DEVELOPMENT
POLICY CENTRE

Why are two in five Australian aid investments rated unsatisfactory on completion?

An investigation into recent trends in Australian aid performance assessments

Stephen Howes, Huiyuan Liu, Terence Wood and Cameron Hill

5 May 2023

Authors

Dr 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.

Huiyuan (Sharon) Liu is a Research Officer at the Development Policy Centre, working in the area of economic development.

Dr Terence Wood is a Research Fellow at the Development Policy Centre and Professor of Economics at the Crawford School of Public Policy, at The Australian National University

Dr Cameron Hill is a Senior Research Officer at the Development Policy Centre. He has previously worked with DFAT, the Parliamentary Library and ACFID.

The authors thank DFAT for providing them with the data and with comments on earlier drafts. They thank Matthew Morris for his analysis and comments, others who provided comments, and Karen Downing for her editing. They also thank the Bill and Melinda Gates Foundation for generous funding support. All views and any remaining errors are their own.

DEVELOPMENT
POLICY CENTRE

Published 5 May 2023

Development Policy Centre
Crawford School of Public Policy
ANU College of Asia and the Pacific
The Australian National University
devpolicy.crawford.anu.edu.au
devpolicy.org

devpolicy@anu.edu.au

Executive Summary

Abstract

In 2021 and 2022, two in five completed aid investments were rated as unsatisfactory by the Department of Foreign Affairs and Trade (DFAT), up from one and a half or even one in five just a few years earlier. Since 2020, the average completed investment rating has been less than satisfactory for both effectiveness and efficiency. There is also a large and growing disconnect between the assessments of ongoing and completed investments. This report documents these recent trends in reported Australian aid performance and shows that a major factor behind them is methodological. DFAT is to be commended for having moved in 2019 to a more independent system of ratings for completed investments. However, further reforms are needed given the significantly lower ratings that have resulted. We recommend that DFAT re-establish its performance-oversight architecture – in the form of the Office of Development Effectiveness and the Independent Evaluation Committee – and strengthen other parts of its performance review and reporting system.

Trends

The data reveal a clear downward trend in the rating of completed investments. Using DFAT's current methodology, from 2014 to 2018, on average 77% of completed investments were rated as satisfactory, but from 2019 to 2022 on average only 66% were. For the last two years, this figure is only about 60%. On DFAT's six-point scale, an investment must get a score of at least four to be regarded as satisfactory. But the average rating for completed investments for the period 2019-2022 has fallen to 3.91 for effectiveness and 3.87 for efficiency: that is, the average completed investment rating is now less than satisfactory. Current performance levels can only be described as poor, whether in absolute terms, or compared to past reported performance or to levels reported by other development agencies with rigorous rating systems, such as the World Bank.

We stress the ratings of completed investments, since DFAT gives them primacy in its own performance framework, as do other development agencies, such as the World Bank. However, DFAT also rates ongoing investments annually, and these ratings have improved over time. As a result, the disconnect between ongoing and final ratings – measured as the share of investments regarded as satisfactory using DFAT's definition – has risen sharply. It was only 3 percentage points between 2014 and 2018, but it rose to 24 percentage points on average between 2019 and 2022, and to 30 percentage points in 2021 and 2022.

Explaining the trends

These trends are analysed using decompositions and regressions, including difference-in-differences analysis. One main factor emerges as explaining both trends – the decline in the performance of completed investments, and the growing disconnect – namely, the shift to a more independent rating system in 2019 for completed investments. This change has reduced the "optimism bias" inherent in self-awarded ratings, and so has led to a decline in the rating of completed investments. However, since it did not apply to them, it has not affected the rating of ongoing investments. As a result, the gap between the two has quadrupled, and is now not much smaller than that observed by the World Bank, whose investments are rated at completion by its independent evaluation agency.

The evidence is also consistent with COVID-19 having a negative impact on the performance of investments, but only for final ratings. The argument here is that the more independent rating system for completed investments allowed for a franker assessment of the impact of the pandemic on aid performance. However, there was also a subtle change in rating instructions in 2021 and 2022, and we are not able to disentangle the impact of this from that of the pandemic.

