CENTER FOR GLOBAL DEVELOPMENT

A Higher Bar or an Obstacle Course?

PEER REVIEW AND ORGANIZATIONAL DECISION-MAKING IN AN INTERNATIONAL DEVELOPMENT BUREAUCRACY

Ranil Dissanayake and Euan Ritchie

Abstract

Many public organizations employ technologies of scrutiny such as peer review or quality assurance to improve their performance and decision-making. Such technologies may affect performance and decision-making directly, through scrutiny, and indirectly, through behavioural responses by agents within the organization. We examine one such technology 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 by agents just around the threshold for review designed to avoid scrutiny. Furthermore, there is no evidence that over- or under-spending or fidelity to the planned completion date is better among reviewed projects, despite this observed manipulation—though project quality is more complex than these simple measure capture. Our results suggest that organisations considering such a technology need to investigate both the naïve effect of the technology, and how agents will respond to its existence, setting a new organisational equilibrium.

KEYWORDS
Bureaucracy, Public Organization, Organizational Behavior

JEL CODES
D02, D23, D73

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

Ranil Dissanayake and Euan Ritchie
Center for Global Development

The authors would like to thank Dan Honig, Lee Crawfurd, Julien Labonne, Clare Leaver, Keith Wood, Stefan Dercon and participants at seminars at the Blavatnik School of Government and Center for Global Development for their insightful comments. Any errors that remain are the authors'.

Ranil Dissanayake and Euan Ritchie. 2022. "A Higher Bar or an Obstacle Course? Peer Review and Organizational Decision-Making In an International Development Bureaucracy." CGD Working Paper 615. Washington, DC: Center for Global Development. https://www.cgdev.org/publication/higher-bar-or-obstacle-course-peer-review-and-organizational-decision-making

CENTER FOR GLOBAL DEVELOPMENT

2055 L Street, NW Fifth Floor

Washington, DC 20036

202.416.4000

www.cgdev.org

Contents

Introduction 1

Setting 6

Data 11

Method and empirical strategy 13

Results 15

Discussion 21

Appendix A 24

References 26

List of Figures

Contents

1. Distribution of project sizes before and after the institution of Quality Assurance 15
2. Plot of McCrary density test outputs 16
3. Plot of CJM manipulation test results 17
4. Intrinsic motivation in DFID, 2014–19 23

List of Tables

1. Data and completeness of data scraped from DevTracker 12
2. Results of McCrary density test 16
3. Results of Cattaneo, Jansson and Ma test of manipulation 17
4. Project performance around the quality assurance threshold, 2011–2020 19
5. Project performance around the future quality assurance threshold, pre-2011 20
6. Test for a discontinuity in project performance around the £40 million threshold 21

A1. Results of McCrary density test at £10 million 24
A2. Results of McCrary density test at £20 million 24
A3. Results of Cattaneo, Jansson and Ma test at £10 million 24
A4. Results of Cattaneo, Jansson and Ma test at £20 million 25

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 better, using better 'technologies', broadly defined, to achieve their objectives.

One commonly-adopted technology 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 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 technologies 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 technology 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 technologies may be positive (if the direct benefits from review outweigh costs of distortion) or negative.

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.

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 that this avoidance behaviour was not associated with substantially worse performance as measured by project review scores, over- or under-spending relative to expectations; or timeliness of project completion. 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, Spagnolo, and Lotti 2021). This study extends such findings to a broader segment of public policy than procurement.

This work contributes to an empirical literature that investigates the effect organizational technologies on decision-making and performance by documenting distortionary unintended consequences of the technology, complementing Martinez et. al. (2015) who find that organizational culture is an important determinant of the impact of such technologies, 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.

The paper proceeds as follows: the next section briefly discusses the literature on the use of organizational technologies 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 technology 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 and proposes a framework for assessing the likely net effect of such a system which policymakers may use when considering adoption of review and assurance.

Taking 'technology' literally, in 1960 Herbert Simon discussed the use of computers[4] as aids to management decision-making, suggesting that they might be as revolutionary to office work as heavy machinery was to manual labour (Simon 1960). However, most of the technologies used in decision-making remain rather analogue: 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.

