DEVELOPMENT POLICY CENTRE

Discussion Paper 108
March 2024
Series ISSN 2206-303X

The impact of independent review on assessments of aid project effectiveness

Terence Wood and Stephen Howes

Abstract

In this paper we test whether increasing the independence of the project appraisal process changes the reported effectiveness of aid projects. We do this using a dataset of Australian aid appraisals and a natural experiment, which occurred when a more independent process involving DFAT's central evaluation unit and external contractors was implemented for the review of performance appraisals of completed projects. Using difference-in-differences and contrasting assessments of ongoing projects, which the appraisal process was not changed for, and completed projects, where the process was changed, we show that introducing a more independence led to a substantial fall in how successful projects were deemed to be. We also show that the change probably led to more accurate recording of COVID-19's impact on Australian aid, as well as more accurate assessments of the quality of Australia's aid to Papua New Guinea, its largest aid partner. As we do this, we take care to demonstrate that our findings are robust to the types of methodological issues that can afflict difference-in-differences studies.

The Development Policy Centre is part of Crawford School of Public Policy at The Australian National University. We undertake research and promote discussion on Australian aid, Papua New Guinea and the Pacific and global development policy.

Australian National University

The impact of independent review on assessments of aid project effectiveness

Terence Wood

Stephen Howes

Terence Wood is a Fellow at the Development Policy Centre.

Stephen Howes is Director of the Development Policy Centre and Professor of Economics at the Crawford School of Public Policy, at The Australian National University.

Wood, T & Howes, S 2024 "The impact of independent review on assessments of aid project effectiveness," *Development Policy Centre Discussion Paper 108,* Crawford School of Public Policy, The Australian National University, Canberra.

This research was undertaken with the support of The Bill & Melinda Gates Foundation. We are also very grateful to Huiyuan Liu and Cameron Hill for their assistance with the research.

The impact of independent review on assessments

of aid project effectiveness

1 Introduction

In an attempt to move away from unhelpfully broad debates about whether aid works or not, scholars have increasingly taken to using datasets of aid project performance ratings to study what influences aid and the success or failure of aid at the project level (Ashton et al., 2022). The data used in this type of analysis come from individual project appraisals produced by aid agencies themselves. In addition to scholars, aid agencies also use these performance ratings, both in their high level reporting, and in some agencies when deciding the fates of individual projects.

In this paper, we contribute to the existing literature on aid project performance by using performance ratings in a different way, focused on better understanding the numbers themselves, their validity, and what they reveal about internal dynamics within aid agencies. We do this by testing whether making the performance appraisal process more independent affects project performance ratings. Our analysis is based on a dataset of performance ratings from Australian aid project appraisals and a natural experiment, which occurred when the process of verifying the end-of-project performance ratings produced by project managers was taken out of the hands of country teams and passed to a central evaluation unit that then tasked an external contractor with reviewing the ratings.

Using difference-in-differences and contrasting assessments of ongoing projects, which continued to be produced as before by project managers and which also continued to be reviewed internally, and assessments of completed projects, which were subjected to the more independent new approach, we show that the shift to greater independence led to a dramatic fall in how successful projects were deemed to be. We also show that the change probably led to more accurate recording of the impact of COVID-19 on Australian aid effectiveness, as well as the effectiveness of aid to Papua New Guinea (PNG), Australia's largest aid partner. As we do this, we take care to demonstrate that our findings are robust to the types of methodological issues that can afflict analysis using difference-in-differences.

Our findings make both practical and theoretical contributions to existing work on the performance of aid projects. Practically, we demonstrate just how vulnerable assessments of aid project performance are to aid staff being unduly generous when assessing their own work, even when assessments are reviewed internally, as was the case in the Australian aid program prior to the 2019 change. This is relevant for high-level public performance reporting that draws on data of this sort (for example, DFAT, 2020c; House of Commons International Development Committee, 2020). It is also relevant in instances when aid agencies use scores produced in project assessments to make decisions about whether projects should be continued, closed or revamped (for example, DFAT, 2022a). If aid project appraisals are to serve as a useful guide either to the politicians and public who are the intended targets of high-level reports, or to agencies as they decide which projects to close and which to continue, project scores

need to reflect actual performance as much as possible. For researchers, our practical contribution is also relevant to work based on project assessments. At present, some scholarly work draws only on project assessment data that has been subject to external review, while other work combines data from project assessments which have been subject to different types of review (for various examples see, Bulman et al., 2017; Denizer et al., 2013; Feeny & Vuong, 2017; Honig, 2018). Our findings demonstrate the potential for future learning using data from assessments subject to different types of review and studying in detail how they differ.

More theoretically, the role of the external contractor in verifying project performance ratings, and the fact that the use of a contractor was associated with a fall in ratings, offers insights relevant to the study of how incentives shape dynamics within aid agencies and when agencies contract tasks to external agents. It is not necessarily surprising to find that aid staff produce excessively positive assessments of the performance of their own projects, but it is more surprising to discover that internal processes, such as those that existed in the Australian aid program prior to the change, designed to serve as a check on this behaviour, sometimes do not. Moreover, a *priori* there were few grounds for anticipating that if internal processes were failing to prevent inaccurate project appraisals, using an external contractor would change matters. Purely in terms of incentives, the most obvious course of action for the contractor would have been to avoid antagonising the aid program by downgrading

project performance scores. Yet the contractor chose otherwise in this case, for reasons we explore in the paper's discussion.[1]

The paper proceeds as follows: it summarises work based on aid project appraisals; it also reviews research on incentives at work within aid agencies and when contractors are used in aid. Then it provides background information on the aid project appraisals that produced our data, summarises the data and details the empirical approaches that we drew on in studying the data. After that, we address possible empirical issues and present our results. We then demonstrate the practical significance of our findings using the example of Papua New Guinea. Finally, we discuss the ramifications of our findings.

2 Literature, research questions and background

2.1 Literature

The existing literature based on aid project performance assessments has tended to focus on one of two core areas. The first is recipient country and project traits associated with more successful projects. A number of country traits have been found in instances to be positively associated with better project performance. Gross domestic product, economic growth and good governance have often been found to be positively

[1] It is worth noting that it could be possible that the external reviewers' revised appraisals were less accurate than the project managers'. However, there are good grounds to believe that this is not the case. The external reviewers had access to the full information on project performance. They could also investigate further and ask for more information if they felt it warranted. In addition, revised project scores were discussed with project teams and the teams themselves could contest them if they felt the final assessments were unfair.

associated with better performance, although not always. Findings for some other traits, such as political and civil liberties have varied widely between different studies (Bulman et al., 2017; Denizer et al., 2013; Feeny & Vuong, 2017; Isham et al., 1997; Kilby, 2000; Wood et al., 2020). Similarly diverse findings exist when looking at project traits: no clear consensus has emerged on, for example, the most likely sectors to produce successful projects. (For a good selection of findings see: Briggs, 2019; Bulman et al., 2017; Denizer et al., 2013; Feeny & Vuong, 2017.)

Some findings are more consistent though. Perhaps the most striking of these is that, even when limiting data to individual donors, project performance appears to vary considerably more within countries than between countries (Briggs, 2019; Bulman et al., 2017; Denizer et al., 2013; Feeny & Vuong, 2017; Wood et al., 2020). In other words, there are few aid recipient countries, or types of aid recipient countries, where projects are destined to fail, and few places where projects will inevitably succeed.

A second type of work has focused on the donor side of aid project delivery, studying, for example, whether divesting more decision-making power to country officers or increasing project supervision affects outcomes (Ashton et al., 2022; Honig, 2018, 2019, 2020; Kilby, 2000). Other studies in this vein include whether project manager capacity and turnover is associated with project performance (Ashton et al., 2022; Bulman et al., 2017).

Most researchers studying project appraisals note the risk that project appraisals could be biased (for an excellent discussion of sources of bias see: Kilby, 2000, p. 239). In

discussing this risk, some authors have contended that overall inflation of project scores is not a source of particular concern, as long as inflation is equal across the board and projects are scored accurately relative to each other (Wood et al., 2020). In other instances, authors have gathered additional evidence in an attempt to demonstrate that bias does not undermine their findings (for example, Denizer et al., 2013; Honig, 2019; Kilby, 2000).

Although no one has to-date studied the impact of changing to more rigorous systems of appraising projects or validating project appraisals, the World Bank has stated that staff-generated appraisals of ongoing projects, and final project appraisals, which are reviewed by its semi-autonomous *Independent Evaluation Group* differed for 18 and 15 percent of its projects in the 2019 and 2020 financial years (World Bank, 2021, p. 15). In their work using World Bank data Kilby and Michaelowa (2019) found that in most cases apprasials became worse when subjected to more stringent reviews. In other scholarly work, both Bulman et al. (2017) and Feeny and Vuong (2017) found that more rigorous appraisals were associated with lower appraisal scores in the ADB.