There are no prominent regional or sectoral explanations for the decline in final ratings.

We also argue that aspects of DFAT's performance management system have put upward pressure on ratings, especially of ongoing investments, and that this also helps explain the growing disconnect. In particular, the additional management effort required if an investment is rated as unsatisfactory helps explain why ongoing ratings continued to increase even after the target for them was achieved and then abandoned.

Recommendations

The shift to a more independent rating system in 2019 for completed investments is a good thing as it enables more frank and objective assessments. Reliance on independent ratings is common for multilaterals but not for bilateral aid agencies. DFAT is to be commended for moving in this direction, and for shifting the focus of performance measurement from ongoing to final assessments. At the same time, it is evident that these positive reforms have had some adverse and unanticipated consequences. The report concludes with six recommendations to address them.

First, DFAT should continue with its more independent rating system introduced in 2019. The introduction of the new methodology was laudable from both an aid management and an aid transparency perspective. Abolishing it would give a lift to reported performance, but at the cost of credibility.

Second, the Office of Development Effectiveness (ODE) and Independent Evaluation Committee (IEC) should be re-established. The problems identified in this report have largely occurred since the abolition

of ODE and its oversight body, the IEC. One of the roles of ODE, prior to its abolition in 2020, was to assess "DFAT's internal performance management system". The absence of any entity playing that role (of assessing the adequacy of the performance system as distinct from managing it) has allowed the current situation to develop. In short, the trends revealed in this report, and the lack of any response to them to date by DFAT, make a compelling case for the reintroduction of ODE and the IEC as aid effectiveness champions and watchdogs within the department.

Third, DFAT needs to take its own performance assessments more seriously, and reverse some of the recent decline. Of course, one would not expect all investments to succeed, but in each of the last two years, the annual value of unsatisfactory investments has been half a billion dollars. This is, by any standard, of a size significant enough to warrant corrective measures.

Fourth, DFAT should guard against solving the problem by grade inflation. The easiest way to lift the satisfactory rate would be to grade more easily. There is no easy way to prevent this, but we recommend against adopting a performance target, and in favour of the retention of consultants to review and finalise the ratings of completed investments.

Fifth, a communication effort is needed. There may be a fear that a higher profile for the low reported performance of completed aid investments could lead to a backlash and a loss of support for Australian aid**.** However, with the change in government and new aid strategy there is an opportunity to draw a line in the sand and get on the offensive. Reporting a low level of performance without explanation is a great vulnerability. A significant number of large and very important investments have been declared to be unsatisfactory, including some former ministerial flagships. There is no explanation in the public domain for why individual investments have been rated unsatisfactory, and only a sentence or two in the annual report in relation to the portfolio as a whole. DFAT should publish all its final rating reports. In addition, any verdict of unsatisfactory performance of a completed investment given in an annual country performance report (which is where the ratings are currently provided) should be accompanied by a statement explaining the reason for the verdict. At the portfolio level, a detailed public performance analysis is needed.

Sixth, ongoing ratings are meaningless due to the large disconnect with final ratings, and the performance assessment system for ongoing investments should be overhauled.

Contents

Introduction

In 2021 and 2022, two in five completed aid investments were rated as unsatisfactory by the Department of Foreign Affairs and Trade (DFAT), up from one and a half or even one in five just a few years earlier. Since 2020, the average completed investment rating has been less than satisfactory for both effectiveness and efficiency. Analysis also reveals a large and growing disconnect between the assessments of ongoing and completed investments, with the former showing improved performance over time. This report documents these recent trends in reported Australian aid performance, tries to explain them, considers the implications, and puts forward some measures to reverse the decline.

DFAT's aid performance system

The system AusAID developed and now DFAT uses to rate the performance of its aid investments (or projects) has undergone some changes over the years but, for the period covered, has consistently required all investments above a certain size to be rated on a scale of 1 to 6 for effectiveness, efficiency and various other dimensions of performance, where 4 or above indicates satisfactory or better performance and 3 or below indicates less than satisfactory performance or worse (Table 1). For convenience, we refer to investments that get a score of 4 or better as satisfactory, and investments that get a score of 3 or less as unsatisfactory.

