychology. Plenum, New York.

Dewatripont, M., Jewitt, I., and Tirole, J. (1999). The Economics of Career Concerns, Part II: Application to Missions and Accountability of Government Agencies. *Review of Economic Studies*, 66(1):199–217.

Diehl, E., Rieger, S., Letzel, S., Schablon, A., Nienhaus, A., Pinzon, L. C. E., and Dietz, P. (2021). The relationship between workload and burnout among nurses: The buffering role of personal, social and organisational resources. *PloS one*, 16(1):e0245798.

Eldor, L. (2018). Public service sector: The compassionate workplace - The effect of compassion and stress on employee engagement, burnout, and performance. *Journal of public administration research and theory*, 28(1):86–103.

Georgellis, Y., Iossa, E., and Tabvuma, V. (2011). Crowding Out Intrinsic Motivation in the Public Sector. *Journal of public administration research and theory*, 21(3):473–493.

Gertler, P. and Vermeersch, C. (2012). Using Performance Incentives to Improve Health Outcomes.

Gottlieb, J. and Bauer, J. (2022). The Municipal Workforce Through the Pandemic: Where Are We Now ? Technical report, Natiopnal League of Cities, Washington, DC.

Grant, A. M. and Hofmann, D. A. (2011). It's Not All About Me: Motivating Hand Hygiene Among Health Care Professionals by Focusing on Patients. *Psychological science*, 22(12):1494–1499.

Honig, D. (2022). Managing for Motivation as Public Performance Improvement Strategy in Education & Far Beyond.

Hosie, P., Willemyns, M., and Sevastos, P. (2012). The impact of happiness on managers' contextual and task performance. *Asia Pacific journal of human resources*, 50(3):268–287.

Iyer, L. and Mani, A. (2012). Traveling agents: political change and bureaucratic turnover in India.(Author abstract)(Report). *Review of Economics and Statistics*, 94(3):723.

Jackson, C. K. and Bruegmann, E. (2009). Teaching students and teaching each other: the importance of peer learning for teachers. *American economic journal. Applied economics*, 1(4):85–108.

Kamenica, E. (2012). Behavioral Economics and Psychology of Incentives. *Annual review of economics*, 4(1):427–452.

Karachiwalla, N. and Park, A. (2017). Promotion incentives in the public sector: Evidence from Chinese schools. *Journal of public economics*, 146:109–128.

Leaver, C. (2009). Bureaucratic Minimal Squawk Behavior : Theory and Evidence from Regulatory Agencies. *American Economic Review*, 99(3):572–607.

Lepper, M. R., Greene, D., and Nisbett, R. E. (1973). Undermining children's intrinsic interest with extrinsic reward: A test of the "overjustification" hypothesis. *Journal of personality and social psychology*, 28(1):129–137.

Linos, E., Ruffini, K., and Wilcoxen, S. (2021). Reducing Burnout and Resignations among Frontline Workers: A Field Experiment. *Journal of public administration research and theory*, 32(3):473–488.

Liss-Levinson, R. (2022). Continued Impact of COVID-19 on Public Sector Employee Job and Financial Outlook , Satisfaction, and Retention. Technical Report March, MissionSquare Research Institute.

McGregor, D. (1966). The human side of enterprise. *Classics of organization theory*, 2(1):6–15.

Muralidharan, K. and Sundararaman, V. (2011). Teacher Performance Pay: Experimental Evidence from India. *The Journal of Political Economy*, 119(1):39–77.

Nikolova, M. and Cnossen, F. (2020). What makes work meaningful and why economists should care about it. *Labour economics*, 65:101847.

Oswald, A. J., Proto, E., and Sgroi, D. (2015). Happiness and Productivity. *Journal of labor economics*, 33(4):789–822.

Park, S. M. and Word, J. (2012). Driven to Service: Intrinsic and Extrinsic Motivation for Public and Nonprofit Managers. *Public personnel management*, 41(4):705–734.

Perrow, C. (1979). *Complex organizations : a critical essay*. Scott, Foresman, Glenview, Ill, 2d ed. edition.

Phillips, C. (2020). Relationships between workload perception, burnout, and intent to leave among medical-surgical nurses. *International journal of evidence-based healthcare*, 18(2):265–273.

Plimmer, G. and Cantal, C. (2016). Workplace Dynamics in New Zealand Public Services. Technical report, Centre for Labour, Employment and Work, Victoria University of Wellington, Wellington.

Reizer, A., Brender-Ilan, Y., and Sheaffer, Z. (2019). Employee motivation, emotions, and performance: a longitudinal diary study. *Journal Of Managerial Psychology*, 34(6):pp415–428.

Roberts, E. P., Mills, D. A., and Stein, A. F. (2018). Dentists' Perceptions of Their Peer Learning Experiences in Dental School and Effects on Practice. *Journal of dental education*, 82(11):1185–1193.

Robinson, G. F. W. B., Switzer, G. E., Cohen, E. D., Primack, B. A., Kapoor, W. N., Seltzer, D. L., and Rubio, D. M. (2014). Shortening the Work Preference Inventory for Use with Physician Scientists: WPI[U+2010]10. *Clinical and translational science*, 7(4):324–328.

Schaufeli, W. B., Salanova, M., González-Romá, V., and Bakker, A. B. (2002). The Measurement of Engagement and Burnout: A Two Sample Confirmatory Factor Analytic Approach. *Journal of happiness studies*, 3(1):71.

Schaufeli, W. B., Shimazu, A., Hakanen, J., Salanova, M., and De Witte, H. (2019). An Ultra-Short Measure for Work Engagement. *European journal of psychological assessment : official organ of the European Association of Psychological Assessment*, 35(4):577–591.

Vroom, V. H. (1964). *Work and motivation*. Wiley, New York.

Wells, M. B. (2015). Predicting preschool teacher retention and turnover in newly hired Head Start teachers across the first half of the school year. *Early childhood research quarterly*, 30(Pt. A):152–159.

Willis-Shattuck, M., Bidwell, P., Thomas, S., Wyness, L., Blaauw, D., and Ditlopo, P. (2008). Motivation and retention of health workers in developing countries: A systematic review. *BMC health services research*, 8(1):247–247.

Wood, P. (2019). Rethinking time in the workload debate. *Management in education*, 33(2):86–90.

Appendix A Statement from co-authors

To whom it may concern:

We confirm that Ranil Dissanayake was the primary author of our article, What motivates public sector workers to form peer networks? Evidence from a survey experiment. Ranil developed the idea of the experiment, led the design of the survey vignette around which the paper is built, conducted all the quantitative analysis and was the primary author of the article itself.

Dan Honig
Associate Professor of Public Policy
University College London
dan.honig@ucl.ac.uk

Sarah Thompson
Strategic Growth Senior Manager
Evidence Action
sarah.thompson@evidenceaction.org

Elizabeth Linos
Emma Bloomberg Associate Professor of Public Policy and Management
Harvard Kennedy School
elizabeth_linos@hks.harvard.edu

Appendix B Survey instrument

Practitioner Circles Pre-Study Final Survey

Start of Block: Consent and Screening

HIRB Consent Script This survey is part of a research project being led by Dr. Dan Honig (dhonig@jhu.edu), a professor at University College London and a fellow of the Johns Hopkins School of Advanced International Studies Foreign Policy Institute. The study has received ethical approval from the Johns Hopkins Homewood IRB (hirb@jhu.edu), who can be contacted if you have any concerns regarding this study. This survey is intended for public sector workers. If you are not currently employed in the public sector, please do not complete this survey. What follows is very short survey on workplace motivation will begin on the next screen. This survey includes approximately 20 questions (including demographic questions), and will take approximately 18 minutes to complete. The survey will ask for no personally identifiable information – your answers will be entirely anonymous. Those who successfully complete the survey will earn 2.70 GBP for this HIT. If you have any concerns about this study, please feel free to contact Dr. Honig or the Johns Hopkins Homewood IRB (hirb@jhu.edu) which has approved this study. **If you have taken this survey already, please do not take it again.** Your subsequent responses will be rejected and you will not be compensated if repeat test-taking is identified. Rejection will also occur if the information you provide is nonresponsive (does not answer the questions asked). By clicking "yes" below you confirm that you are currently a public sector employee. You also confirm that you are at least 18 years old. Many thanks for considering participation in this study.

Consent Do you agree to participate in a study for full-time public sector workers?

- ○ Yes
- ○ No

Eligibility Are you a full-time public sector worker over the age of 18 who speaks English?

- ○ Yes
- ○ No

End of Block: Consent and Screening

Start of Block: Prolific ID

Prolific ID Autofill What is your Prolific ID?
Please note that this response should auto-fill with the correct ID.