These findings are relevant to our work and give us some cause to anticipate that the 2019 change in Australia would bring changes in appraised project performance in its wake. However, the findings in existing research all come from the context of multilateral development agencies, not a government aid program, and all either stem from simple comparisons between appraisals that were not reviewed and appraisals that were (Kilby & Michaelowa, 2019; World Bank, 2021) or dummy variables added as controls in multiple regressions focused on other aspects of aid agencies' work (Bulman

et al., 2017; Feeny & Vuong, 2017). The findings in existing work do not emerge from changes in appraisal processes, nor do they involve contracting out the review of appraisals to external contractors.

Other work informed by the theory-driven analysis of case studies, provides additional insights about what could have occurred when the project appraisal process was made more independent in Australia. In particular, Martins et al. (2002) and Gibson et al. (2005) provide rich studies of how incentives influence dynamics within aid agencies as well as how issues such as principal agent problems affect relationships between donors and contractors.

If the staff of aid agencies reliably respond to incentives and pursue objectives such as advancing their careers, it is easy to see the potential problems that might arise when staff appraise the performance of their own projects (Martins, 2002; Seabright, 2002). Staff seeking to advance their careers might, for example, inflate scores in appraisals of projects they manage with a mind to the next promotion round. Or staff may simply wish to avoid the lengthy process associated with redesigning a project that is seen to be failing (Kilby, 2000). Less rationally, even staff who believe they are being honest in their appraisals may be inadvertently biased and overly inclined to see the merits of a project they have worked hard to deliver. Of course, internal review processes, such as the process Australia had in place prior to 2019 should serve as a check on such inflation. But it is also possible that internal dynamics within aid agencies might prevent this from occurring if, for example, aid management themselves want positive numbers to provide to the public and politicians (Martins, 2002). Even absent external pressures

to perform, internal reviewers might be reluctant to downgrade appraisals for the sake of managing relationships with colleagues. Some of the evidence from existing studies of project appraisals (for example, Feeny & Vuong, 2017; Kilby & Michaelowa, 2019) suggests both that staff do provide overly positive appraisals of projects they have worked on and that internal reviews serve as something of a check on this. This existing evidence comes from multilateral aid agencies though. Circumstances could be different in a bilateral donor such as Australia, where the need to please politicians and publics could be more acute (Seabright, 2002). What is more, in the World Bank and ADB the external evaluation units have a high degree of formal independence. This is not the case in the Australian aid program.

While having a central evaluation unit contract out the review and validation of project managers' own aid appraisals to a third party, as Australia did for final project appraisals in 2019, might seem like a sensible solution to problems with internal dynamics within aid programs, there are also reasons to doubt its efficacy. As agents, consultants have strong incentives to keep the aid donor — the principal in this relationship — happy, which is not necessarily the same thing as delivering high quality outcomes (in this case accurate validations) (Gibson et al., 2005). If the consultants involved wish to keep the contract, or win other contracts from the donor, they may have had good reasons not to revise scores for fear of making the donor look worse or adding to the workload of donor staff and thereby gaining a reputation as a problematic partner (Martens et al., 2002).

The existing literature provides some grounds to anticipate that the Australian change to more independent project appraisal validation might lead to lower project performance scores. And yet existing research also provides reasons to doubt the change would have had an effect. Reflecting this, the central question motivating our research was:

> Did the change in the project performance validation process in 2019 lead to significant changes in assessed project performance for those projects affected by it?

2.2 Background on aid appraisals in the Australian aid program

Like the other donors that have been the main focus of research on aid project performance to-date, Australia assesses the performance of its individual aid projects. All aid projects run by Australia's aid program with a total budget of over AU$3,000,000 (approximately $2,000,000 USD) are required to be appraised each year (DFAT, 2022a, p. 66).[2] Both ongoing and closing projects are appraised. Final project appraisals of

[2] Core funding to multilateral agencies is assessed via a different process and could not be covered in our analysis. Similarly, funding for humanitarian emergencies is also assessed in a somewhat different manner, and so was excluded from our analysis. Also, a small amount of Australian aid is aid delivered by parts of the Australian government other than its foreign ministry. This aid is not normally assessed, and therefore is excluded from our analysis.

closing projects occur in the final year of a project's operation or shortly after it closes (DFAT, 2022a).[3]

Appraisals contain qualitative descriptions of the aid project such as challenges encountered, as well as scores on a one to six scale reflecting perceived project performance (DFAT, 2022a). Performance is assessed in a range of areas, some of which have changed over time. However, project effectiveness and project efficiency were assessed throughout the years covered by this study. Effectiveness and efficiency also play a particularly prominent role in how the aid program reports on its overall performance to the public (for example, DFAT, 2020c). In addition, the two indicators play a central role in determining whether underperforming projects go through a formal remediation process, and whether they are closed early (DFAT, 2022a).[4]

Project ratings for ongoing and final reports are drafted by projects' managers along with their supervisors. Project ratings for ongoing projects are then reviewed and signed off by senior managers (DFAT, 2020a). Until 2019, the final assessments of closing projects went through a similar internal moderation process (DFAT, 2020b).

---

[3] If a project is slated for closing and a final appraisal is scheduled for a year no ongoing appraisal is conducted in that year.

[4] The DFAT programming guide outlines the aid programme's rules: "Investments with unsatisfactory ratings in their IMR (scores of 3 or below) for effectiveness and efficiency criteria…must provide Development Effectiveness and Enabling Division (PRD) with an IRI Remediation Plan… If performance against both the effectiveness and efficiency criteria remains unsatisfactory after one year, the FAS [head] of the program area will decide whether the investment should be cancelled" (DFAT 2022, p. 67).

However, from 2019 onwards the moderation process of completed projects was coordinated by the aid program's evaluation and performance unit which outsourced the review of appraisal scores to external consultants, which had the power to change appraisal scores (DFAT, 2020c, 2022a).

It is this 2019 change that provides the primary analytical leverage for our study. No other dramatic changes occurred to Australian aid around this point of time: there was an election, but the government did not change, there were no major changes in aid strategy, Australia did not shift its sectoral or regional aid focus (Wood et al., 2021). AusAID, the Australian aid agency had been disbanded, but that was six years prior, and not likely to cause a sudden change in Australian aid quality beginning in 2019.[5]

Of key importance for our analysis, the process of reviewing ongoing projects did not change in 2019. Only the appraisals conducted at the completion of aid projects were affected. As we discuss below, this means we have both pre-treatment and treatment periods. We also have a control group (ongoing projects) as well as a treatment group (completed projects) in the treatment period.

Also, importantly, throughout both the pre-treatment period and the treatment period, the project performance appraisals for ongoing and completed projects were completed

[5] There were projects started when AusAID existed in our sample, and these slowly dwindled over time; however, there was no sudden break in 2019. Moreover, AusAID projects did not receive more positive assessments on average, even under external validation.

using very similar templates, which contained nearly identical questions. They were also generated following very similar processes. In practice, performance assessments for ongoing projects and completed projects were not meaningfully different types of assessment. Moreover, in the pre-treatment period the two types of performance assessment did not typically generate different scores. Prior to the commencement of the external review in the treatment period, the median difference between projects' final performance scores and the average of their ongoing performance scores was zero.

3 Data, estimation and potential concerns

3.1 Data

Measures of assessed project effectiveness and efficiency play a central role in how the Australian aid program assesses its overall performance and the performance of specific projects. Reflecting this, we focus on these two measures. In our central analysis we use the arithmetic mean of the efficiency and effectiveness scores for individual projects in individual years as the dependent variable. In addition, in Online Appendix 4 we provide alternate results based on an approach that does not involve taking the mean of the two variables. All findings are substantively very similar.

We built the dataset used in this study by taking an existing dataset covering 2014 to 2017 produced by earlier researchers (Wood et al., 2020). We requested more recent data from the Australian foreign ministry who provided us data covering the period from 2017 to 2022. We then combined the two datasets. The combined data covered all appraised projects from 2014 to the end of 2022. As we built the dataset we also

gathered data on other project traits such as project size. At points in our analysis we included these traits as control variables. Summary statistics of the dependent variable of interest as well as the various control variables are provided in Table 1.