Table 1: DFAT's 6-point rating scale

Satisfactory (4, 5 and 6)
6 = Very good; satisfies criteria in all or almost all areas.
5 = Good; satisfies criteria in most areas.
4 = Adequate; on balance, satisfies criteria; does not fall in any major area.
Less than satisfactory (1, 2 and 3)
3 = Less than adequate; on balance does not satisfy criteria but does not fail in any major area
2 = Poor; does not satisfy criteria in major areas
1 = Very poor; does not satisfy criteria in many major areas.

Source: DFAT Aid Program Performance Report South-East Asia Regional Program, November 2015, last page. Note: the exact definitions for the 1-6 scores may have changed slightly over time.

All aid investments managed by DFAT with a value of at least $3 million are rated every year, either with an ongoing rating[1] or, if the investment has concluded, a final rating.[2] An investment is only rated once in

[1] Investment Monitoring Ratings (IMRs); before 2021, AQCs or Aid Quality Checks.
[2] Final Investment Monitoring Ratings (FIMRs); before 2021, FAQCs or Final Aid Quality Checks.

a given year, either through an ongoing rating or through a final rating, but not both. The same type of questions provide the basis for ratings of effectiveness and efficiency for both ongoing and completed investments. Some subtle changes were made in the ratings instructions in 2021, when DFAT moved from a system based on what were called AQCs (Aid Quality Checks) to one based on what are now called IMRs (Investment Monitoring Reports). The old and new ratings questions are slightly different. Both are shown in Table 2. We return to the possible influence of the changes in wording later in the report.

Table 2: Questions against which effectiveness and efficiency of aid investments are assessed for ongoing and completed investments

	Ongoing	Completed
Effectiveness	Are we achieving the outputs and outcomes that we expected? (Has the investment achieved the outputs and outcomes expected at this time?)	Have we achieved the outputs and outcomes that we expected over the lifetime of the investments? (Did the investment achieve the end-of-investment outcomes that were expected over the lifetime of this investment?)
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?)	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 partners' time and resources to achieve the end-of-investment outcomes?)

Source: Final rating (FAQC/FIMR) templates and ongoing rating (AQC/IMR) templates, with IMR/FIMR in brackets. See footnotes 1 and 2 for these acronyms; the former (AQC/FAQC) were replaced by the latter (IMR/FIMR) in 2021.

DFAT has provided us with performance ratings for ongoing and completed investments from 2014 (2013-14) to 2022 (2021-22). Earlier versions of this data set have already been used for research, with a focus on comparing the performance of Australian aid to that of other donors[3] and on aid to the Pacific.[4]

DFAT's ratings include both emergency (humanitarian) investments and development investments. There is a separate rating template for the former. To keep the analysis simple, we analyse only development investments in this report.[5] For the period of analysis, 93% of the investments worth 95% of the portfolio are development investments, and therefore covered by this report.[6]

There are 3,080 (development) ratings over the nine-year period: 2,586 ongoing ratings and 494 final ratings. Over this period of nine years, on average every year 287 ongoing investments and 55 completed ones were given a rating (Figure 1). These investments were worth, on average, respectively, $11.8 and $1.8 billion (Figure 2). The annual value of investments covered in this report is about 40% of the annual

[3] Wood, Otor and Dornan, (2020), "Australian aid projects: what works, where projects work and how Australia compares", *Asia and Pacific Policy Studies*, *7(3)*, 1–16.
[4] Wood, Otor and Dornan, (2021), "Why are aid investments less effective in the Pacific?" *Development Policy Review, 40(3),* 1–21.
[5] In particular, it can be difficult to determine whether an emergency investment is completed or still ongoing.
[6] We consider investments with at least one humanitarian rating in any given year to be emergency investments and exclude them from our analysis.

value of the Australian aid program. Projects of other departments, and core funding for multilaterals are not rated; and, as mentioned, humanitarian investments are excluded.