_____

End of Block: Prolific ID

Start of Block: Attn_Check

Attn_Check From NYT: When Bodega, a streetwear shop in the Back Bay neighborhood of Boston, released a hyped, limited-edition New Balance 997S sneaker in 2019, the entire stock sold out online in under 10 minutes. There was one problem, though: About 60 percent of Bodega's sales went to shoppers gaming the system with bots, timesaving automation software used to speed through checkout. The bots had claimed hundreds of pairs of New Balances for a single customer; many other shoppers failed to secure just one. This question is an attention check. To show you're not a bot choose answer number two below. Shoppers armed with specialized sneaker bots can deplete a store's inventory in the time it takes a person to select a size and fill in shipping and payment information.

Based on the passage above, please choose an option for shoe stores to mitigate the effects of automated bots on their sales.

- (1) Nothing, let bots buy the shoes--at least they're still making sales
- (2) Make all sales in-person only
- (3) Hire more sophisticated IT teams that can build defenses against automated bots
- (4) Team up with other shoe stores to streamline the purchase process
- (5) Require all purchasers to enter certain information that would make it harder for bots to navigate
- (6) Give up, the bots will always win

End of Block: Attn_Check

Start of Block: Treatment & Control

CONTROL
We are designing a new initiative to support public sector workers through peer-to-peer engagement, organized in small groups. The goal of these groups is to support you as a person and as a public sector employee.

Engaging in these groups would require a small but regular commitment of time and effort. By joining one of these groups, you will be supported personally and professionally by the other members of the group. **Based on past experience and feedback, we expect these groups to have a number of benefits for participants.**

If offered the opportunity to join a group like this, how likely would you be to join?

- 1 - Very Unlikely (1)
- 2 - Unlikely (2)
- 3 - Moderately Likely (3)
- 4 - Likely (4)
- 5 - Very Likely (5)

USEFUL FOR CAREER We are designing a new initiative to support public sector workers through peer-to-peer engagement, organized in small groups. The goal of these groups is to support you as a person and as a public sector employee.

Engaging in these groups would require a small but regular commitment of time and effort. By joining one of these groups, you will be supported personally and professionally by the other members of the group. **Based on past experience and feedback, we expect these groups to bring career benefits to participants, by sharing information and advice that help you advance in your career.**

If offered the opportunity to join a group like this, how likely would you be to join?

- ○ 1 - Very Unlikely (1)
- ○ 2 - Unlikely (2)
- ○ 3 - Moderately Likely (3)
- ○ 4 - Likely (4)
- ○ 5 - Very Likely (5)

PUBLIC BENEFIT We are designing a new initiative to support public sector workers through peer-to-peer engagement, organized in small groups. The goal of these groups is to support you as a person and as a public sector employee.

Engaging in these groups would require a small but regular commitment of time and effort. By joining one of these groups, you will be supported personally and professionally by the other members of the group. **Based on past experience and feedback, we expect these groups to help participants better support the people your agency serves, by sharing information and advice that helps your organization improve the way it supports these end-users.**

If offered the opportunity to join a group like this, how likely would you be to join?

- 1 - Very Unlikely (1)
- 2 - Unlikely (2)
- 3 - Moderately Likely (3)
- 4 - Likely (4)
- 5 - Very Likely (5)

PERSONAL HAPPINESS We are designing a new initiative to support public sector workers through peer-to-peer engagement, organized in small groups. The goal of these groups is to support you as a person and as a public sector employee.

Engaging in these groups would require a small but regular commitment of time and effort. By joining one of these groups, you will be supported personally and professionally by the other members of the group. **Based on past experience and feedback, we expect these groups to provide emotional support and help you become a happier, more balanced, more satisfied person.**

If offered the opportunity to join a group like this, how likely would you be to join?

- ○ 1 - Very Unlikely (1)
- ○ 2 - Unlikely (2)
- ○ 3 - Moderately Likely (3)
- ○ 4 - Likely (4)
- ○ 5 - Very Likely (5)

End of Block: Treatment & Control

Start of Block: Program Context

Program Context We call these peer groups Practitioner Circles, and they may have a number of benefits:

They may help you advance in your career.
They may help you support the end users of your organization better.
They may help you become a happier, more balanced, more satisfied person.

For the rest of the survey, when we refer to Practitioner Circles, these small peer-to-peer networks are what we mean.

End of Block: Program Context

Start of Block: Org Culture & PSM

Transition Text Please consider the following questions about your experience in your current organization.

Trajectory1 On the whole, do you think things are getting better or getting worse at your organization?

- ○ Much better
- ○ Somewhat better
- ○ About the same
- ○ Somewhat worse
- ○ Much worse

Trajectory1_Why Why do you think that?

_____

_____

_____

_____

_____

Trajectory3 Does your work have the potential to impact the organization's trajectory?

- ○ Definitely yes
- ○ Probably yes
- ○ Might or might not
- ○ Probably not
- ○ Definitely not

Oversight In your opinion, what is the balance in your organization between providing freedom for staff to pursue organizational goals (however they are defined) as they see fit on the one hand, and tightly controlling staff through targets, monitoring, oversight and incentive schemes?

	0 10 20 30 40 50 60 70 80 90 100
0 (full freedom for staff) to 100 (tight oversight by management) ()	

Org Culture How much do you agree with the following statement?

	1 - Strongly Disagree (1)	2 - Disagree (2)	3 - Neutral (3)	4 - Agree (4)	5 - Strongly Agree (5)
Government is the best place for those who want to do public service. (6)	○	○	○	○	○

Learning Network Think about the following groups of people in your life. How frequently...

	Immediate colleagues	Supervisor	Broader members of organization	Other public sector workers

...do you learn from this group? (1)	▼ 1 - Never (1 ... 5 - Always (5)	▼ 1 - Never (1 ... 5 - Always (5)	▼ 1 - Never (1 ... 5 - Always (5)	▼ 1 - Never (1 ... 5 - Always (5)
...do you share learning with this group? (2)	▼ 1 - Never (1 ... 5 - Always (5)	▼ 1 - Never (1 ... 5 - Always (5)	▼ 1 - Never (1 ... 5 - Always (5)	▼ 1 - Never (1 ... 5 - Always (5)
...do you lean on this group for support? (3)	▼ 1 - Never (1 ... 5 - Always (5)	▼ 1 - Never (1 ... 5 - Always (5)	▼ 1 - Never (1 ... 5 - Always (5)	▼ 1 - Never (1 ... 5 - Always (5)
...is this group influenced by how you go about doing your work? (4)	▼ 1 - Never (1 ... 5 - Always (5)	▼ 1 - Never (1 ... 5 - Always (5)	▼ 1 - Never (1 ... 5 - Always (5)	▼ 1 - Never (1 ... 5 - Always (5)
...do you discuss decisions you need to make at work with this group? (5)	▼ 1 - Never (1 ... 5 - Always (5)	▼ 1 - Never (1 ... 5 - Always (5)	▼ 1 - Never (1 ... 5 - Always (5)	▼ 1 - Never (1 ... 5 - Always (5)

End of Block: Org Culture & PSM

Start of Block: Engagement

Intent to Leave Which of the following statements best describes your future plans?

- ○ I want to leave my organization as soon as possible. (1)
- ○ I want to leave my organization within the next 12 months. (2)
- ○ I want to stay working for my organization for at least the next year. (3)
- ○ I want to stay working for my organization for at least the next three years. (4)

UWES-3 How often do you agree with the following statements?

	0 - Never (1)	1 - Very Rarely (2)	2 - Rarely (3)	3 - Neutral (4)	4 - Frequently (5)	5 - Very Frequently (6)	6 - Always (7)
At my work, I feel bursting with energy. (4)	○	○	○	○	○	○	○
I am enthusiastic about my job. (5)	○	○	○	○	○	○	○
I am immersed in my work. (6)	○	○	○	○	○	○	○

WPI How often are the following statements true for you?

	0 - Never or almost never true for me (1)	1 (2)	2 (3)	3 - Always or almost always true for me (4)
I am strongly motivated by the recognition I can earn from other people.	○	○	○	○
I want other people to find out how good I really can be at my work.	○	○	○	○
To me, success means doing better than other people.	○	○	○	○
I am keenly aware of the promotion goals I have for myself.	○	○	○	○
I am keenly aware of the income goals I have for myself.	○	○	○	○
I enjoy tackling problems that are completely new to me.	○	○	○	○
I enjoy trying to solve complex problems.	○	○	○	○
The more difficult the problem, the more I enjoy trying to solve it.	○	○	○	○
What matters most to me is enjoying what I do.	○	○	○	○

It is important for me to be able to do what I most enjoy.	○	○	○	○

End of Block: Engagement

Start of Block: Goal of PCs

Program Context For the following questions, think again about Practitioner Circles, these small virtual group meetings of peers in public sector roles.