Decision-making technologies have been designed and adopted to aid or guide such human deliberation. Martinez et. al. 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 technology 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 (Martinez et al. 2015). Nicholas Bloom and co-authors argue that management systems and practices constitute a technology for production (Bloom, Sadun, and Van Reenen 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. 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 (Dunsch et al. 2021). 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; Kahneman, Lovallo, and Sibony 2019; Sibony, Lovallo, and Powell 2017).

The role of evaluation and audit as a technology 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 decision-making technology, evaluating 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

4 So exciting were they to him in 1960 that the abstract to his paper described them as 'startling even in a world that takes atomic energy and the prospect of space travel in its stride'. Typically, he was correct, and far ahead of his time: the computing power and access to information have fundamentally changed the calculus of decision-making and even the bounds to rationality that he himself had proposed.

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 technologies 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, Szaszi, and Holcombe 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, Saez, and Sándor 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.[5]

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 technologies 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 2021b), or sub-standard proposals submitted due to pressure to spend allocated funds[6]) 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 the effect of organizational monitoring (Gibbons 2016)). It may apply additional information or cognitive capacity to proposals, a form of cognitive redundancy

5 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

6 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 2021a) , 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.

(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, Dercon, and Gauri 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 (Aczel et al. 2021) is worthwhile. 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 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 technology to support 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 technology and those that are not. In examining how a new technology to aid decision-making for the purpose of improved performance was implemented we document the existence of distortionary unintended consequences of the technology, 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 technology, 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).

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.[7] 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 (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).

7 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

The decision-making technologies used in DFID had substantial public policy and global welfare consequences. In 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, Hughes, and Ritchie 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 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 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. Each of these forms of scrutiny can be conceived as a technology to aid decision-making in the department.

This paper focuses on the third of these technologies: 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.[8] 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 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 sat 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 DFIDs Investment Committee, chaired by the Director-General for Finance and Corporate Performance.

8 Sometimes using formal cost-benefit analysis.

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."[9] 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.[10] Ministers were in turn accountable to Parliament through and the public 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

9 This section draws on personal communication by email with the Head of the Quality Assurance Unit, dated 02/06/21.
10 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.

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 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.[11] 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 the 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.[12] It may have had some function of internal theatre and signaling of values, but certainly, it's 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.[13] 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.

11 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

12 I 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. I searched Hansard for mentions of 'Quality Assurance Unit' or 'QAU' and no relevant matches were returned.

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

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 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.

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.
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.

The table below summarises the data available:

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	4/1/87	4/1/24	2/23/11
Planned completion date	5006	99.4	8/18/09	3/31/45	5/28/15
Actual completion date	1400	27.8	11/2/15	11/26/21	7/26/18
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

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). 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.

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.

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).

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 (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:

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

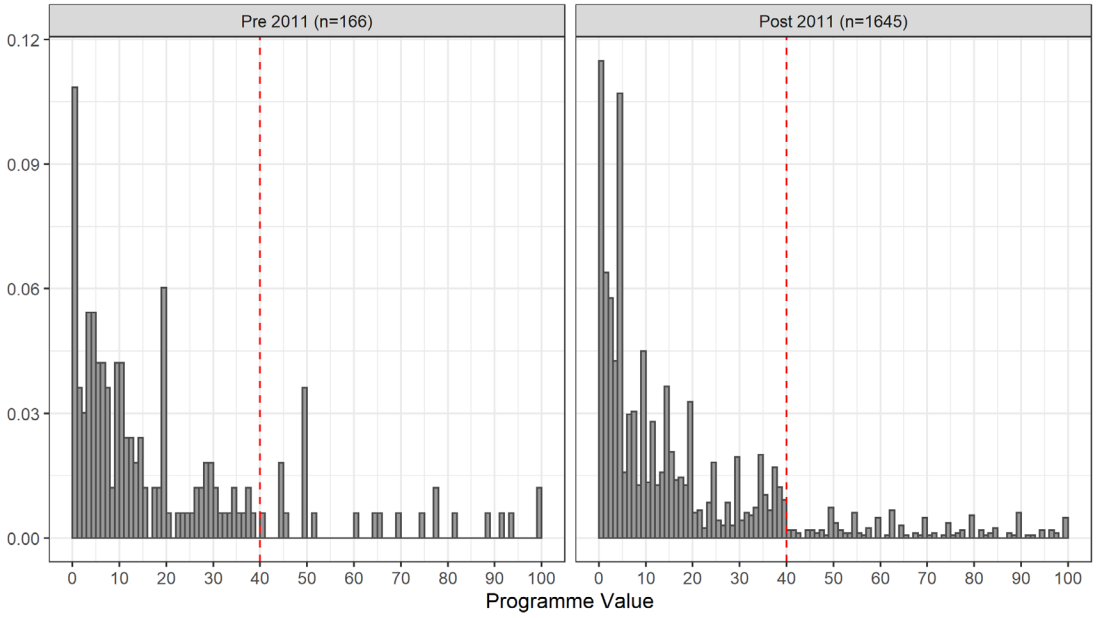
Where QAU is a dummy variable taking the value 0 for a project above the quality assurance threshold, ProgVal is the running variable (programme value of the project), Sector, Recipient and Start Year refer to the location, primary objective and year of actual commencement of each project and εi is an 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. 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.