Figure 1: Annual number of ongoing and completed investments assessed

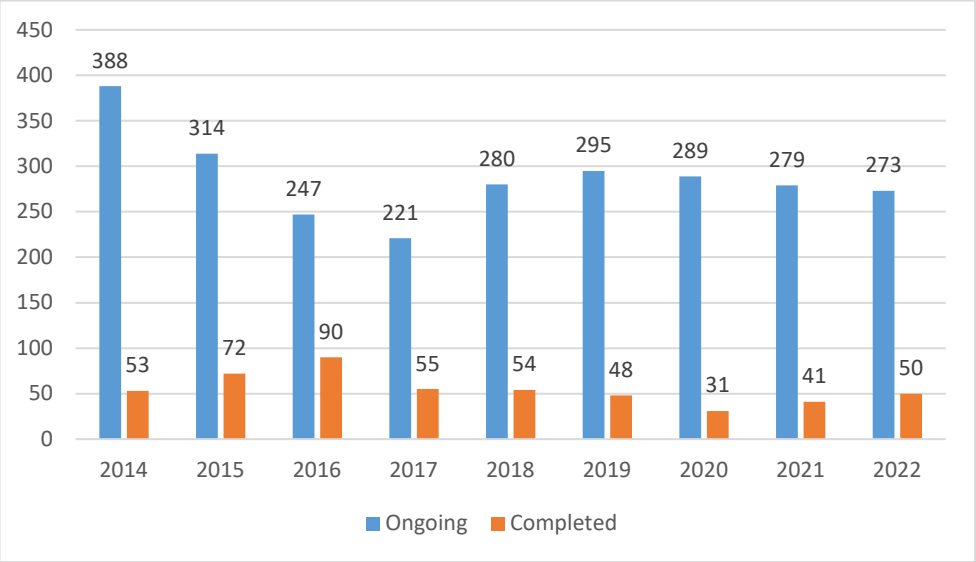
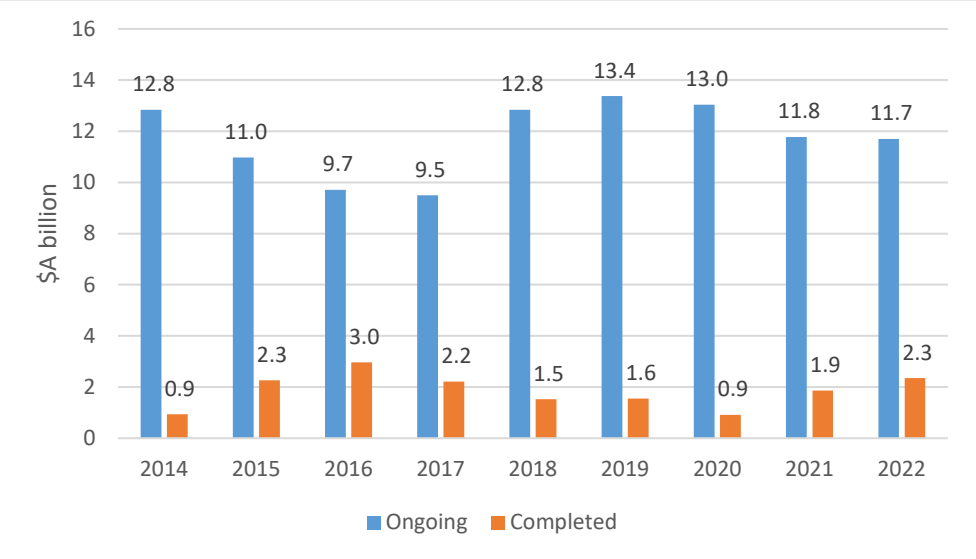