Goal of Connecting If you were to join a Practitioner Circle group, would each of the following be a goal of joining or not be a goal of joining?

	Would be a goal	Would not be a goal
Learning from this group (1)	○	○
Sharing learning with this group (2)	○	○
Leaning on this group for support (3)	○	○
Influencing this group by how you go about doing your work (4)	○	○
Discuss decisions you need to make at work with this group (5)	○	○
Other (Select "Would not be a goal" if this doesn't apply) (6)	○	○

PC Focus Given the opportunity to participate in one of these virtual groups called Practitioner Circles, what would you like it designed to focus on?

	Most Focus On	Some Focus On	No Focus On
Career benefits (1)	○	○	○
Benefits to users of public services (7)	○	○	○
Happiness and wellbeing of participants (8)	○	○	○
Decision making and the decisions you face in your job (12)	○	○	○
Other (Select "No Focus On" if this doesn't apply) (5)	○	○	○

Interest Intensity How enthusiastic would you be about joining a Practitioner Circle?

	0 10 20 30 40 50 60 70 80 90 100
0 (not enthusiastic) to 100 (extremely enthusiastic) ()	

End of Block: Goal of PCs

Start of Block: Matching

Matching1
There are many ways groups of participants in Practitioner Circles can be constructed. Think about the **characteristics of fellow participants in the group that would make the group most useful to you.**

Which type of characteristics of your group's fellow participants are most important to you?

	1 - Not important at all (1)	2 - Unimportant (2)	3 - Moderately important (3)	4 - Important (4)	5 - Very important (5)
The organization they work for (1)	○	○	○	○	○
The country they are from (2)	○	○	○	○	○
The practice area / sector they work in (3)	○	○	○	○	○
Their rank in their organization (4)	○	○	○	○	○
The native language they speak (5)	○	○	○	○	○
Their gender (6)	○	○	○	○	○
The religion they practice (7)	○	○	○	○	○
Number of years of work experience they have (9)	○	○	○	○	○
Their race or ethnicity (10)	○	○	○	○	○

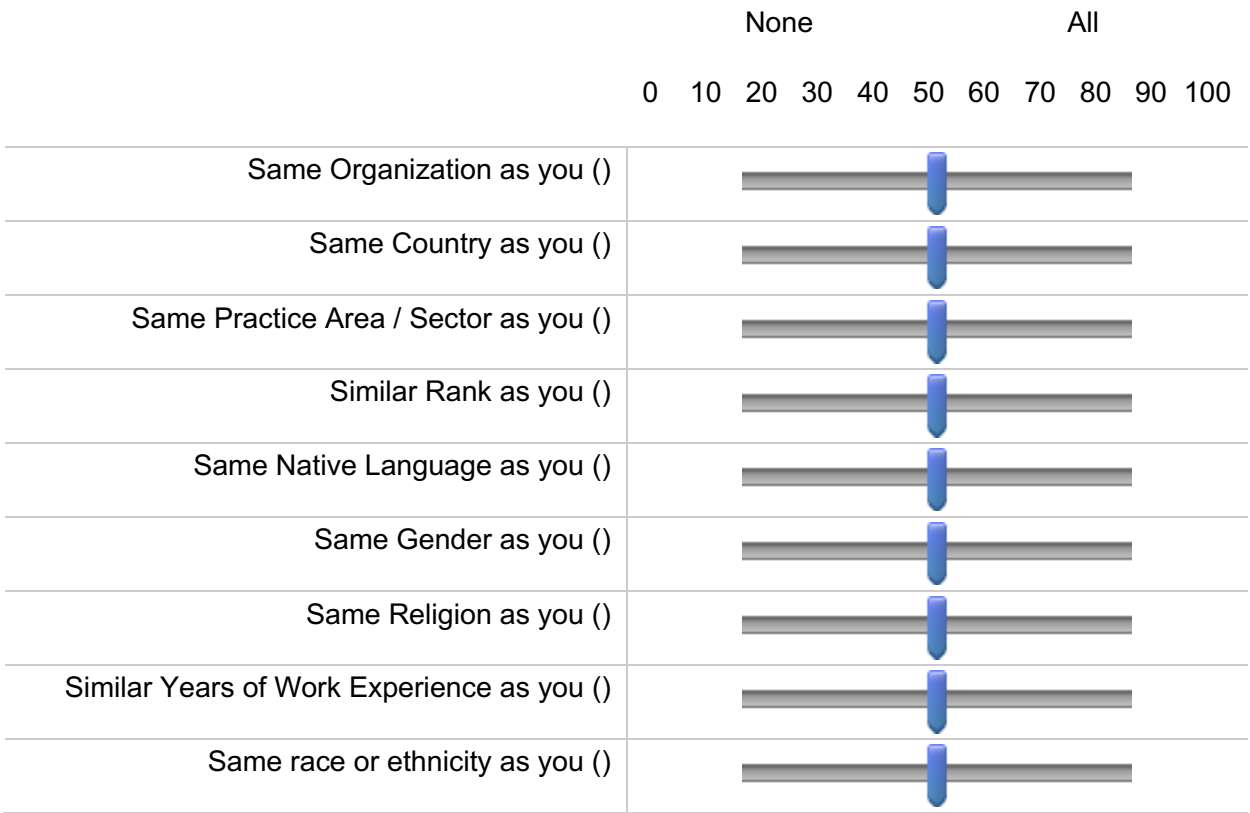
Page Break

Display This Question:

If There are many ways groups of participants in Practitioner Circles can be constructed. Think abou... = 4 - Important

Or There are many ways groups of participants in Practitioner Circles can be constructed. Think abou... = 5 - Very important

Matching2 What percentage of the group members would you like to have the same characteristic as you?



Matching_Other What other participant characteristics or factors matter to you?

_____

End of Block: Matching

Start of Block: Functioning Mechanism

Program Context Think again about these small virtual group meetings of people in other public sector roles. For the following questions, imagine you have a choice about how your group interacts.

Convening Which of these ideas might be most interesting to you to convene around? (Check all that apply.)

- [ ] Navigating my organization
- [ ] Navigating my personal career path
- [ ] Learning new tangible things
- [ ] Emotional support
- [ ] Professional networking
- [ ] ⊗ None of the abov

Mechanisms1 How much do the following aspects of the program matter to whether or not you will participate?

	1 - Very unimportant (1)	2 - Unimportant (2)	3 - Moderately important (3)	4 - Important (4)	5 - Very important (5)
Timing and Frequency of meeting (1)	○	○	○	○	○
Size of group (2)	○	○	○	○	○

Display This Question:

If How much do the following aspects of the program matter to whether or not you will participate? = Timing and Frequency of meeting [4 - Important]

Or How much do the following aspects of the program matter to whether or not you will participate? = Timing and Frequency of meeting [5 - Very important]

Timing of Meeting How likely would you be to participate in a program that:

	1 - Very unlikely (1)	2 - Unlikely (2)	3 - Moderately Likely (3)	4 - Likely (4)	5 - Very Likely (5)
Meets bi-weekly (1)	○	○	○	○	○
Meets monthly (2)	○	○	○	○	○
Meets during work hours (3)	○	○	○	○	○
Online message boards instead of live meetings (4)	○	○	○	○	○

Facilitation How likely would you be to participate in a program with the following facilitation styles for leading group discussion?

	1 - Very unlikely (1)	2 - Unlikely (2)	3 - Moderately Likely (3)	4 - Likely (4)	5 - Very Likely (5)
No facilitator or agenda -- peers lead conversations (1)	○	○	○	○	○
No facilitator but suggested agenda provided (2)	○	○	○	○	○
Peers take on structured facilitator role after a brief training (3)	○	○	○	○	○
Facilitated through a pre-recorded video (4)	○	○	○	○	○
External facilitator leads live group discussion (5)	○	○	○	○	○

Display This Question:

If How much do the following aspects of the program matter to whether or not you will participate? = Size of group [4 - Important]

Or How much do the following aspects of the program matter to whether or not you will participate? = Size of group [5 - Very important]

*

Group Size How many people would you like to have in a group?

_____

End of Block: Functioning Mechanism

Start of Block: Respondent Characteristics

Transition Text Now please consider the following questions about you and your current role.

Gender What is your gender?