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. 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.

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 1 shows this clearly.

FIGURE 1. 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 2 presents the plot visualizing the results of the McCrary test. Table 3 and Figure 3 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 > |T|
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 < 0.1, ** p < 0.05, *** p < 0.01

FIGURE 2. Plot of McCrary density test outputs

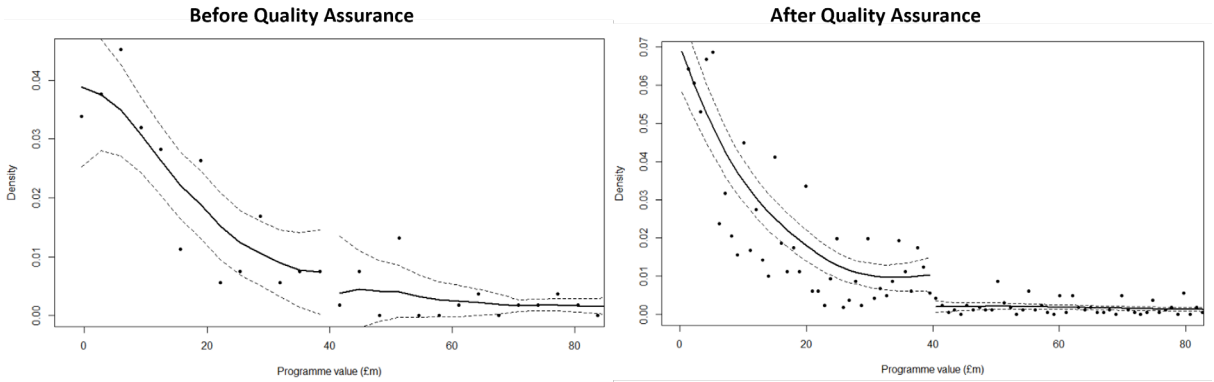
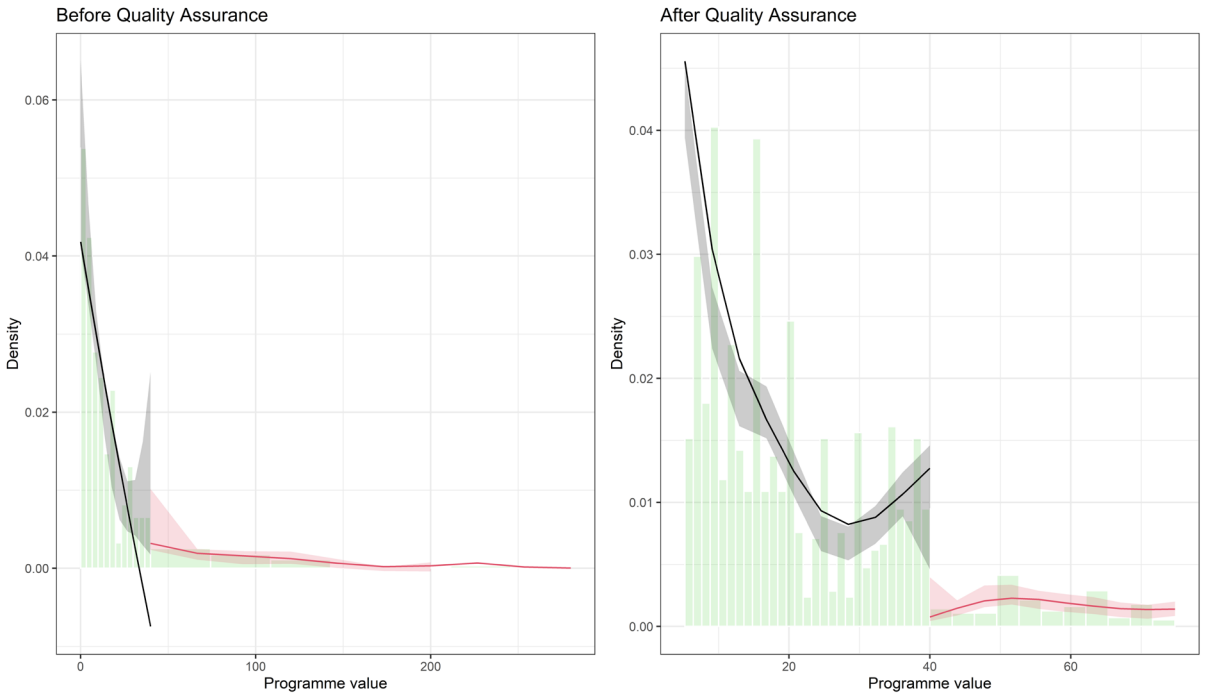