Figure 2: Annual total value of ongoing and completed investments assessed

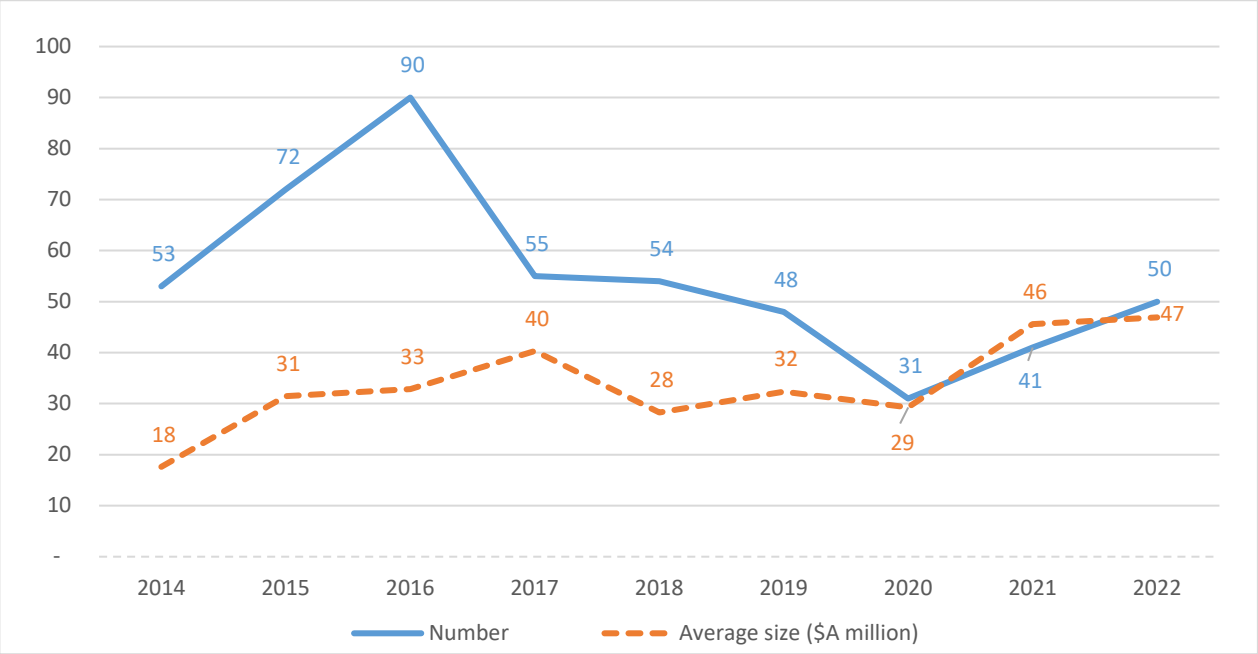


Although we also consider assessments of ongoing investments, the main focus of this report is on the assessment of completed investments. This is for several reasons. First, DFAT itself now gives primacy to the assessment of completed investments when reporting the performance of its portfolio. Second, as we discuss, the assessments of completed investments are now more independent. Third, it makes sense to look at investments as a whole rather than progress reports. Fourth, completed investments are typically the focus of investment performance assessments for development agencies other than DFAT (e.g., the World Bank).

The next graph (Figure 3) shows the annual number and average value of completed investments. For the period as a whole, these go in opposite directions (analysis of ongoing investments gives the same result). The number of investments assessed falls sharply from 2016 to 2020 and then slightly recovers. Comparing the first five to the second four years, the average number of investments assessed declines from 65 in 2014-2018 to 43 in 2019-2022. In contrast, the average value increases from $30 to $39 million.