- ○ Male
- ○ Female
- ○ Other _____

Race What is your race?

_____

Ethnicity What is your ethnicity?

_____

Country In what country do you work?

_____

Org_Level What is your level in your organization?

- ○ Leader of organization (1)
- ○ Leader of department (2)
- ○ Some Management Responsibility (3)
- ○ No Management Responsibility (4)

People_Mgmt How many people do you manage directly?

_____

ServiceDelivery Does your job require direct service delivery to citizens?

- ○ Yes (1)
- ○ No (2)

Org_Size What is the size of your organization?

- ○ 0-49 (1)
- ○ 50-99 (2)
- ○ 100-199 (3)
- ○ 200-499 (4)
- ○ 500+ (5)

Org_Sector In what sector do you work?

- ○ Healthcare (1)
- ○ Transportation (2)
- ○ Housing (3)
- ○ Education (4)
- ○ Agriculture (5)
- ○ Diplomacy (6)
- ○ International Aid (7)
- ○ Finance (8)
- ○ Law (9)
- ○ Infrastructure (10)
- ○ Environment (11)
- ○ Recreation/Tourism (12)
- ○ Economic development (13)
- ○ Utilities (14)
- ○ Fire fighters/military/police (15)
- ○ Immigration (17)
- ○ Other (16) _____

Role Function What is the primary function of your role?

- ○ Finance (2)
- ○ Talent/HR (3)
- ○ Strategy (4)
- ○ IT/Data (5)
- ○ Communications (6)
- ○ Analytics (7)
- ○ Program Management (8)
- ○ Service Delivery (9)
- ○ Research/Analysis (10)

Years Experience How many years of full time work experience do you have?

_____

Work Challenges Please indicate whether each of the following is a challenge you're currently facing at work.

	Current challenge (1)	Not a challenge (2)
I do not have clear objectives (2)	○	○
I have to spend my own money (3)	○	○
The organization finds it difficult to take risks (4)	○	○
My line management changes regularly (5)	○	○
Senior leadership changes regularly (6)	○	○
My supervisor is not good (7)	○	○
I am struggling with a new supervisor (8)	○	○
Prioritization of initiatives or projects are not clear (9)	○	○
Working across departments/agencies/organizations (10)	○	○
Changing leadership priorities or organizational goals (11)	○	○
My organization isn't helping the people it is meant to reach (16)	○	○
My organization isn't focused on the right goals (17)	○	○
Lack of resources (12)	○	○
Lack of time (13)	○	○
Building/leading a team (14)	○	○

Other topic (15)	○	○

End of Block: Respondent Characteristics

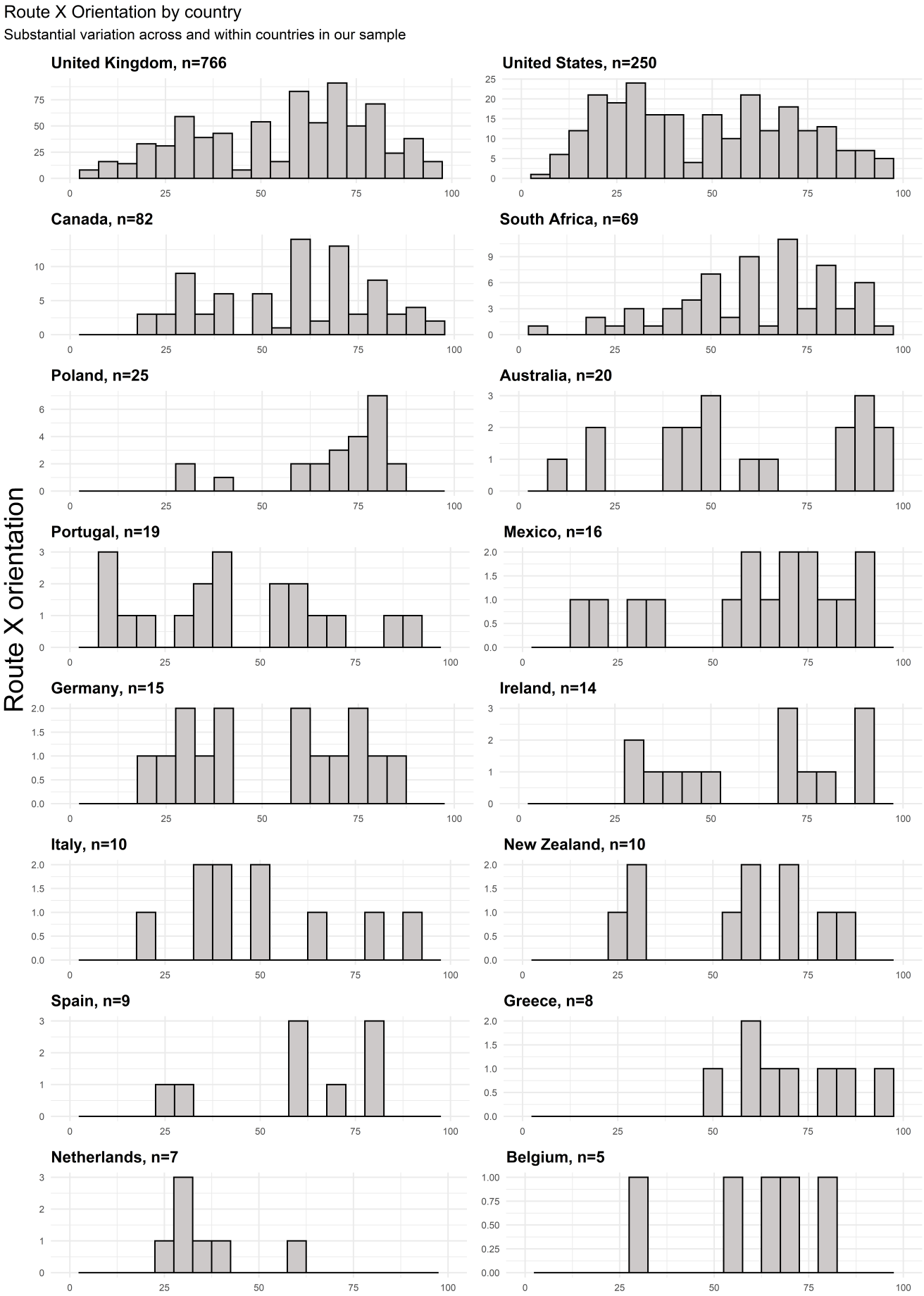
Start of Block: Thank you

Thank you Thank you for taking this survey. Click below to be redirected to Prolific.

End of Block: Thank you

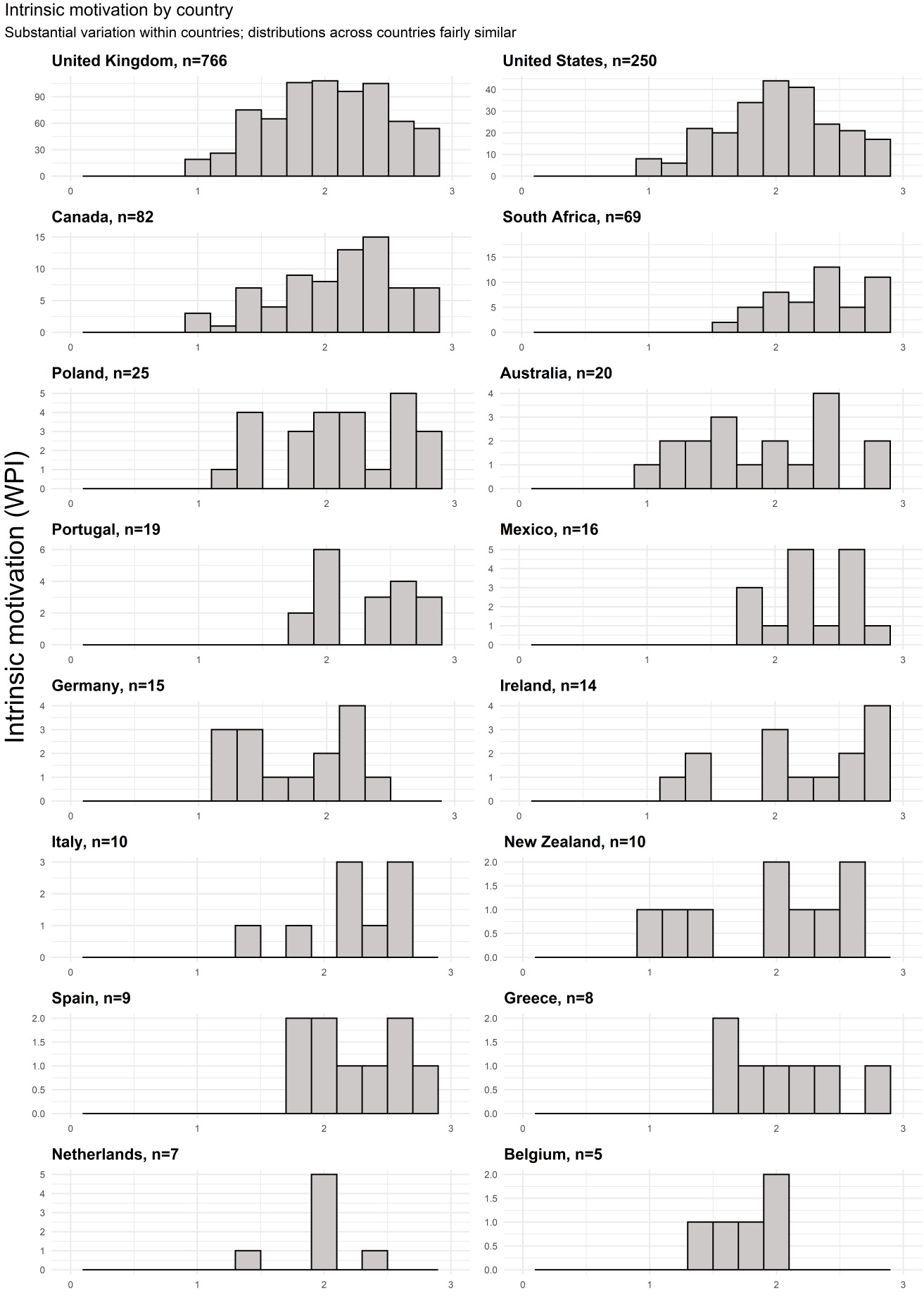
Appendix C Respondent characteristics by country