TABLE 3. Results of Cattaneo, Jansson and Ma test of manipulation

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 > |T|	*0.2712*		*0.0189***	

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

FIGURE 3. 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% 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 A.

These results are consistent with widespread manipulation of project sizes beginning once the new technology of peer review and assurance of proposals is introduced. It is not consistent with peer review and assurance picking up sub-optimal projects around the threshold and dramatically increasing their chances of rejection. 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.

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 either because of negative selection 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.

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)	H_0:EQUAL MEANS
			p
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.

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

TABLE 5. Project performance around the future quality assurance threshold, pre-2011

PROJECT IMPLEMENTATION QUALITY INDICATORS, PRE-QAU PERIOD	ALL PROJECTS		T-TEST
	BELOW CUTOFF (OBS)	ABOVE CUTOFF (OBS)	H_0:EQUAL MEANS
			p
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.

Table 5 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. 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% and 40% of original project size, respectively).

The results of the regression discontinuity analysis largely confirm this simple comparison of means. Table 6, overleaf, 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 otherwise there were no significant differences around the threshold. In the years during which peer review was applied, reviewed projects had significantly smaller deviations from their planned completion date (that is, they were significantly less likely to overrun), but otherwise there were no significant differences between reviewed and not reviewed projects. The results do not suggest that either through selection or a causal effect of review that reviewed projects performed systematically better than those not reviewed, even in comparison to the period before peer review was implemented.

TABLE 6. Test for a discontinuity in project performance around the £40 million threshold

	PRE-QAU			POST-QAU		
	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

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++)
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 proivde the number of observations on each side of the discontinuity investigated.
The MSE-Optimal bandwidth was used in each of these regressions.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

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 no evidence of an associated effect on project quality, either through selection effects or the causal effect of review.

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

The latter finding suggests that even the naïve effect should not be taken for granted. The fact that there is no apparent difference on performance metrics between reviewed and not-reviewed projects, 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 no value in this context. This could be the case, for example, if the reviewers suffer from incentive problems, for example, withholding 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 no value to proposals (unlike the effect of 'cognitive redundancy' identified in Hutchins (1995)). 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 unrelated 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]