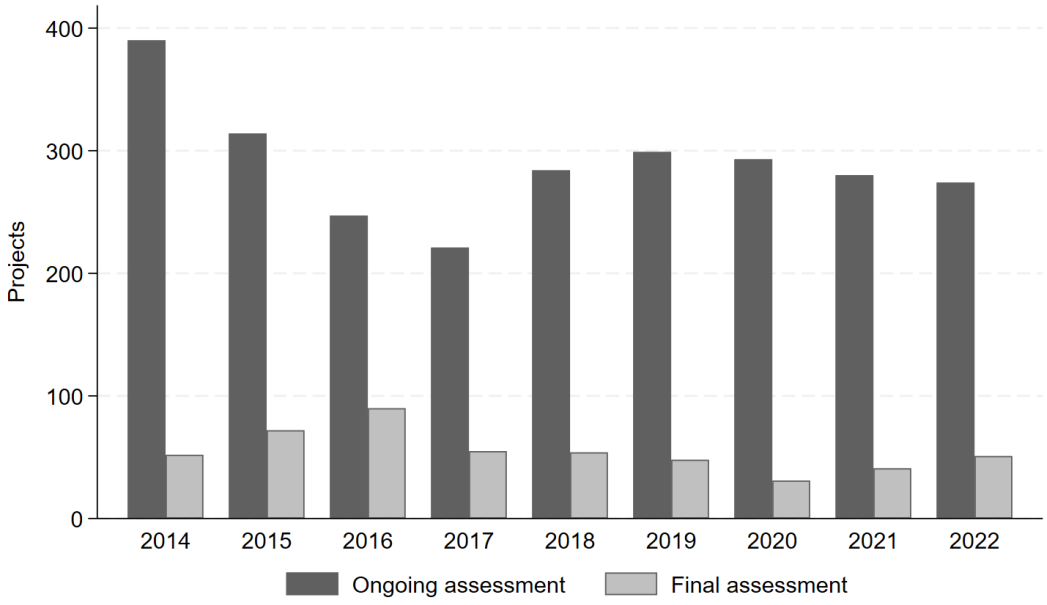
In Figure 1 we show the number of project appraisals completed annually over the years covered by our study. In any year there are considerably more projects underway than coming to an end, hence there are more appraisals of ongoing projects than of completed projects. Australian aid was repeatedly cut between 2015 and 2017. As a result the number of ongoing projects fell during these year, while completed projects rose. However, from 2018 onwards trends in appraisal numbers were fairly flat.

Table 1: Descriptive statistics

	Min	Max	Mean	Std. Dvn.	N	%
Effectiveness & Efficiency	1.00	6.00	4.27	0.65	3,096	
Budget (natural log)	12.03	20.25	16.80	1.14	3,096	
Project duration (planned)	60.00	9,861.00	2,525.71	1,088.16	3,096	
Assessment type						
Ongoing						84%
Final						16%
Sector						
Agriculture						8%
Resilience						12%
Education						18%
Governance						25%
General						6%
Health						14%
Economic						17%
Is project in Pacific?						
Elsewhere						63%
Pacific						37%

Notes: the data come from 2014 to 2022 and covers all aid projects over AUD $3m, operated by the Australian aid program during this period. Humanitarian emergency responses are excluded as is core funding to multilateral organisations: both are assessed through different processes.

Figure 1: Project assessments by type and year



3.2 Identification strategy

To identify whether adopting external validation affected appraised project performance we used difference-in-differences. Our analysis took several forms.

Our most basic approach involved a simple two by two analysis. The form of the analysis is shown in Equation 1.

$$(1)\ Y_{rt} = \propto\ + \gamma Type_r + \ \lambda Period_t + \ \beta Type \cdot Period_{rt} + \ \varepsilon_{rt}$$

Y is each project's appraised performance. Type is a dummy variable capturing the type of appraisal (either that of an ongoing project or of a completed project). Period is dummy variable capturing whether the project occurred before or during the external validation years. And β is the interaction term capturing the difference-in-differences across appraisal types and validation periods.

The basic intuition underpinning this analytical approach is that if average scores awarded to ongoing and completed projects had similar trends prior to the advent of external validation, and if no other changes occurred which caused ongoing and completed projects' performance to differ, β will capture the impact of more independent validation on appraised project performance.

We address the issue of parallel trends in scores awarded to ongoing and completed projects in Section 3.3. In addition, to reduce the risk that any observed findings were spurious and actually the results of other changes in Australian aid such as more projects in more difficult sectors or increased work on parts of the world where it was

harder to deliver aid effectively, we also included controls for a suite of project traits in some of our regression models. When we did this, the regression specifications took the form shown in Equation 2:

$$(2)\ Y_{rt} = \propto\ + \gamma Type_r + \lambda Period_t + \beta Type \cdot Period_{rt} + \delta Controls + \varepsilon_{rt}$$

Controls is a vector of control variables covering project duration, (the natural log) of project size, sector, and whether the project is in the Pacific region or not. We added these controls because other work, including work focused on Australian aid, has also found larger projects tend to be assessed as more effective (Wood et al., 2020). Other work has also found project duration associated with effectiveness (Feeny & Vuong, 2017). The plurality of Australia's aid goes to the Pacific region, a region where both Australian aid projects and projects from other donors have been found to be less successful (Feeny & Vuong, 2017; Wood et al., 2022).

To further reduce the risk that changes in the nature of projects were driving findings, or that other unobservable trends in project performance were behind observed changes, we also included project fixed effects in some of our models. Because different projects come to an end and receive their final appraisal in different years, the analytical approach for this analysis we use two-way fixed effects and the regression model takes the following form.

$$(3)\ Y_{pt} = \propto\ + \gamma Project_p + \lambda Year_t + \beta Treated_{pt} + \varepsilon_{pt}$$

Project and Year are project and year fixed effects, Treated, the variable of interest, is a dummy variable coded one if the appraisal is a treated appraisal (in effect, the final

appraisal of a project when the appraisal occurs after 2018). In this model the coefficient for Treated value is derived from the extent to which individual projects' appraisal scores in a year differ from their scores in the previous year, and, specifically, the extent to which this difference differs between externally validated appraisals (treated appraisals) and other appraisals.

Finally, to examine dynamic effects (the extent to which the effect of appraisal validation changed over time) we also undertook a series of event studies. The event study version of the project fixed-effects model shown in Equation 4 (its application is discussed in more detail when we report on the event study results). Project and Year are project and year fixed effects. Final is a dummy variable capturing whether the assessment is an ongoing or final assessment. Reflecting this, leverage in the analysis comes from the difference in projects' final appraisals and the rating in the ongoing appraisal from the year prior, and how this difference differs once treatment is introduced:

$$(4)\ Y_{pt} = \propto\ + \gamma Project_p + \ \lambda Year_t + \ \delta Final_p + \beta Final \cdot Year_{pt} + \ \varepsilon_{pt}$$

3.3 Addressing potential concerns

The assumption that potential outcomes of the treated and control groups would have evolved in a parallel absent treatment is central to difference and difference analysis. If trends in the outcome variable are not parallel prior to treatment, the internal validity of any difference in difference model is open to question (Angrist & Pischke, 2009).

Figures 3, 4 and 5 provide visual evidence that the appraised performance of completed projects trended broadly in tandem with the appraised performance of final projects prior to the introduction of external validation. In Table 2 we provide a formal test of this across years 2014 to 2018, testing to see whether the trend in ongoing projects' scores over the years differs from the trend in completed projects. Model 1 in the table shows a comparison from a regression without any control variables included. Model 2 shows the same regression but with a suite of project traits in the control. Neither model provides evidence of any statistically significant deviation from parallel trends.

Table 2 – Testing for parallel trends pre-treatment

	(1)	(2)
Difference in trends	0.03	0.04
	(0.03)	(0.03)
Year	0.01	0.00
	(0.01)	(0.01)
Final report	-0.10	-0.11
	(0.07)	(0.07)
Project Controls	No	Yes
Observations	1779	1779

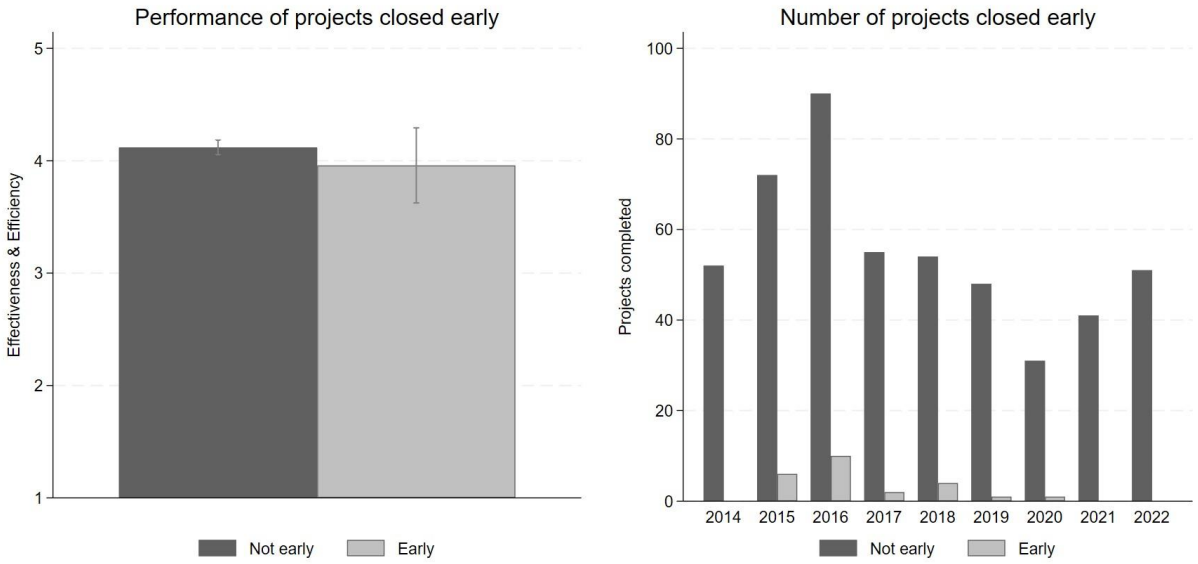
Robust standard errors clustered at the project level. The "parallel trends issue" coefficient is the interaction of year and final report. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Another possible threat to the internal validity of our approach could plausibly stem from poorly performing projects being closed in advance of the 2019 change in an attempt to avoid the impending more rigorous appraisal validation system. The Australian aid program announced in advance that it was making the change (DFAT, 2019). As a result, this could have been possible, although unlikely in our view. Regardless of the likelihood, we were able to check the performance and number of

projects that were closed early by comparing projects' planned closure dates at opening, or when they first appeared in the data, with actual closing dates.