Figure 3: Route X Orientation



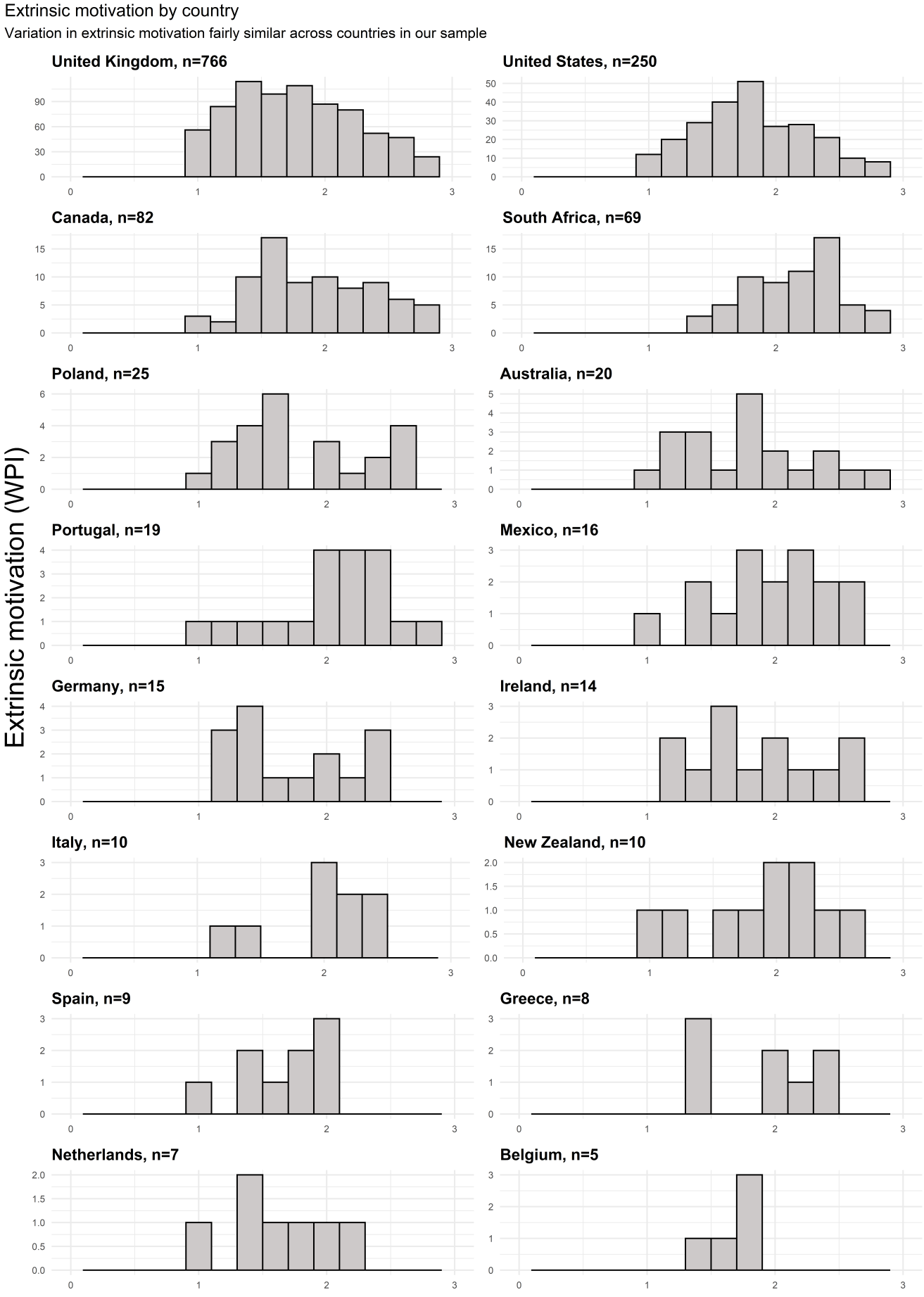
Bin width = 5. Only countries with 5 or more observations reported. Route X orientation is increasing in the tightness of control imposed by the organization on workers.

Figure 4: Intrinisic Motivation



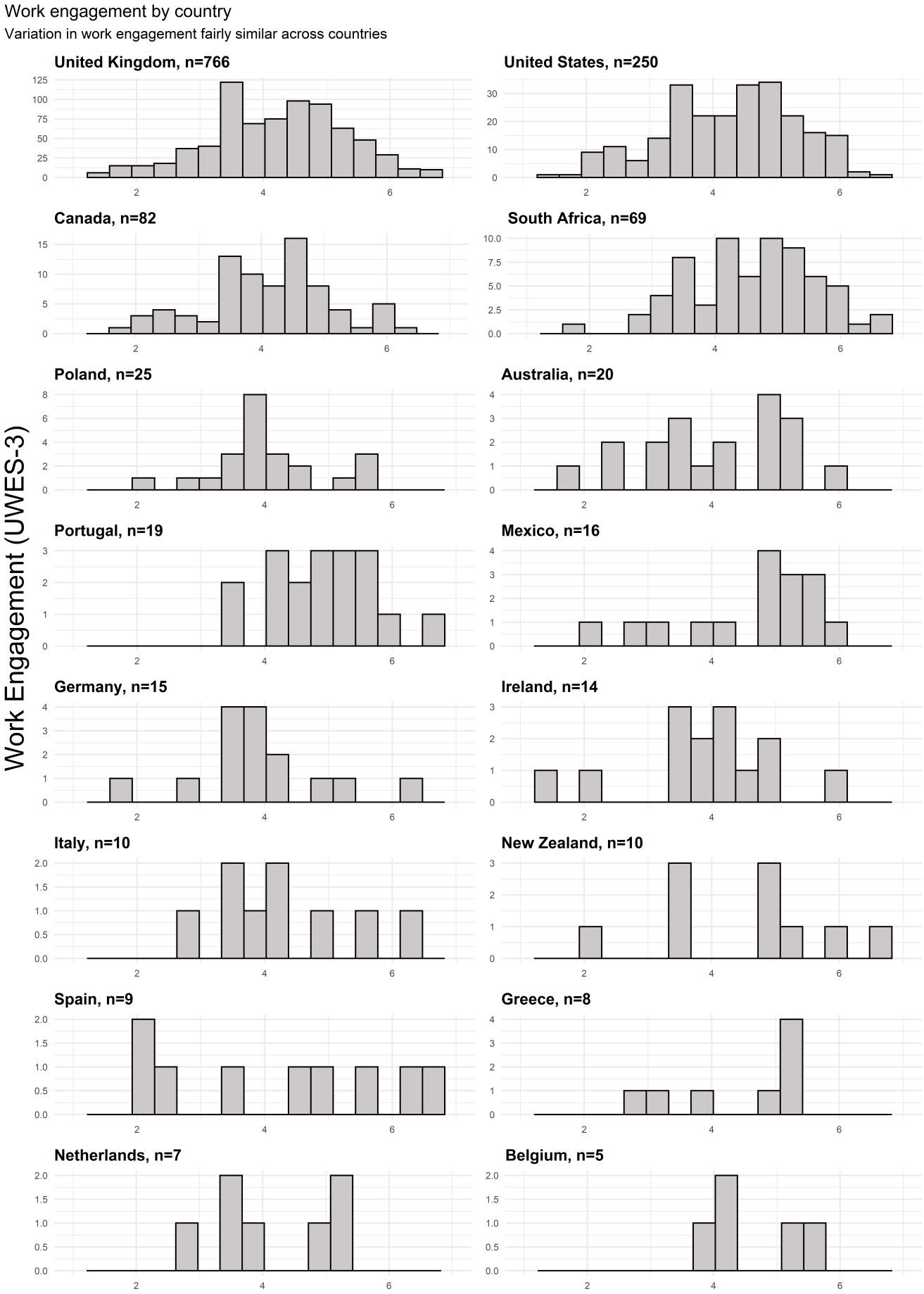
Bin width = 0.2. Only countries with 5 or more observations reported. The WPI-Intrinsic score is increasing in intrinsic motivation.