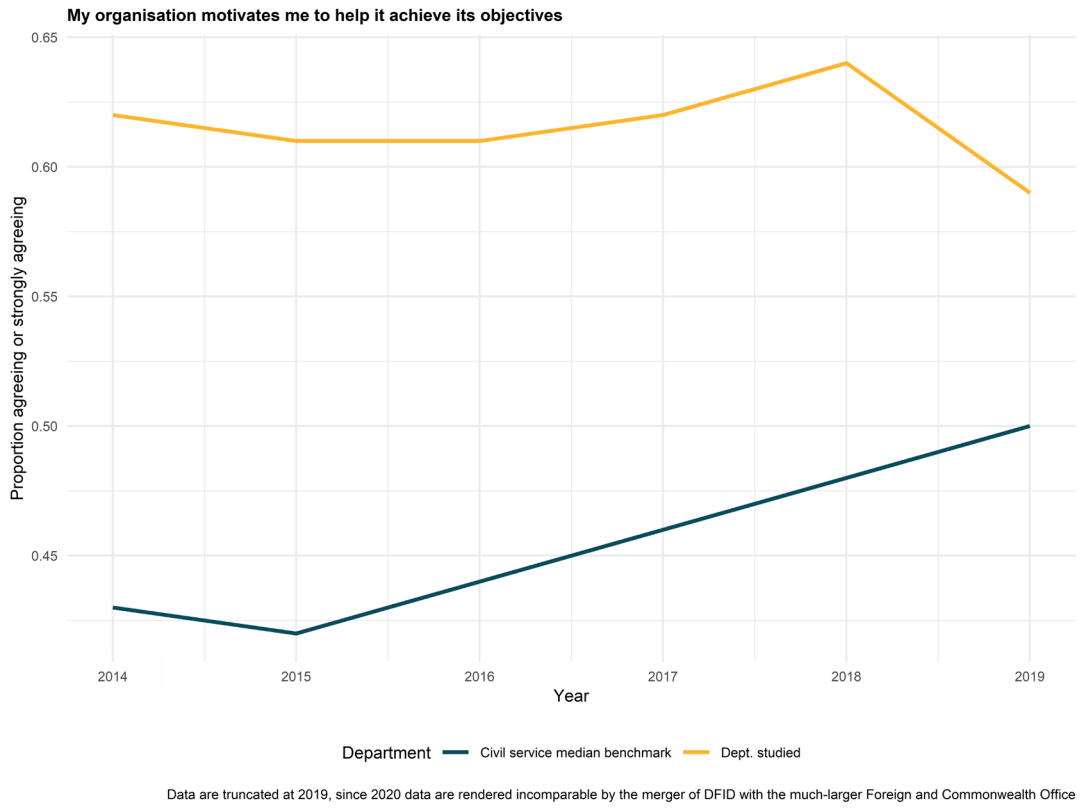
A third possibility is that the metrics tracked are simply unable to adequately assess project quality. Annual Review scores may be endogenous to proposal quality (though it should be noted that they have been used to demonstrate differences between more and less successful programmes, for example in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

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, Schmelz, and Ploner 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 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

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.

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 4. Intrinsic motivation in DFID, 2014–19



Data are truncated at 2019, since 2020 data are rendered incomparable by the merger of DFID with the much-larger Foreign and Commonwealth Office

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 outwardly plausible technologies for improving organizational function needs greater scrutiny. Such technologies may fail on design grounds or because agents within the organization adjust to their presence. The example examined here illustrates both possible problems.

Appendix A

TABLE A1. Results of McCrary density test at £10 million

MCCRARY (2008) TEST RESULTS - £10 MILLION	THETA	SE	Z	BIN SIZE	BANDWIDTH	P > |T|
Before peer review instituted	0.18	0.36	0.50	1.00	13.67	0.61
After peer review instituted	0.01	0.11	0.12	0.98	15.55	0.90

TABLE A2. Results of McCrary density test at £20 million

MCCRARY (2008) TEST RESULTS - £20 MILLION	THETA	SE	Z	BIN SIZE	BANDWIDTH	P > |T|
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

TABLE A3. Results of Cattaneo, Jansson and Ma test at £10 million

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	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	5.6	5.6
Method	Robust		Robust	
T	1.6918		0.3893	
P > |T|	*0.091**		*0.697*	

TABLE A4. Results of Cattaneo, Jansson and Ma test at £20 million

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	106	78	1151	608
Effective number of observations	72	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		-102899	
P > |T|	*0.936*		*0.1971*	

References

Aczel, Balazs, Barnabas Szaszi, and Alex O. Holcombe. 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, Stéphanie and Jean-Louis Chandon. 2013. "Will Monitoring Systems Kill Intrinsic Motivation? An Empirical Study." *Revue de Gestion Des Ressources Humaines* N° 90(90):35–53.