The results of these comparisons are visible in Figure 2. As the left panel shows, the average performance of projects which were closed early does appear marginally worse than projects that were not closed early. However, as can be seen in the second panel, there is no evidence of a large surge in projects being closed early in the years 2018 or 2019. Indeed, in all years, projects which closed early were only a very small share of all the projects completed in that year. There is no evidence that a large number of poorly performing projects were closed early in an attempt to avoid more rigorous appraisal.

Figure 2: Projects which close early



Notes: "Not early" are projects that closed in the year they were originally planned to cease. "Early" are projects that closed in an earlier year than that initially planned. The rise in project closures in 2015 and 2016 is associated with major cuts to the aid budget in 2015.

Another possible source of spurious findings would have been changes in the nature of Australian aid over time. Perhaps, for example, Australia began increasingly delivering

its aid to the Pacific, a region where it is harder to deliver aid effectively (Wood et al., 2022). This could plausibly impact project performance, and changes over time might have contributed to differing trends in the performance of ongoing and completed projects. If changes were significant and occurred prior to 2019, their effect ought to have been captured in parallel trends tests, but changes that straddled the introduction of independent project reviews could still bias findings. In Online Appendix 1 we show the results of tests that examine differences in key project traits between completed and ongoing projects prior to treatment and once treatment was introduced. We also test for any difference in these differences across the two periods. We do find some differences between completed and ongoing projects. However, the only differences that differ between the between the pre-treatment and treatment periods are to do with project sectors, and other work on Australian aid performance has shown that performance itself differs little sectors (Wood et al., 2020). Nevertheless, to reduce the risk of any bias emerging in our findings from the changing nature of Australian aid, we control for a range of project traits, including sector, in our analysis as well as applying project fixed effects in instances.

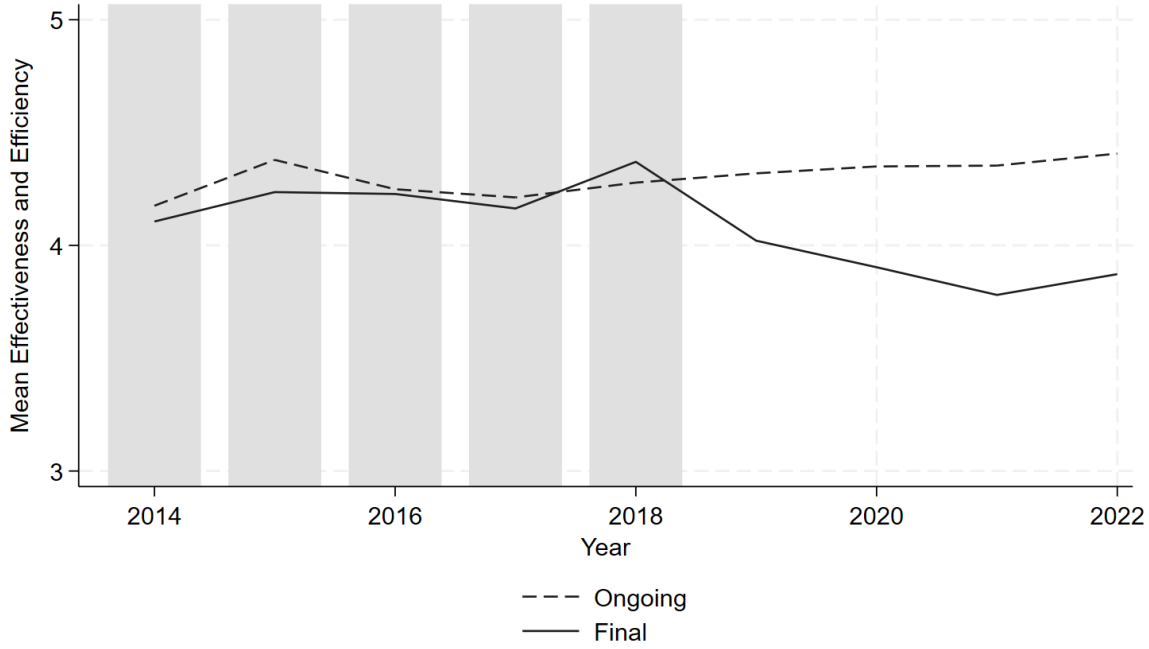
One final threat to the internal validity of our analysis that is worth discussing are problems associated with so-called "forbidden comparisons", which can occur in studies involving treatments applied to multiple groups at different times (for a clear discussion of the problem see, Goodman-Bacon, 2021). Forbidden comparisons occur when two-way fixed-effects models are used in studies of this nature because the method inadvertently leads to comparisons between newly treated groups and already treated

groups, in addition to the desired comparison between treated groups and untreated groups. This can be a serious issue for the accurate estimation of treatment effects. However, it is not a problem in our work. In our initial models, which are based solely on comparisons between appraisals of ongoing projects and completed projects, there is just one treatment period: 2019 onwards. All appraisals of completed projects after that date were treated, while no appraisals of ongoing projects are ever treated. The situation is different, however, when we add project fixed effects to our models. When this occurs, the treatment date does vary: projects are completed on different dates. However, forbidden comparisons still do not occur because projects are appraised at completion and are never appraised again. As a result, they do not occur in subsequent years' data, which means they never take on the role of early treated projects being inappropriately compared with newly treated projects.

4 Results

Figure 3 compares the mean project performance scores for ongoing and completed projects over time. Those years prior to the introduction of external validation are shaded grey. Two features of the chart stand out. First, although there are minor deviations, in accordance with the results of the parallel trends test above, appraised performance for ongoing and completed projects trend broadly in tandem prior to the introduction of the external validation process. However, in the wake of the introduction of the new system, scores deviate considerably: completed projects receive notably worse scores than ongoing projects.

Figure 3: Appraised project performance over time



Notes: The value for "Ongoing" is the mean score for all ongoing projects assessed in that year. The value for "Final" is the mean score for all projects that came to a close in that year, and which received their final assessment. In the shaded years on the chart ongoing and final project assessments were reviewed internally. In the unshaded years, the assessments of final projects were sent to external consultants for review.

4.1 Basic difference-in-differences analysis

Table 3 contains formal difference-in-differences analysis. Models 1 and 2 are standard two by two difference-in-differences regressions. Model 1 is run without control variables. In Model 2 controls for project duration, size and sector, as well as a dummy variable for whether projects are in the Pacific or not. In Models 1 and 2 standard errors are clustered at the project level. This would seem to be a natural unit for clustering – serial correlation, for example, would clearly be most likely within the same projects over time. However, a case could be made that clustering should simply be based on whether a project is treated or not. This approach leaves only two clusters though. Far

too few for standard errors to be calculated correctly (Bertrand et al., 2004). Fortunately, a reasonable alternative approach exists which can accommodate circumstances such as these. This is aggregation of Donald and Lang (2007). Results from this approach are shown in Models 3 and 4. Once again, the treatment effect is very similar.

Table 3: Difference-in-differences, standard and with Donald and Lang aggregation

	(1)	(2)	(3)	(4)
Diff in Diff	-0.42***	-0.42***	-0.43***	-0.42***
	(0.07)	(0.07)	(0.07)	(0.08)
Final	-0.03	-0.03		
	(0.04)	(0.04)		
After 2018	0.10***	0.08**		
	(0.03)	(0.03)		
Project Controls	No	Yes	No	Yes
Observations	3096	3096	18	18

Notes: Robust standard errors clustered at the project level in Models 1 & 2; Donald and Lang Aggregation Models 3 & 4: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. The control variables included in Models 2 & 4 were the natural log of project size, project duration, sector, and whether the project was in the Pacific or not.