Figure 5: Extrinisic Motivation



Bin width = 0.2. Only countries with 5 or more observations reported. The WPI-Extrinsic score is increasing in intrinsic motivation.

Figure 6: Work Engagement



Bin width = 0.35. Only countries with 5 or more observations reported. The UWEs-3 score is increasing in work engagement.

Appendix D Robustness Tests

Table 5: Robustness test 1 (DV = Dummy variable for 'likely' or 'very likely' to participate in a peer network)

	Basic		With interactions	
	(1)	(2)	(3)	(4)
T1: Career value	0.033	0.027	0.070	0.002
	(0.030)	(0.024)	(0.192)	(0.198)
T2: Public benefit	0.037	0.030	-0.087	-0.054
	(0.042)	(0.044)	(0.207)	(0.226)
T3: Emotional satisfaction	-0.017	-0.015	0.074	0.061
	(0.018)	(0.017)	(0.242)	(0.277)
Work Engagement		0.052***	0.054**	0.077***
		(0.011)	(0.020)	(0.018)
Intrinsic Motivation		0.086***	0.040	0.017
		(0.023)	(0.049)	(0.050)
Extrinsic Motivation		0.013	0.088*	0.029
		(0.039)	(0.047)	(0.042)
Work Engagement x T1:CV			-0.027	-0.038*
			(0.022)	(0.022)
Intrinsic motivation x T1:CV			0.078	0.115*
			(0.047)	(0.058)
Extrinsic motivation x T1:CV			-0.046	-0.031
			(0.070)	(0.075)
Work Engagement x T2:PB			0.009	-0.007
			(0.035)	(0.027)
Intrinsic motivation x T2:PB			0.126***	0.098**
			(0.043)	(0.037)
Extrinsic motivation x T2:PB			-0.095*	-0.049
			(0.055)	(0.053)
Work Engagement x T3:ES			-0.037	-0.050
			(0.036)	(0.033)
Intrinsic motivation x T3:ES			0.055	0.061
			(0.071)	(0.086)
Extrinsic motivation x T3:ES			-0.023	0.005
			(0.051)	(0.046)
Num.Obs.	1354	1354	1354	1354
R2	0.072	0.166	0.114	0.169
R2 Adj.	0.051	0.117	0.085	0.115
R2 Within	0.002	0.103	0.047	0.106
R2 Within Adj.	0.000	0.070	0.036	0.068
RMSE	0.47	0.45	0.46	0.45
Std.Errors	by: country	by: country	by: country	by: country
Controls	N	Y	N	Y
FE: country	X	X	X	X

* p < 0.1, ** p < 0.05, *** p < 0.01. Standard errors are clustered by country and in parentheses. CV = T1, career value; PB = T2, public benefit; ES = T3, Emotional satisfaction. Controls not reported are: Number of organizational challenges identified by the respondent, dummy variable for female respondents, years of experience in the public sector, role type, size of organization, sector of organization, level of seniority, a dummy for highly networked respondents, a dummy for respondents working in service delivery, Route X orientation of the organization, belief that Government is the place for those who want to do public service and the number of staff managed by the respondent.

Table 6: Brant (1990) Test for Proportionality

Brant test results:			
	X2	df	alpha
Omnibus	425.14	249	0

When alpha <0.05, we reject the proportional odds assumption

Table 7: Robustness test 2 (Ordered probit, DV = Likelihood of participating in a peer network, 1-5)

	Basic		With interactions	
	(R1)	(R2)	(R3)	(R4)
T1: Career value	0.083	0.083	0.335	0.221
	(0.057)	(0.048)	(0.316)	(0.357)
T2: Public benefit	0.075	0.069	-0.158	-0.080
	(0.061)	(0.063)	(0.277)	(0.327)
T3: Emotional Satisfaction	-0.058	-0.045	-0.161	-0.178
	(0.025)	(0.043)	(0.335)	(0.447)
Work Engagement		0.154***	0.137**	0.219***
		(0.023)	(0.030)	(0.028)
Intrinsic Motivation		0.221***	0.191	0.119
		(0.050)	(0.102)	(0.100)
Extrinsic Motivation		0.090	0.175	0.038
		(0.072)	(0.108)	(0.091)
Work Engagement × T1:CV			-0.046	-0.083
			(0.049)	(0.050)
Intrinsic Motivation × T1:CV			-0.011	0.071
			(0.111)	(0.146)
Extrinsic Motivation × T1:CV			-0.010	0.037
			(0.082)	(0.099)
Work Engagement × T2:PB			0.019	-0.032
			(0.067)	(0.052)
Intrinsic Motivation × T2:PB			0.253	0.204
			(0.090)	(0.090)
Extrinsic Motivation × T2:PB			-0.196	-0.078
			(0.104)	(0.092)
Work Engagement × T3:ES			-0.075	-0.132*
			(0.031)	(0.032)
Intrinsic Motivation × T3:ES			0.083	0.140
			(0.151)	(0.175)
Extrinsic Motivation × T3:ES			0.141	0.220
			(0.077)	(0.064)
Num.Obs.	1354	1354	1354	1354
RMSE	3.14	3.14	3.14	3.14
Std.Errors	by: country			

Note: * p < 0.1, ** p < 0.05, *** p < 0.01. Standard errors in parentheses. CV = T1, career value; PB = T2, public benefit; ES = T3, Emotional satisfaction. Controls not reported are: Number of organizational challenges identified by the respondent, dummy variable for female respondents, years of experience in the public sector, role type, size of organization, sector of organization, level of seniority, a dummy for highly networked respondents, a dummy for respondents working in service delivery, Route X orientation of the organization, belief that Government is the place for those who want to do public service and the number of staff managed by the respondent.

5

Conclusions and Policy Implications

At the beginning of this thesis, I set out the central questions I sought to investigate: how does the behaviour and decision-making of public servants affect the organizations they work for and their functioning? And how do these effects vary according to the ability, motivation and incentives facing these public servants? Here, I take stock of how far towards this ambition I have been able to travel, the limitations of my work, and the directions it suggests for both further academic inquiry and public policy.

I have sought, throughout, to apply a quantitative lens to my efforts. Doing so has helped the work make a novel contribution; but also set some limitations. In my first paper, I used a survey experiment making use of randomly-varied vignettes, of the type increasingly popular with large organizations (Neumark 2018) to ask how exposure to political preferences orthogonal to organizational mandate can affect the decision-making of civil servants, and the extent to which this effect varies with seniority in the organization. This study generated causal estimates of the effect of exposure to politician preference, even when the organizational mandate should be unaffected by it; and shows senior decision-makers are insulated from such an effect. This has implications for how civil services might optimally structure decision-making when Ministerial engagement is important. In this case, we consider decision-making given a specific objective, but the general mechanism may also be important when the objectives themselves are wrong (Elston and Bevan 2021). In my second paper, I look at another kind of 'bureaucratic' adaption: the use of peer

review and quality assurance processes to—nominally—improve the choices made in a large Government department. I use statistical tests of manipulation around a threshold to document substantial avoidance behaviour by the public servants affected by this system. Equally strikingly, available metrics of quality show only weak effects on quality of projects around this threshold, whether they might be driven by selection (avoidance behaviour) or the causal effect of review. Some organizational adaptations aimed to address performance have limited or negative effects: public bureaucracies, constrained by budget, time, attention, staffing and so on should be aware of this possibility and invest in learning about their systems. While systems of audit or scrutiny are almost impossible to make completely immune to gaming, understanding the behavioural response they engender is of first-order importance in their design—or they may fall prey to the 'hollow ritual' critique of performance auditing (Kells 2011).