The decline in the number of investments and the increase in their average value reflect the fact that DFAT engaged in a process of consolidation, with a target adopted in 2014 to reduce the number of investments by 20% by 2017. This was achieved, with the number of investments in fact reduced by 32% by 2018.[7]

Figure 3: Number and average value of completed investments

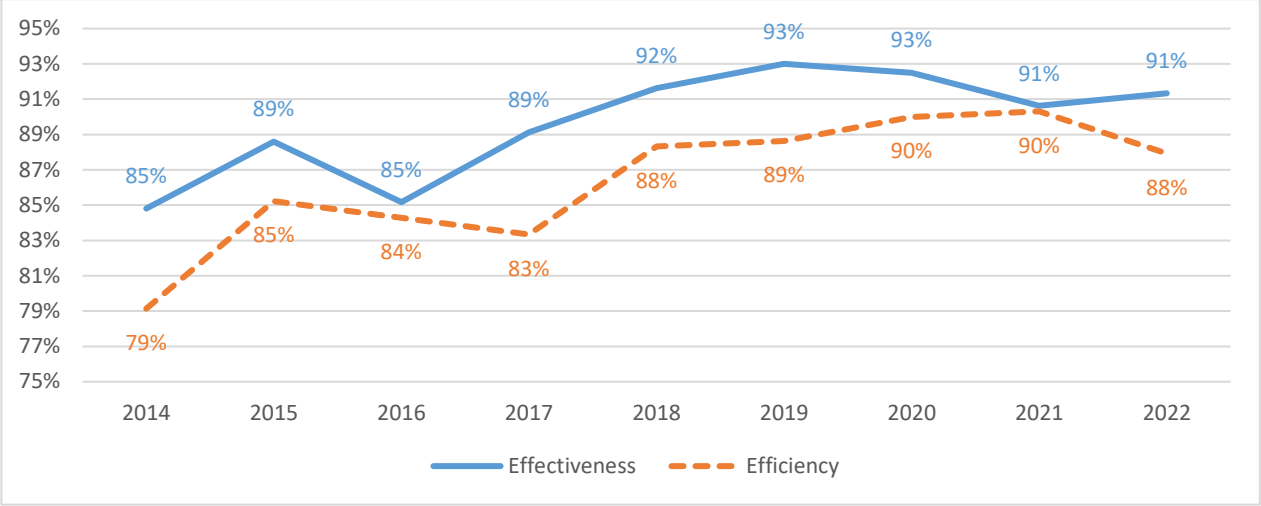


Trends in performance ratings

DFAT's 2014 performance framework included the target that the share of investments with satisfactory efficiency ratings and the share with satisfactory effectiveness ratings should both be at least 85%. Performance against this target was measured by looking at all ratings – for both ongoing and completed investments. The target has been met every year from 2018 onwards (Figure 4). The overall performance trend is positive, with average efficiency ratings reaching 90% in 2020 and 2021 and average effectiveness ratings crossing the 90% threshold in 2018. For the last four years, on average 92% of investments have been assessed as effective, a remarkable result.

[7] DFAT, (2018), *The performance of Australian Aid 2017-18*, Table 1.

Figure 4: Percentage of investments with satisfactory ratings for effectiveness and efficiency: all investments