Aucoin, Peter and Ralph Heintzman. 2000. "The Dialectics of Accountability for Performance in Public Management Reform." *International Review of Administrative Sciences* 66(1):45–55.

Banerjee, Abhijeet, Tahir Andrabi, Sally Grantham-McGregor, Hirokazu Yoshikawa, Jaime Saavedra, Rukmini Banerji, Kwame Akyeampong, Susan Dynarski, Rachel Glennerster, Karthik Muralidharan, Sylvia Schmelkes, and Benjamin Piper. 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?*

Banuri, Sheheryar, Stefan Dercon, and Varun Gauri. 2017. "Biased Policy Professionals." *World Bank Policy Research Working Paper* 8113(June).

Bertelli, Anthony 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, Anthony M. 2007. "Determinants of Bureaucratic Turnover Intention: Evidence from the Department of the Treasury." *Journal of Public Administration Research and Theory* 17(2):235–58.

Besley, Timothy. 2007. *Principled Agents? : The Political Economy of Good Government*. Oxford University Press.

Black, Fischer. 1986. "Noise." *Journal of Finance* 41(3):529–43.

Bloom, Nicholas, Renata Lemos, Raffaella Sadun, and John Van Reenen. 2015. "Does Management Matter in Schools?" *Economic Journal* 125(584):647–74.

Bloom, Nicholas, Raffaella Sadun, and John Van Reenen. 2016. "Management as a Technology?" *NBER Working Paper*.

Bovens, Mark. 2007. "Analysing and Assessing Accountability: A Conceptual Framework." *European Law Journal : Review of European Law in Context* 13(4):447–68.

Bowles, Samuel. 2016. *The Moral Economy : Why Good Incentives Are No Substitute for Good Citizens*.

Brehm, John and Scott Gates. 1997. *Working, Shirking, and Sabotage : Bureaucratic Response to a Democratic Public*. Michigan: University of Michigan Press, c1997.

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

Briggs, Ryan C. 2020. "Results from Single-Donor Analyses of Project Aid Success Seem to Generalize Pretty Well across Donors." *Review of International Organizations* 15(4):947–63.

Bright, Liam Kofi. 2021. "Why Do Scientists Lie?" *Royal Institute of Philosophy Supplement* 89(May):117–29.

Buchanan, James and Gordon Tullock. 1962. *The Calculus of Consent*. Ann Arbor, MI: University of Michigan Press.

Calonico, Sebastian, Matias D. Cattaneo, and Max H. Farrell. 2021. "Optimal Bandwidth Choice for Robust Bias-Corrected Inference in Regression Discontinuity Designs." *The Econometrics Journal* 23(2):192–210.

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

Cattaneo, Matias D., Ann Arbor, Michael Jansson, and Xinwei Ma. 2018. "Manipulation Testing Based on Density Discontinuity." *The Stata Journal* 18(1):234–61.

Chetty, Raj, Emmanuel Saez, and László Sándor. 2014. "What Policies Increase Prosocial Behavior? An Experiment with Referees at the Journal of Public Economics †." *Journal of Economic Perspectives* 28(3):169–88.

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

Coviello, Decio, Giancarlo Spagnolo, and Clarissa Lotti. 2021. "Rules, Bunching and Discretion in Emergency Procurement: Evidence from an Earthquake." Pp. 13–22 in *Procurement in Focus: Rules, Discretion, and Emergencies*, edited by O. Bandiera, E. Bosio, and G. Spagnolo. London: CEPR Press.

Crijns, Tom J., Janna S. E. Ottenhoff, and David Ring. 2021. "The Effect of Peer Review on the Improvement of Rejected Manuscripts." *Accountability in Research* (ahead-of-print):1.

Deci, Edward L. and Wayne F. Cascio. 1972. "Changes in Intrinsic Motivation as a Function of Negatrive Feedback and Threats 24p .; Paper Presented at the Eastern Psychological Association Meeting in Boston , Massachusetts , April * Behaviora1 Science Research ; Learning Motivation ; * Low Motivatio." (February 2016).