In my final paper, I look at the commonly-used practice of organizational peer networks, and ask whether different ways of framing the benefits of these networks affect the likelihood public servants report of joining them. Using another experimentally varied survey vignette, I find framings making career benefits, benefits to end users and the emotional satisfaction and personal happiness of public sector workers themselves have no overall effect. However, we find some evidence of differential responses to these treatments according to pre-existing heterogeneity among respondents. These results suggest that some public servants do respond differentially to framings that appeal to their motivation or engagement levels, even when there is no overall effect. The extensive literature on the optimal selection of civil servants provides great insights into how to recruit more effective or motivated public servants (Ashraf et al. 2020; Banuri and Keefer 2015b; Leaver et al. 2021), but conditional on already being selected into public service, at the margin, motivation to engage in costly effort may not follow the same patterns.

Collectively, my three papers make a novel contribution to the literature examining how the behaviour and decision-making of public servants affects public sector performance. An extensive quantitative literature on management in the public sector finds that managerial quality and characteristics often have significant and important effects on organizational function (Bloom et al. 2015a,b; Rasul et al. 2017; Rasul and Rogger 2018); but there

has been less quantitative exploration of hierarchy and decision-making specifically. My first paper contributes to this literature, and points to further academic directions; for example, do the effects of hierarchy vary when senior officials are political appointees? Does it depends on tenure or proximity to promotion or other career milestones? What precisely is the mechanism through which this effect occurs: ability, probity or willingness to challenge, or organizational culture? And does the finding of 'speaking truth to power' extend to field studies (as opposed to survey experiments), and when the objectives, rather than the means of achieving them, are contested (Elston and Bevan 2021)? Lab and field studies (and indeed lab-in-the-field studies) may shed light on different aspects of these questions.

Similarly, there is a strong public administration literature which investigates the effect of audit and assurance on organizational function, some of which suggests that it is often a rather blunt and limited tool (Kells 2011; Morin 2001), and a similar literature which finds relatively small effects of peer review in academic settings Crijns et al. 2021; Higgs and Gelman 2021; and a literature which documents gaming of certain aspects of audit compliance in the UK (Elston and Zhang 2022). My paper contributes to this body of work by providing a specific, empirically demonstrated example of a negative behavioural response to the existence of an assurance process in the UK, and quantitative estimates of the effect of review on project quality, though it cannot distinguish between the behavioural response the causal effect of review. Again, more work is suggested by this study, particularly to unpack the precise mechanism by which quality effects occur.

My third paper took a different approach, looking bottom-up at what motivates public sector workers, from a large, 28 country sample, to participate in peer networks, again using a causally identified survey vignette experiment. This study built on two extensive literatures: one on the effects of peer networks themselves (Jackson and Bruegmann 2009; Linos et al. 2021) and one on the motivation and selection of public sector workers (Banuri and Keefer 2015b,a; Banuri et al. 2018; Ashraf et al. 2020; Deci 1975). This paper builds on these literatures by filling a knowledge gap around why public sector employees might join or participate in peer networks, and also explores the differential effects of using framings that make different kinds of motivation more salient according to pre-existing

heterogeneity among respondents. This result also suggests directions for further research, particularly in investigating whether these effects are replicated in a field study.

Across these papers, I have sought to demonstrate that the behaviour and decision-making of public servants responds to the organizational and personal incentives, motivations and circumstances they face; and that these can matter for how the organization functions.

That said, my research has certain limitations, some imposed by the difficult circumstances under which my work was conducted. My original thesis plan was to conduct a field experiment to obtain causal estimates of a novel organizational deliberation structure on policy decisions, but the onset of the Coronavirus pandemic and the rapid reprioritization of UK civil service activities put this work on (to date) indefinite hold. Only one of my final papers investigates actual (rather than reported or inferred) organizational function. Though it still provides new information for scholars to build on, in the future, I aim to undertake field experiments both in the UK and abroad to further investigate this relationship between organizational technologies and public sector worker behavioural responses, and their effect on decision-making and performance. Secondly, my methodological approach has been to focus on specific empirical questions, with a view—as far as possible—to causal identification. A complementary approach could use the rich qualitative material available, in particular for my first two papers, to supplement and extend the findings of my empirical analysis. For example, if the true reasoning behind the use of quality assurance in a public bureaucracy is 'theatre' designed to satisfy public scrutiny, it may explain some of the effects I find in my second paper. In order to keep the research manageable and focused, I explicitly took the decision to take at face value the stated objective of the system. Further work could look for empirical ways to causally test these alternative theses from the literature.

The policy implications of my work are rather direct, and have already led to engagement with UK public sector bodies. My second paper, on the quality assurance system, has been the focus of intensive discussion between the researchers and the department studied, and I have been invited to help them determine an alternative and improved evaluatory function, specifically focusing on the development of new, better metrics of project quality. Beyond this, the findings of my first paper paint a dark picture with respect to recent

trends in the UK. If progress in the hierarchy selects civil servants better able to resist and argue with their political masters, or indeed has some causal effect on their ability to do so, either directly or via organizational culture, the recent trend of politicians sacking senior civil servants or selecting ones perceived as more pliable (see for example, the recent sacking of Tom Scholar[1], or the appointment of Simon Case as the Cabinet Secretary[2]) is worrying. If the quality of public sector decision-making depends in part on the good functioning of its organizational safeguards, including the use of hierarchical decision-making, then probity in appointments is of central importance.

My third paper suggests that public sector organizations can better understand and engage with the different workers they employ; not just at the rarefied heights of leadership but in the rank and file. Collectively, with my other two papers, it suggests that by understanding the specific interaction between organizational choices and the public servants who make up those organizations can provide avenues for better performance, or indeed to avoid wasteful effort.

Throughout this thesis, I have sought to understand how public bureaucracies function, and how the organizational choices they make can help or hinder their ability to discharge their mandate. These questions are of first order importance for public welfare: regulation, the provision of public services and sovereign transactions all depend to a greater or lesser extent on them. And yet, there remains much more to be understood. I hope that my work here has shed some light on them, and suggested new ways to continue doing so.

[1] see this coverage in The Guardian, for example

[2] https://www.ft.com/content/670d1f85-5173-44dc-aedb-170c6b0f0713

Bibliography

Acemoglu, D., Kremer, M., and Mian, A. (2007). Incentives in Markets, Firms, and Governments. *Journal of law, economics, & organization*, 24(2):273–306.

Ali, A. J., Fuenzalida, J., Gómez, M., and Williams, M. J. (2021). Four lenses on people management in the public sector: An evidence review and synthesis. *Oxford review of economic policy*, 37(2):335–366.

Amabile, T. M., Hill, K. G., Hennessey, B. A., and Tighe, E. M. (1994). The Work Preference Inventory. *Journal of personality and social psychology*, 66(5):950–967.

Arrow, K. J. (1974). *The limits of organization*. Fels lectures on public policy analysis (New York, N.Y.). Norton, New York ; London.

Ashraf, N., Bandiera, O., Davenport, E., and Lee, S. S. (2020). Losing prosociality in the quest for talent? sorting, selection, and productivity in the delivery of public services. *The American economic review*, 110(5):1355–1394.

Ashraf, N., Bandiera, O., and Jack, B. K. (2014). No margin, no mission? A field experiment on incentives for public service delivery. *Journal of Public Economics*, 120:1–17.

Bandiera, O. and Lee, S. S. (2015). Do-Gooders and Go-Getters : Career Incentives , Selection , and Performance.

Banuri, S. and Keefer, P. (2015a). Pro-social motivation, effort and the call to public service. *European Economic Review*, 83:139–164.

Banuri, S., Keefer, P., and de Walque, D. (2018). Love the Job... or the Patient? : Task vs. Mission-Based Motivations in Health Care.

Banuri, S. and Keefer, P. E. (2015b). Was Weber right ? the effects of pay for ability and pay for performance on pro-social motivation, ability and effort in the public sector.

Belle, N. and Cantarelli, P. (2015). Monetary Incentives, Motivation, and Job Effort in the Public Sector. *Review of public personnel administration*, 35(2):99–123.

Besley, T. (2007). *Principled agents? : the political economy of good government*. Oxford University Press.

Besley, T. and Ghatak, M. (2005). Competition and incentives with motivated agents.