Another possible approach to difference-in-differences analysis with the project data we have is to add project fixed effects. With these added, there is now no longer any question about the appropriate level for clustering standard errors – standard errors should be clustered at the project level. Adding project fixed effects also brings the benefit of controlling for any unobserved differences in Australian aid projects' characteristics, most importantly characteristics which may have changed over time. Table 4 shows the results from regressions with project fixed effects added. The coefficient for the change in appraisal validation procedures is very close in magnitude to those produced by the other models. In all models it is worth noting that the effect is non-trivial. Project scores can vary between one and six, meaning a coefficient of -0.45

is about 7.5 per cent of potential variance. However, in reality project scores cluster. The interquartile range of the data is itself only one – the change associated with treatment is 45 per cent of this.

Table 4: Project fixed effects

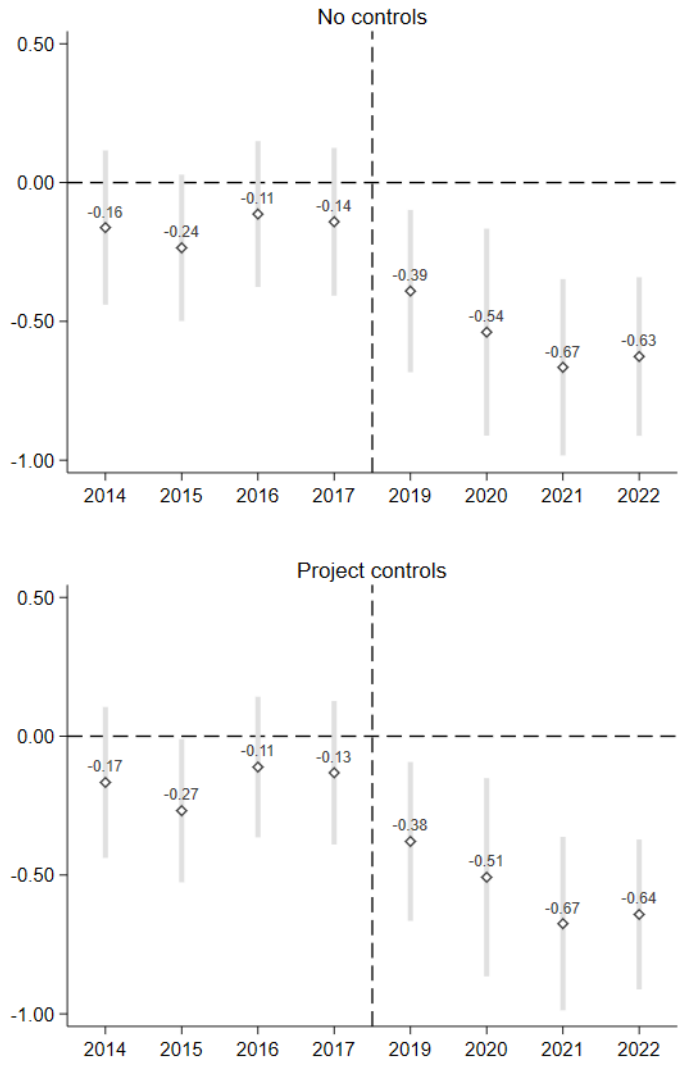
	(1)
Diff in Diff	-0.45***
	(0.07)
Project FE	Yes
Observations	2781

Notes: Robust standard errors clustered at the project level. This model differs from those in the previous table in that project fixed effects are added. The sample size is smaller because some projects could not be tracked over time. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

4.2 Effects over time

In our formal analysis so far we have relied on simple before and after comparisons. However, inspection of Figure 3 suggests the treatment effect itself changed over time. Average scores for ongoing and completed projects diverge in 2019, but they do not thereafter return to paralleling each other. Rather, performance of completed projects continues to trend worse than that of ongoing projects in 2020 and 2021, only returning to what might possibly be a parallel trend in 2022. Figure 4 shows event studies, which plot the difference between ongoing and completed projects over time. The first panel comes from a simple two way fixed effects regression model, the second comes from the same model but with control variables included. Results are very similar.

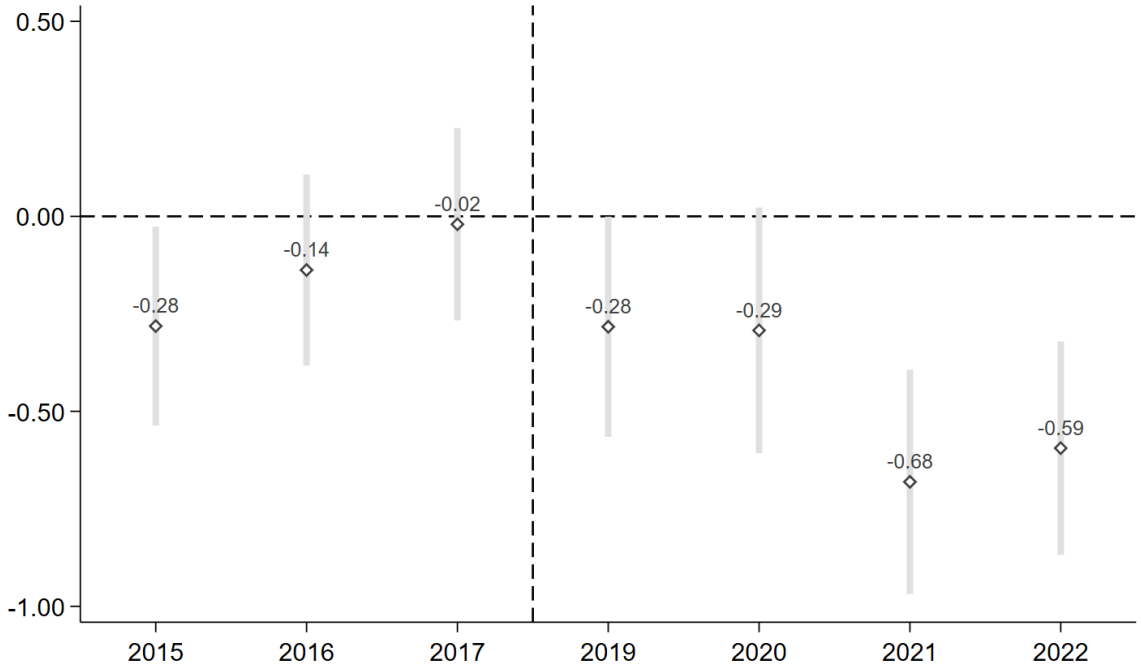
Figure 4: Event study, with and without project controls



Notes: the value plotted in the y-axis is the coefficient from the interaction between year and report type (ongoing or final). The year prior to the introduction of external appraisal review (2018) is the omitted comparator year. 95 percent confidence intervals are plotted. Analysis was run with robust standard errors clustered at the project level.

Figure 5 is similar to Figure 4, but contains project fixed effects. Results are similar, although the drop from 2019 to 2020 is somewhat less.

Figure 5: Event study, project fixed effects



Notes: the value plotted in the y-axis is the coefficient from the interaction between year and report type (ongoing or final). The year prior to the introduction of external appraisal review (2018) is the omitted comparator year. Some projects (those which changed project codes or with issues such as multiple final appraisals, or ongoing appraisals after final appraisals) were excluded from analysis. Analysis was run with robust standard errors clustered at the project level. 95 percent confidence intervals are plotted. The p-value for the coefficient for 2019 is 0.049. The p-value for 2020 is just above the conventional threshold for statistical significance, although it is close at 0.069.

Briefly, it is worth addressing the values for 2015 and 2016.[6] In both years completed projects had worse recorded performance than ongoing appraisals (compared to the comparator year of 2018). This almost certainly stems from aid cuts affecting those years and associated project closures. Taken together the coefficients for 2015 to 2017 seem to suggest a trend (although only one coefficient, that of 2015 is statistically

[6] There is no coefficient for 2014 in the project fixed effects models as we have no data prior to 2014 and so, therefore, projects ending in this year had no ongoing appraisal scores from earlier years to serve as comparators.

significant). However, even if such a trend exists it would lead to the expectation of that completed projects would score higher than ongoing projects in the treatment period. In fact, the opposite occurs.

Beyond these matters, it is worth noting what these figures are not showing us – they are not showing that the effect of external validation on individual projects accumulates over time. Projects receive only one externally reviewed final appraisal, then exit the dataset. Rather the event studies show us that, something, or things, is causing increasingly negative appraisals of final projects over time, while at the same time not affecting appraisals of ongoing projects.

With project fixed effects added the increased divergence between ongoing and final appraisals between 2019 and 2020 largely vanishes, suggesting perhaps that this change could be a product of (unobserved) project traits, possibly even deteriorating project quality, or falling aid capacity (for a discussion of these concerns with respect to Australian aid see, Moore, 2019). However, the divergence increases again in 2021. Australian aid program staff have advised us that 2021 (which is the 2020/21 financial year) is the first year any effect of COVID-19 might have been expected to show up in appraised project performance. The fact that such a dramatic fall in the appraised performance of completed projects occurs in this year is strongly suggestive: Covid may well have affected project performance, with its effect only being captured by the more rigorous external appraisal validation process. However, another possible candidate for the fall is a change in wording to the questions asked about effectiveness and efficiency in project appraisals which took effect in 2021. This change slightly differed between

ongoing and final projects too, meaning that if could have been the source of change in this year if the wording change influenced how aid program staff filled out their final appraisals, but not how they filled out the appraisals of ongoing projects. (We provide the wording of relevant questions from the appraisal forms in online Appendix 2).