DFID. 2011. *DFID's Approach to Value for Money (VfM)*.

DFID. 2020. *Smart Rules*.

Dissanayake, Ranil. 2021a. "Hitting 0.7 For 0.7's Sake: The Perils of the Global Aid Funding Target | Center For Global Development." *Views from the Center: Center for Global Development*. Retrieved July 8, 2021 (https://www.cgdev.org/blog/hitting-07-perils-global-aid-funding-target).

Dissanayake, Ranil. 2021b. *The Importance of Being Earnest: Noise, Incentives and Hierarchy in Public Sector Decision-Making*.

Dunsch, Felipe, David Evans, Ezinne Eze-Ajoku, and Mario Macis. 2021. "Management, Supervision, and Health Care: A Field Experiment." *NBER Working Paper Series* 23749.
Ellison, Glenn. 2002. "The Slowdown of the Economics Publishing Process." *Journal of Political Economy* 110(5):947–93.
Falk, Armin and Michael Kosfeld. 2006. "The Hidden Costs of Control." *American Economic Review* 96(5):1611–30.
Flesher, Dale L. and Marilyn Taylor Zarzeski. 2002. "The Roots of Operational (Value-for-Money) Auditing in English-Speaking Nations." *Accounting and Business Research* 32(2):93–104.
Frey, Bruno S. 1993. "Does Monitoring Increase Work Effort? The Rivalry with Trust and Loyalty." *Economic Inquiry* 31(4):663–70.
Gavas, Mikaela; and Rachael; Calleja. 2020. "DfID Is a World Leader in Tackling Poverty. Our International Standing Is Weakened without It | Aid | The Guardian." *The Guardian*.
Gawande, Atul. 2010. *The Checklist Manifesto : How to Get Things Right. Profile.*
Hares, Susannah and Pauline Rose. 2021. "As It Assumes Leadership of the Global Education Agenda, the UK Slashes Its Own Aid to Education | Center For Global Development." *Views from the Center: Center for Global Development*. Retrieved July 8, 2021 (https://www.cgdev.org/blog/it-assumes-leadership-global-education-agenda-uk-slashes-its-own-aid-education).
Higgs, MeganD. and Andrew Gelman. 2021. "Research on Registered Report Research." *Nature Human Behaviour.*
Hoggett, Paul. 1996. "New Modes of Control in the Public Sector." *Public Administration (London)* 74(1):9–32.
Honig, Dan. 2018. *Navigation by Judgment : Why and When Top-down Management of Foreign Aid Doesn't Work*. New York.
Hutchins, Edwin. 1995. *Cognition in the Wild*. MIT Press.
ICAI. 2018. *DFID's Approach to Value for Money in Programme and Portfolio Management A Performance Review.*
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.*
Ioannidis, John P. .. 2019. "Why Most Published Research Findings Are False." *Chance (New York)* 32(1):4–13.
Kahneman, Daniel, Dan P. Lovallo, and Olivier Sibony. 2019. "A Structured Approach to Strategic Decisions." *MIT Sloan Management Review* 60(1):1–12.
Kahneman, Daniel, Andrew M. Rosenfield, Linnea Gandhi, and Tom Blaser. 2016. "Noise: How to Overcome the High, Hidden Cost of Inconsistent Decision Making." *Harvard Business Review* 2016(October).
Kaufman, Daniel. 2021. "It's Complicated: Lessons from 25 Years of Measuring Governance." *Brookings Future Development*. Retrieved January 7, 2022 (https://www.brookings.edu/blog/future-development/2021/09/30/its-complicated-lessons-from-25-years-of-measuring-governance/).
Kells, Stuart. 2011. "The Seven Deadly Sins of Performance Auditing: Implications for Monitoring Public Audit Institutions." Australian Accounting Review 21(4):383–96.
Kennedy McDade, Kaci and Wenhui Mao. 2021. "UK Aid Cuts Will Put Global Health Systems at Risk—The BMJ." *The BMJ Opinion*. Retrieved July 8, 2021 (https://blogs.bmj.com/bmj/2021/06/11/uk-aid-cuts-will-put-global-health-systems-at-risk/).
Kerr, Steven. 1995. "On the Folly of Rewarding A, While Hoping for B." *Academy of Management Perspectives* 9(1):7–14.
Lankester, Tim. 2013. *The Politics and Economics of Britain's Foreign Aid : The Pergau Dam Affair*. Routledge.
Lindblom, Charles 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, Elizabeth A., Nancy Beaulieu, Robert Gibbons, Peter Pronovost, and Thomas Wang. 2015. "Organizational Culture And Performance." *American Economic Review: Papers & Proceedings* 3(4):512–27.
McCrary, Justin. 2008. "Manipulation of the Running Variable in the Regression Discontinuity Design: A Density Test." *Journal of Econometrics* 142(2):698–714.
Mitchell, Ian and Arthur Baker. 2019. "How Effective Is UK Aid? Assessing the Last 8 Years of Spending | Center For Global Development." *Views from the Center: Center for Global Development*. Retrieved July 8, 2021 (https://www.cgdev.org/blog/how-effective-uk-aid-assessing-last-8-years-spending).
Mitchell, Ian, Sam Hughes, and Euan Ritchie. 2021. "An Overview of the Impact of Proposed Cuts to UK Aid | Center For Global Development." *Views from the Center: Center for Global Development*. Retrieved July 8, 2021 (https://www.cgdev.org/publication/overview-impact-proposed-cuts-uk-aid).
Molander, Per. 2014. "Public Procurement in the European Union: The Case for National Threshold Values." *Journal of Public Procurement* 14(2):181–214.
Morin, Danielle. 2001. "Influence of Value for Money Audit on Public Administrations: Looking Beyond Appearances." *Financial Accountability & Management* 17(2):99–117.
Palguta, Ján and Filip Pertold. 2017. "Manipulation of Procurement Contracts: Evidence from the Introduction of Discretionary Thresholds." *American Economic Journal*. Economic Policy 9(2):293–315.
Power, Michael. 1999. *The Audit Society: Rituals of Verification*. Oxford: Oxford University Press.