Bloom, N., Eifert, B., Mahajan, A., McKenzie, D., and Roberts, J. (2013). Does managment matter? Evidence from India. *The Quarterly Journal of Economics*, 128(1):1–51.

Bloom, N., Lemos, R., Sadun, R., and Van Reenen, J. (2015a). Does Management Matter in schools? *Economic Journal*, 125(584):647–674.

Bloom, N., Propper, C., Seiler, S., and Van Reenen, J. (2015b). The Impact of Competition on Management Quality: Evidence from Public Hospitals. *The Review of Economic Studies*, 82(2):457–489.

Bloom, N., Sadun, R., and Van Reenen, J. (2016). Management as a Technology? *NBER Working Paper*, Working Pa.

Bloom, N. and Van Reenen, J. (2007). Measuring and Explaining Management Practices across Firms and Countries. *The Quarterly Journal of Economics*, 122(4):1351–1408.

Bloom, N. and Van Reenen, J. (2010). Why Do Management Practices Differ across Firms and Countries? *Journal of Economic Perspectives*, 24(1):203–224.

Buchanan, J. and Tullock, G. (1962). *The Calculus of Consent*. University of Michigan Press, Ann Arbor, MI.

Coase, R. (1937). The Nature of the Firm. *Economica*, 4(16):386–405.

Cohen, M. D., March, J. G., and Olsen, J. P. (1972). A Garbage Can Model of Organizational Choice. 17(1):1–25.

Coviello, D., Spagnolo, G., and Lotti, C. (2021). Rules, bunching and discretion in emergency procurement: Evidence from an earthquake. In Bandiera, O., Bosio, E., and Spagnolo, G., editors, *Procurement in Focus: Rules, Discretion, and Emergencies*, pages 13–22. CEPR Press, London.

Crijns, T. J., Ottenhoff, J. S. E., and Ring, D. (2021). The effect of peer review on the improvement of rejected manuscripts. *Accountability in research*, ahead-of-p(ahead-of-print):1.

Dal Bo, E., Finan, F., and Rossi, M. A. (2013). Strengthening state capabilities: The role of financial incentives in the call to public service. *The Quarterly journal of economics*, 128(3):1169–1218.

Deci, E. L. (1975). *Intrinsic motivation*. Perspectives in social psychology ; v 1. Plenum Press, New York.

Deci, E. L. and Ryan, R. M. (1985). *Intrinsic motivation and self-determination in human behavior*. Perspectives in social psychology. Plenum, New York.

Dewatripont, M., Jewitt, I., and Tirole, J. (1999). The Economics of Career Concerns, Part II: Application to Missions and Accountability of Government Agencies. *Review of Economic Studies*, 66(1):199–217.

Dunsch, F., Evans, D., Eze-Ajoku, E., and Macis, M. (2021). Management, Supervision, and Health Care: A Field Experiment. *NBER Working Paper Series*, page 23749.

Elston, T. and Bevan, G. (2021). Using opportunity costs to counter "one-shot bias" in policy innovation. In Sullivan, H., Dickinson, H., and Henderson, H., editors, *The Palgrave Handbook of the Public Servant*. Palgrave Macmillan.

Elston, T. and Zhang, Y. (2022). Implementing Public Accounts Committee Recommendations: Evidence from the UK Government's 'Progress Reports' since 2012. *Parliamentary Affairs*.

Gawande, A. (2010). *The checklist manifesto : how to get things right*. Profile.

Gertler, P. and Vermeersch, C. (2012). Using Performance Incentives to Improve Health Outcomes.

Gibbons, R. (2003). Team theory, garbage cans and real organizations: some history and prospects of economic research on decision-making in organizations. *Industrial and Corporate Change*, 12(4):753–787.

Gibbons, R. and Henderson, R. (2012). What Do Managers Do? Exploring Persistent Performance Differences among Seemingly Similar Enterprises. In *The Handbook of Organizational Economics*, chapter What Do Ma. Princeton University Press, Princeton, NJ.

Gneezy, U. and Rustichini, A. (2000). Pay Enough or Don't Pay at All. *The Quarterly journal of economics*, 115(3):791–810.

Grant, A. M. and Hofmann, D. A. (2011). It's Not All About Me: Motivating Hand Hygiene Among Health Care Professionals by Focusing on Patients. *Psychological science*, 22(12):1494–1499.

Higgs, M. and Gelman, A. (2021). Research on registered report research. *Nature Human Behaviour*.

Hutchins, E. (1995). *Cognition in the wild*. MIT Press.

Iyer, L. and Mani, A. (2012). Traveling agents: political change and bureaucratic turnover in India.(Author abstract)(Report). *Review of Economics and Statistics*, 94(3):723.

Jackson, C. K. and Bruegmann, E. (2009). Teaching students and teaching each other: the importance of peer learning for teachers. *American economic journal. Applied economics*, 1(4):85–108.

Kahneman, D., Lovallo, D. P., and Sibony, O. (2019). A structured approach to strategic decisions. *MIT Sloan Management Review*, 60(1):1–12.

Kamenica, E. (2012). Behavioral Economics and Psychology of Incentives. *Annual review of economics*, 4(1):427–452.

Karachiwalla, N. and Park, A. (2017). Promotion incentives in the public sector: Evidence from Chinese schools. *Journal of public economics*, 146:109–128.

Kells, S. (2011). The Seven Deadly Sins of Performance Auditing: Implications for Monitoring Public Audit Institutions. *Australian Accounting Review*, 21(4):383–396.

Lazear, E. P. (2000). Performance Pay and Productivity. *The American economic review*, 90(5):1346–1361.

Leaver, C. (2009). Bureaucratic Minimal Squawk Behavior : Theory and Evidence from Regulatory Agencies. *American Economic Review*, 99(3):572–607.

Leaver, C., Serneels, P., Zetlin, A., and Ozier, O. (2021). Recruitment, effort, and retention effects of performance contracts for civil servants: experimental evidence from Rwandan primary schools. *American Economic Review*, 111:2213–2246.

Lepper, M. R., Greene, D., and Nisbett, R. E. (1973). Undermining children's intrinsic interest with extrinsic reward: A test of the "overjustification" hypothesis. *Journal of personality and social psychology*, 28(1):129–137.

Lindblom, C. E. (1959). The Science of " Muddling Through ". *Public Administration Review*, 19(2):79–88.

Linos, E., Ruffini, K., and Wilcoxen, S. (2021). Reducing Burnout and Resignations among Frontline Workers: A Field Experiment. *Journal of public administration research and theory*, 32(3):473–488.

Martinez, E. A., Beaulieu, N., Gibbons, R., Pronovost, P., and Wang, T. (2015). Organizational Culture And Performance. *American Economic Review: Papers & Proceedings*, 3(4):512–527.

Morin, D. (2001). Influence of Value for Money Audit on Public Administrations: Looking Beyond Appearances. *Financial Accountability & Management*, 17(2):99–117.

Muralidharan, K. and Sundararaman, V. (2011). Teacher Performance Pay: Experimental Evidence from India. *The Journal of Political Economy*, 119(1):39–77.

Neumark, D. (2018). Experimental research on labor market discrimination. *Journal of Economic Literature*, 56(3):799–866.

Park, S. M. and Word, J. (2012). Driven to Service: Intrinsic and Extrinsic Motivation for Public and Nonprofit Managers. *Public personnel management*, 41(4):705–734.

Perrow, C. (1979). *Complex organizations : a critical essay*. Scott, Foresman, Glenview, Ill, 2d ed. edition.

Rasul, I. and Rogger, D. (2018). Management of Bureaucrats and Public Service Delivery: Evidence from the Nigerian Civil Service. *The Economic journal (London)*, 128(608):413–446.

Rasul, I., Rogger, D., and Williams, M. J. (2017). Management and Bureaucratic Effectiveness: a Scientific Replication.

Rogger, D. and Somani, R. (2018). Hierarchy and Information.

Simon, H. A. (1983). *Reason in human affairs*. Harry Camp lectures at Stanford University ; 1982. Stanford University Press, Stanford, Calif.

Simon, H. A. (1997). *Administrative Behavior: A study of decision- making processes in administrative organizations*.

Simon, H. A. H. A., Smithburg, D. W., and Thompson, V. A. (1991). *Public administration*. Transaction Publishers.

Weibel, A., Rost, K., and Osterloh, M. (2009). Pay for Performance in the Public Sector—Benefits and (Hidden) Costs. *Journal of public administration research and theory*, 20(2):387–412.

Williamson, O. E. (1999). Public and Private Bureaucracies : A Transaction Cost Economics Perspective. *Journal of Law, Economics, & Organization*, 15(1):306–342.