In our view, the wording change is a possible but not particularly likely source of the 2021 fall in recorded completed aid project performance. The changes in wording are real and they do differ between questions asked in appraisals of final and ongoing reports. However, the differences between the two types of reports are comparatively minor, whereas the layout and format of the appraisal forms changed a lot for both ongoing and final appraisals in 2021, something that had no impact on the appraised performance of ongoing projects. What is more, the Australian aid program itself notes the significant challenges posed by the pandemic when discussing its performance in annual reports (DFAT, 2021a, 2022b). In 2021, for example, it stated that effectiveness and efficiency were lower and that, "This was primarily due to changes in the development context, including as a result of COVID-19" (DFAT, 2021a, p. 71). The balance of evidence suggests the pandemic did affect the performance of Australian aid, a fact that was only captured in final performance appraisals, and which was presumably only captured as a result of the more robust external appraisal validation process.

5 Practical significance: Papua New Guinea

Papua New Guinea is Australia's closest neighbour. It has a population of over 9 million people and is one of the poorest countries in the Pacific (Secretariat for the Pacific Community, 2023). It is the largest recipient of Australian aid (Development Policy Centre, 2023).

The one existing detailed study of Australian aid project effectiveness is based on data from before the period of external validation of appraisals. The authors of that paper noted that, "Of all the countries in the Pacific, the mean appraisal is second highest in Papua New Guinea, which will likely come as a surprise to anyone who has worked in that country's challenging context" (Wood et al., 2020, p. 174).

The surprise they note stems from the fact that Papua New Guinea is a very challenging place to deliver aid in. Governance is poor and violence an ongoing issue (Forsyth et al., 2023; May, 2022; Pieper, 2012; Reilly et al., 2014; Standish, 2007). Existing econometric work has found little evidence that aid promotes economic growth in Papua New Guinea (Feeny, 2005). It would be both unsurprising and understandable if Australian aid were underperforming in Papua New Guinea. It would also be useful to Australian aid policymakers if available aid performance data accurately reflected performance in its largest aid partner.

The first panel in Figure 6 shows the appraised performance of Australian aid projects based on the data in our dataset; it draws its data from all available appraisals that were not been subject to external validation. Each bar in the chart is a country. Unlike quote

above, which is solely focused on Pacific countries, all countries that Australia has had aid projects over AU $3 million are included in the chart. The mean score for that country is indicated by the bar's height. Papua New Guinea is labelled on the x-axis of the chart. The second panel is identical except that focuses only on data from externally validated appraisals.

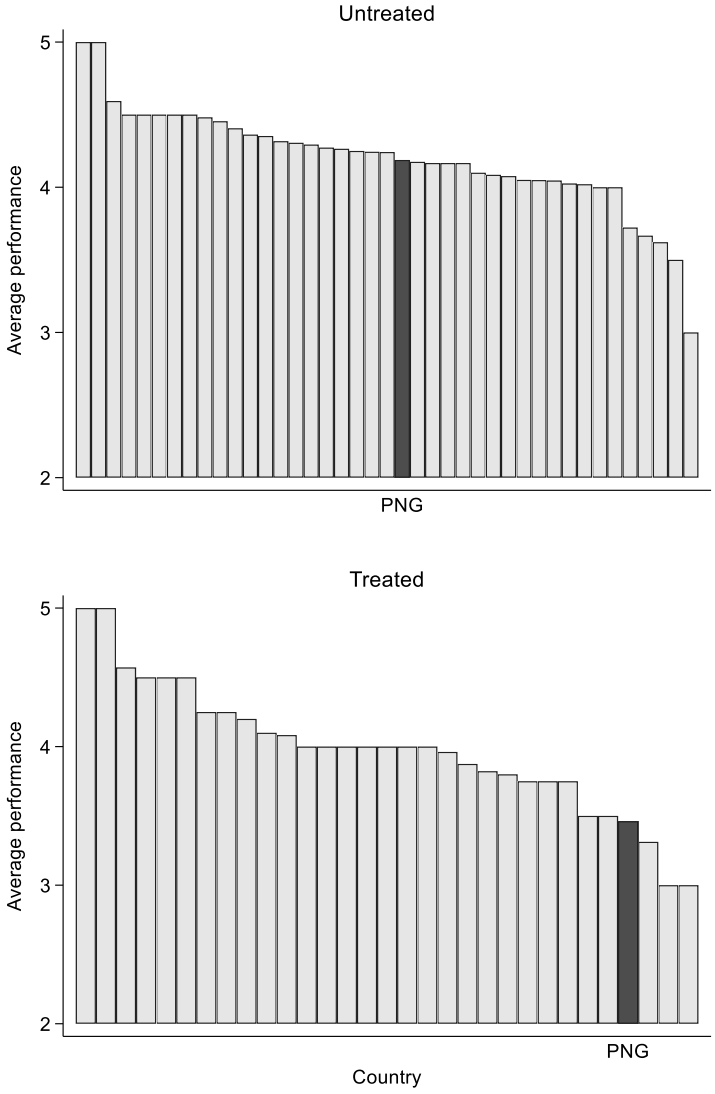
When using non-validated appraisals, Papua New Guinea is just below the median country globally. Among validated appraisals, Papua New Guinea's performance falls considerably and is the fourth lowest country globally.

In Online Appendix 3 we systematically test the effect of project validation on appraised performance in Papua New Guinea controlling for timing and appraisal type, as well as project traits. We do this to ensure that the differences evident in Figure 6 are not the product of some other factor such as average reported performance of all projects in Papua New Guinea falling over time regardless of whether their appraisals were validated or not. As the appendix shows, even with other factors controlled for, external validation clearly causes average appraised project performance to fall in Papua New Guinea.

If the 2019 change in project the appraisal of final Australian aid project assessments only led to lower aid success scores overall, this would still be useful to politicians and the public as it would presumably provide a more accurate sense of the overall performance of Australian aid. However, the Papua New Guinea case demonstrates something else: the more rigorous project appraisal validation process has helped

identify the underperformance of Australian aid in its most important partner. This has the potential to be useful to policymakers who can now, if they so wish, use the information to as the catalyst for a reconfiguration of their aid to Papua New Guinea.

Figure 6: Performance of projects in Papua New Guinea



Notes: Each bar in the chart is a country that received Australian aid projects, and where aid projects were large enough to be assessed. The height of the bars reflects the mean project assessment in that country. There are fewer projects in the treated pane because fewer countries where host to sufficiently large Australian completed aid projects in the period 2019 onwards than across the full 2014-2022 period. Papua New Guinea's bar is shaded and indicated with "PNG".

6 Discussion and conclusion

The changes how final project appraisals how were validated in 2019 led to a substantial fall in the reported performance of recently completed Australian aid projects. A marked fall in reported performance occurred as soon as the validation process was introduced, a further fall occurred in 2021. At the same time as these dramatic changes occurred, measured performance in ongoing project appraisals has not fallen at all, indeed it has risen slightly. The difference in differences analysis that we have reported on in this paper provides clear evidence that a shift to a more independent process involving validation of performance appraisals coordinated by a central evaluation unit and undertaken by external contractors was the source of much of the observed change. Although we cannot be as certain, there is also some evidence that the change led to more accurate reporting of the impact of the COVID-19 pandemic on the performance of Australian aid projects. It also appears to have led to more accurate assessment of project performance in Papua New Guinea.

As we discussed in our summation of existing research, existing evidence and theory provided us with some reason to believe that Australian aid program staff might have been overly generous when assessing the performance of their own projects. However, existing work provided little by the way of clear guidance of what to expect when a more independent system of validation was introduced involving external consultants contracted by the air program's evaluation unit reviewing staff-produced appraisals. It seemed quite possible that outsourcing the task of reviewing appraised performance would have no impact on appraised scores. It seemed all too possible that contractors

would not risk irritating the aid program that contracted them by routinely revising performance scores downwards. Yet scores were downgraded, which raises the question why?

Two explanations seem likely. The first involves the contractors. The contracting company is well-established and promotes itself as having a values-based ethic. Key personnel have a long history in the Australian aid community. Its evaluators are also members of the Australian Evaluation Society (Bluebird Consultants, n.d.). As a result, they are part of two normative communities: one in which effective aid is valued, the other in which sound evaluations are valued. This might possibly explain why the consultants have took a risk and opted to press for accurate project performance scores when validating appraisals.