Rainey, Hal G. 1993. "A Theory of Goal Ambiguity in Public Organizations." Pp. 121–66 in *Research in public administration*. Vol. 2.

Sasse, Tom and Emma Norris. 2019. "Moving On: The Costs of High Staff Turnover in the Civil Service." *Institute for Government* 78.

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

Simon, Herbert A. 1960. *The New Science of Management Decision*. New York.

Simon, Herbert A. 1997. *Administrative Behavior: A Study of Decision-Making Processes in Administrative Organizations*.

Soderberg, CourtneyK., TimothyM. Errington, SarahR. Schiavone, Julia Bottesini, FelixSingleton Thorn, Simine Vazire, KevinM. Esterling, and BrianA. Nosek. 2021. "Initial Evidence of Research Quality of Registered Reports Compared with the Standard Publishing Model." *Nature Human Behaviour*.

Tas, Bedri Kamil Onur. 2019. *Bunching Below Thresholds to Manipulate Public Procurement*. RSCAS 2019/17.

UK Parliament. 2020. *Effectiveness of UK Aid: Potential Impact of FCO/DFID Merger—International Development Committee—House of Commons*.

Della Vigna, Stefano and Elizabeth Linos. 2020. *RCTs to Scale: Comprehensive Evidence from Two Nudge Units*. Cambridge, Mass: National Bureau of Economic Research.

Watts, Richard. 2021. "The Latest View of UK Aid: Death by a Thousand Cuts | Save the Children UK." *Save the Children Blogs*. Retrieved July 8, 2021 (https://www.savethechildren.org.uk/blogs/2021/the-impact-of-uk-aid-cuts).

Williamson, Oliver E. 1999. "Public and Private Bureaucracies : A Transaction Cost Economics Perspective." *Journal of Law, Economics, & Organization* 15(1):306–42.

Williamson, Oliver E. 2002. "The Theory of the Firm as Governance Structure: From Choice to Contract." *Journal of Economic Perspectives* 16(3):171–95.

Ziegelmeyer, Anthony, Katrin Schmelz, and Matteo Ploner. 2012. "Hidden Costs of Control: Four Repetitions and an Extension." *Experimental Economics* 15(2):323–40.