The other possible explanation involves aid program staff themselves. In 2013, Australia's aid agency AusAID, was fully integrated into its foreign ministry, a move which has in the eyes of many reduced aid effectiveness (Moore, 2019; Wood et al., 2017). However, there are still professional aid workers in the aid program, including some who have now risen to senior positions in the foreign ministry. It is possible that there has been a growing desire within the aid program to improve aid quality, and as part of this, a desire to see more accurate project appraisals produced. What is more, external validation of project appraisals was coordinated by a specialist evaluation unit within the aid program, not country teams themselves. The evaluation unit may well have valued rigour in appraisals more highly, and may not have been inclined to

pressure the consultants in the same way country teams might have had they been the central point of contact with the consultants.

In addition, in May 2020, the Australian foreign ministry lowered the prominence afforded to project effectiveness scores in its high-level reporting. Average project performance became a "third tier" indicator accompanied by many other indicators and there was no specific target for the share of projects that needed to be appraised as successful (DFAT, 2021b). These reporting changes lessened the importance of both completed and ongoing project appraisals in DFAT's high-level reporting and so cannot have been the source of the difference in differences between ongoing and completed project performance that we have documented. However, this broader change in prominence may well have produced an environment in which it was felt to be safe to allow completed project performance scores to drop.

While one newspaper article covered the reported fall in Australian aid performance (Packham, 2023) that was the sum total of Australian media interest in the matter. The reputational cost for the aid program, and the politicians that it answers to, has, predictably, been very small. Such circumstances may well have further added to the willingness to allow the consultants to downgrade performance scores.

Determining which of these explanations is accurate will be an important subject for future research. Another question for future research will be about the broad state of validation processes across the various donors that appraise project performance and how much validation changes scores amongst different donors (for a, non-exhaustive,

list of countries which conduct project appraisals see: Honig, 2018). For those donors, such as the World Bank, which have validation systems in place, and where those systems regularly appear to change scores, it will be important to learn whether they change scores to an equal extent across all types of projects, or whether particular types of projects, or projects in particular countries, are changed more often.

For now though, in this paper we have demonstrated the clear impact that changes in the process of validating aid appraisals has had for the Australian aid program. We have also demonstrated that this impact has not been homogenous across projects. In Papua New Guinea, a particularly challenging country context, the impact was larger than elsewhere.

Superficially, this discovery may appear disappointing: Australian projects have not been as successful as previously reported in the country that receives the largest share of Australian aid. And yet the case of Papua New Guinea also demonstrates the potential of more rigorous aid appraisals. It should be easier for aid programs to improve aid effectiveness if they start by openly facing up to failures.

7 References

Angrist, J., & Pischke, J.-S. (2009). *Mostly Harmless Econometrics: an empiricist's companion*. Princeton University Press.

Ashton, L., Friedman, J., Goldemberg, D., Hussain, M. Z., Kenyon, T., Khan, A., & Zhou, M. (2022). A Puzzle with Missing Pieces: Explaining the Effectiveness of World Bank Development Projects. *The World Bank Research Observer*, *38*(1), 115-146. https://doi.org/10.1093/wbro/lkac005

Bertrand, M., Duflo, E., & Mullainathan, S. (2004). How Much Should We Trust Differences-In-Differences Estimates?*. *The Quarterly Journal of Economics*, *119*(1), 249-275. https://doi.org/10.1162/003355304772839588

Bluebird Consultants. (n.d.). *Bluebird consultants*. Retrieved 14/8/2023 from https://www.bluebirdconsultants.com.au/

Briggs, R. C. (2019). Results from single-donor analyses of project aid success seem to generalize pretty well across donors [journal article]. *The Review of International Organizations*. https://doi.org/10.1007/s11558-019-09365-x

Bulman, D., Kolkma, W., & Kraay, A. (2017). Good countries or good projects? Comparing macro and micro correlates of World Bank and Asian Development Bank project performance [journal article]. *The Review of International Organizations*, *12*(3), 335-363. https://doi.org/10.1007/s11558-016-9256-x

Denizer, C., Kaufmann, D., & Kraay, A. (2013). Good countries or good projects? Macro and micro correlates of World Bank project performance. *Journal of Development*

Economics, *105*, 288-302. https://doi.org/https://doi.org/10.1016/j.jdeveco.2013.06.003

Development Policy Centre. (2023). *Australian Aid Tracker - Destinations*. Retrieved 21/7/2023 from https://devpolicy.org/aidtracker/destinations/

DFAT. (2019). *Performance of Australian Aid 2017-18*. https://www.dfat.gov.au/sites/default/files/performance-of-australian-aid-2018-19.pdf

DFAT. (2020a). *Aid quality check template*. Retrieved 19/07/2023 from https://www.dfat.gov.au/sites/default/files/aid-quality-check-template.pdf

DFAT. (2020b). *Final aid quality check template*. Retrieved 19/07/2023 from https://www.dfat.gov.au/sites/default/files/final-aid-quality-check-template.pdf

DFAT. (2020c). *Performance of Australian Aid 2018-19*. https://www.dfat.gov.au/sites/default/files/performance-of-australian-aid-2018-19.pdf

DFAT. (2021a). *Annual Report: 2020-21*. https://www.dfat.gov.au/sites/default/files/dfat-annual-report-2020-21.pdf

DFAT. (2021b). *Australia's Development Program – performance assessment*.

DFAT. (2022a). *Aid Programming Guide*. https://www.dfat.gov.au/sites/default/files/aid-programming-guide.pdf

DFAT. (2022b). *Annual Report: 2021-22*. https://www.dfat.gov.au/sites/default/files/dfat-annual-report-2021-22.pdf

Donald, S. G., & Lang, K. (2007). Inference with Difference-in-Differences and Other Panel Data. *The Review of Economics and Statistics*, *89*(2), 221-233. http://www.jstor.org/stable/40043055

Feeny, S. (2005). The impact of foreign aid on economic growth in Papua New Guinea. *The Journal of Development Studies*, *41*(6), 1092-1117. https://doi.org/10.1080/00220380500155403

Feeny, S., & Vuong, V. (2017). Explaining aid project and program success: findings from Asian Development Bank interventions. *World Development*, *90*, 329–343. https://doi.org/https://doi.org/10.1016/j.worlddev.2016.10.009

Forsyth, M., Kipongi, W., & Gibbs, P. (2023). *How to address escalating violence in PNG*. Retrieved 21/7/2023 from https://devpolicy.org/how-to-address-escalating-violence-in-png-20230714/

Gibson, C. C. A., Krister, Ostrom, E., & Shivakumar, S. (2005). *The samaritan's dilemma: the political economy of development aid*. Oxford University Press.

Goodman-Bacon, A. (2021). Difference-in-differences with variation in treatment timing. *Journal of Econometrics*, *225*(2), 254-277. https://doi.org/https://doi.org/10.1016/j.jeconom.2021.03.014

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

Honig, D. (2019). When Reporting Undermines Performance: The Costs of Politically Constrained Organizational Autonomy in Foreign Aid Implementation. *International Organization*, *73*(1), 171-201. https://doi.org/10.1017/S002081831800036X

Honig, D. (2020). Information, power, and location: World Bank staff decentralization and aid project success. *Governance*, *33*(4), 749-769. https://doi.org/https://doi.org/10.1111/gove.12493

House of Commons International Development Committee. (2020). *Effectiveness of UK aid: interim findings*. https://committees.parliament.uk/publications/1373/documents/12634/default/

Isham, J., Kaufmann, D., & Pritchett, L. H. (1997). Civil Liberties, Democracy, and the Performance of Government Projects. *The World Bank Economic Review*, *11*(2), 219-242. https://doi.org/10.1093/wber/11.2.219

Kilby, C. (2000). Supervision and performance: the case of World Bank projects. *Journal of Development Economics*, *62*(1), 233-259. https://doi.org/https://doi.org/10.1016/S0304-3878(00)00082-1

Kilby, C., & Michaelowa, K. (2019). What Influences World Bank Project Evaluations? In N. Dutta & C. R. Williamson (Eds.), *Lessons on Foreign Aid and Economic Development: Micro and Macro Perspectives* (pp. 109-150). Palgrave Macmillan. https://doi.org/https://doi.org/10.1007/978-3-030-22121-8

Martens, B., Mummert, U., Murrell, P., Seabright, P., & Ostrom, E. (2002). *The Institutional Economics of Foreign Aid*. Cambridge University Press.

Martins, B. (2002). The role of evaluation in foreign aid programmes. In B. Martens, U. Mummert, P. Murrell, P. Seabright, & E. Ostrom (Eds.), *The Institutional Economics of Foreign Aid* (pp. 154-177). Cambridge University Press.

May, R. (2022). *State and Society in Papua New Guinea, 2001–2021*. ANU E-Press.

Moore, R. (2019). *Strategic Choice A future-focused review of the DFAT-AusAID integration*. http://devpolicy.org/publications/reports/DFAT-AusAIDIntegrationReview-FullVersion.pdf

Packham, B. (2023). DFAT Boost for 'ineffective' diplomats. *The Australian*. https://www.theaustralian.com.au/nation/politics/plan-dfat-boost-for-ineffective-diplomats/news-story/72f9f31c7227248016536023c6a05b6a

Pieper, L. (2012). Deterioration of public administration in Papua New Guinea – views of eminent public servants. *Development Policy Centre Discussion Paper*, *2012*(23), 1-19.

Reilly, B., Flower, S., & Brown, M. (2014). Political Governance and Service Delivery in Papua New Guinea (PNG): A Strategic Review of Current and Alternative Governance Systems to Improve Service Delivery. *National Research Institute Discussion Papers*, *2014*. https://tinyurl.com/3tjke3x2

Seabright, P. (2002). Conflict of objectives and task allocation in aid agencies. In B. Martens, U. Mummert, P. Murrell, P. Seabright, & E. Ostrom (Eds.), *The Institutional Economics of Foreign Aid* (pp. 34-68). Cambridge University Press.

Secretariat for the Pacific Community. (2023). *Pocket Summary*. Retrieved 21/7/2023 from https://stats-data-viewer.pacificdata.org/?chartId=205

Standish, B. (2007). The dynamics of Papua New Guinea's democracy: an essay. *Pacific Economic Bulletin*, *22*(1), 135-157.

Wood, T., Burkot, C., & Howes, S. (2017). Gauging Change in Australian Aid: Stakeholder Perceptions of the Government Aid Program. *Asia & the Pacific Policy Studies*, *4*(2), 237-250. https://doi.org/10.1002/app5.173

Wood, T., Dornan, M., & Muller, S. (2021). *Change and continuity in Australian aid: what the aid flows show*.

Wood, T., Otor, S., & Dornan, M. (2020). Australian aid projects: what works, where projects work, and how Australia compares. *Asia & the Pacific Policy Studies*, *7*(2), 171–186.

Wood, T., Otor, S., & Dornan, M. (2022). Why are aid projects less effective in the Pacific? *Development Policy Review*, *40*(3), e12573. https://doi.org/https://doi.org/10.1111/dpr.12573

World Bank. (2021). *Results and Performance of the World Bank Group 2021.* https://ieg.worldbankgroup.org/sites/default/files/Data/Evaluation/files/RAP2021.pdf

Online Appendices

Online Appendix 1

Table A1: Difference in covariates between by report type, and across periods

Table A1 shows the mean values of possible traits (size, duration, location and sector) which could independently affect project performance. The difference between these values in ongoing and final reports is shown for both the pre-2019 period and the 2019 onwards period, the difference in these differences across periods is also shown.

	Pre-2019 ongoing	Pre-2019 final	Diff	p-value	2019 onwards ongoing	2019 onwards final	Diff	p-value	Diff in diff	p-value
Budget(ln)	16.76	16.58	0.17	0.01	16.94	16.68	0.26	0.00	-0.09	0.45
Duration	2,432	2,267	165	0.01	2,697	2,667	30	0.74	135	0.22
Pacific	0.40	0.34	0.06	0.04	0.35	0.33	0.03	0.52	0.04	0.46
Sector										
Agriculture	0.09	0.07	0.03	0.13	0.07	0.10	-0.03	0.23	0.05	0.06
Resilience	0.10	0.15	-0.05	0.02	0.12	0.12	0.00	0.94	-0.05	0.14
Education	0.21	0.18	0.03	0.19	0.15	0.16	-0.02	0.56	0.05	0.21
Governance	0.24	0.25	-0.01	0.64	0.28	0.25	0.03	0.45	-0.04	0.37
Other	0.05	0.07	-0.02	0.15	0.07	0.05	0.01	0.52	-0.03	0.18
Health	0.13	0.18	-0.05	0.03	0.14	0.10	0.04	0.16	-0.09	0.01
Economic	0.18	0.11	0.07	0.00	0.17	0.21	-0.04	0.22	0.10	0.01

Online Appendix 2: Wording of effectiveness and efficiency between before and after change

	Final appraisal before 2021	Final appraisal 2021 onwards
Effectiveness	Have we achieved the outputs and outcomes that we expected over the lifetime of the investment?	Did the investment achieve the end-of-investment outcomes that were expected over the lifetime of this investment?
Efficiency	Did the investment make appropriate use of Australia's and our partners' time and resources to achieve outcomes?	Did the investment make appropriate and efficient use of Australia's and our partners' time and resources to achieve the end-of-investment outcomes?
	Ongoing appraisal before 2021	**Ongoing appraisal 2021 onwards**
Effectiveness	Are we achieving the outputs and outcomes that we expected?	Has the investment achieved the outputs and outcomes expected at this time?
Efficiency	Is the investment making appropriate use of Australia's and our partners' time and resources to achieve outcomes?	Is the investment making an efficient use of Australia's and our partners' time and resources to achieve outputs and expected outcomes?

Online Appendix 3: Testing changes in the performance of projects in Papua New Guinea

In Table A2 the key coefficient of interest is that associated with the interaction of Papua New Guinea and external assessment (PNG*External assess). Model 1 is a simple regression with the interaction as well as individual coefficients for PNG and externally validated assessments. In Models 2 and 3 additional controls are added for the post 2018 period and report types. These controls account for the possibility that project performance in Papua New Guinea has gotten worse in recent years and for the possibility that final project appraisals have always been worse in Papua New Guinea than ongoing appraisals. In all models the coefficient for the interaction term is negative, comparatively large and statistically significant, reflecting the fact that external validation had a larger than average negative impact on the appraised performance of projects in Papua New Guinea.

Table A2: Impact of external validation on appraised project performance in Papua New Guinea

	(1)	(2)	(3)
PNG * External assess	-0.36**	-0.35**	-0.34**
	(0.15)	(0.15)	(0.15)
PNG	-0.11	-0.12	-0.07
	(0.07)	(0.07)	(0.08)
External Assess	-0.37***	-0.40***	-0.39***
	(0.06)	(0.08)	(0.07)
After 2018		0.10***	0.08**
		(0.03)	(0.03)
Final		-0.03	-0.03
		(0.04)	(0.04)
Constant	4.30***	4.26***	3.07***
	(0.02)	(0.02)	(0.27)
Project Controls	No	No	Yes
Observations	3096	3096	3096

Robust standard errors clustered at the project level.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Online Appendix 4: Binary Dependent Variable

One possible issue with our analysis is our use of the mean of the (effectively ordinal) effectiveness and efficiency scores as our dependent variable in regression models. As a robustness test we conducted further analysis using a binary dependent variable of satisfactory effectiveness or not. In this analysis projects which were coded as scoring 4 or more for effectiveness were coded as being satisfactory. Those which did not were coded as unsatisfactory. We used this categorisation based on DFAT's own definition of satisfactory project effectiveness (DFAT, 2021b).

Because non-linear models are problematic when used in difference in difference analysis. We ran the resulting regressions as linear probability models estimated using OLS. In all analysis standard errors were clustered at the project level. The results are shown in the tables below. All results fit with the findings presented in the main body of the text.

Table A2: Binary Dependent Variable: two by two difference in differences

	(1) Basic	(2) Project controls
Diff in Diff	-0.21***	-0.20***
	(0.04)	(0.04)
Final	-0.00	0.00
	(0.02)	(0.02)
After 2018	0.07***	0.07***
	(0.01)	(0.01)
Project Controls	No	Yes
Observations	3096	3096

Regression models are OLS with a binary dependent variable based on satisfactory project effectiveness.
Robust standard errors clustered at the project level.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table A3: Binary Dependent Variable: project fixed effects difference in differences

	(1) Project FE
Diff in Diff	-0.21***
	(0.04)
Project FE	Yes
Observations	2781

Regression models are OLS with a binary dependent variable based on satisfactory project effectiveness.
Robust standard errors clustered at the project level.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table A4: Binary dependent variable: event studies

	(1) Basic	(2) Project controls	(3) Project FE
2014	-0.08	-0.09	0.00
	(0.06)	(0.06)	(.)
2015	-0.02	-0.03	-0.06
	(0.05)	(0.05)	(0.05)
2016	-0.08	-0.08	-0.03
	(0.06)	(0.06)	(0.06)
2017	-0.10*	-0.11*	-0.07
	(0.06)	(0.06)	(0.06)
2019	-0.20***	-0.19***	-0.21**
	(0.07)	(0.07)	(0.08)
2020	-0.23***	-0.22***	-0.08
	(0.08)	(0.08)	(0.09)
2021	-0.34***	-0.35***	-0.27***
	(0.08)	(0.08)	(0.08)
2022	-0.30***	-0.31***	-0.21***
	(0.07)	(0.07)	(0.07)
Project Controls	No	Yes	No
Project FE	No	No	Yes
Observations	3096	3096	2781

Regression models are OLS with a binary dependent variable based on satisfactory project effectiveness.
Robust standard errors clustered at the project level.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$