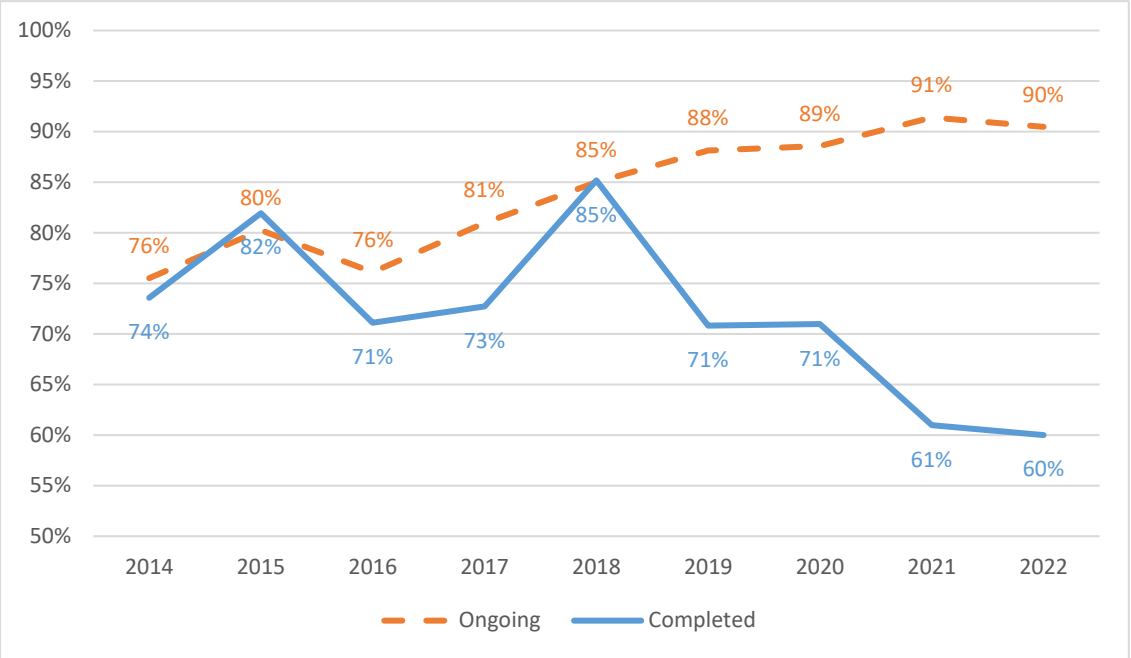
In May 2020, DFAT adopted a new performance framework as part its new strategy developed in response to COVID-19. Among other things, this framework has shifted the focus from **all** to **completed** investments, and slightly changed the way in which 'satisfactory' is defined. The new measure is the "percentage of completed investments assessed as satisfactory against both effectiveness and efficiency quality criteria".[8] There is no target for this indicator, which we label the satisfactory indicator. (The 'satisfactory rate' is the share of investments rated satisfactory, thus defined.)

Figure 5 plots results against the new standard for both completed investments (as per the DFAT definition) and, for comparison, for ongoing investments. For ongoing investments, the new indicator gives a very similar result to the old indicator (Figure 4): there is a clear upward trend. (The two trends are similar because ongoing ratings make up the great bulk of all ratings.) But the percentage of completed investments assessed as satisfactory has declined over time. In 2021 and 2022, only three in five investments are assessed as satisfactory on completion.[9]

[8] This is one of a number of third tier indicators of "operational and organisational effectiveness", sitting below second tier results that are meant to be directly attributable to Australian aid and first tier results that set out the Indo-Pacific development context. Other third-tier quantitative indicators are: percentage of investments effectively addressing gender issues; and percentage of investments effectively addressing disability inclusive development. There are no targets attached to these indicators either.

[9] The DFAT *Annual Report 2021-22* gives the satisfactory rate of 85% for all investments, and 62% for completed investments (p. 66). This is only slightly different from the 86% and 60% figures that we obtain. The *Annual Report 2020-21* figures are 88% and 61% respectively (p. 71), the same as ours.

Figure 5: Percentage of satisfactory investments: ongoing and completed investments



Note: Investments are rated as satisfactory, if they are rated as satisfactory or better (a score of 4 or more) with regards to both effectiveness and efficiency.

2021 and 2022 were pandemic years, which was a difficult time, and this might help explain the decline in the satisfactory rate for completed investments for these two years (note, however, that ongoing investments are not thus affected). If we take those two years out, the performance of completed investments doesn't show a strong overall decline as measured by Figure 5. The satisfactory rate bounces around between 71% and 85%, but there is no clear pre-COVID downwards trend.

One problem with this indicator is its binary nature. It takes no account of the extent to which investments do better (or worse) than pass. Rather, this indicator will only track whether the average quality of investments is getting worse (or better) to the extent that more investments are unsatisfactory (or satisfactory). Half of all investments get an average effectiveness-efficiency score of 4.5 of better, and 10% get a score of 3 or less (see Annex Figure 1). So, for more than half the investments average improvements or deteriorations of one in either effectiveness or efficiency (but not both) will have no impact on the satisfactory score. Such changes though are certainly important when considering overall performance.

An alternative indicator is the average score given to investments, whether for effectiveness or efficiency. This is influenced by all investment ratings and is shown in Figure 6.

Figure 6: Average ratings for efficiency and effectiveness for completed and ongoing investments

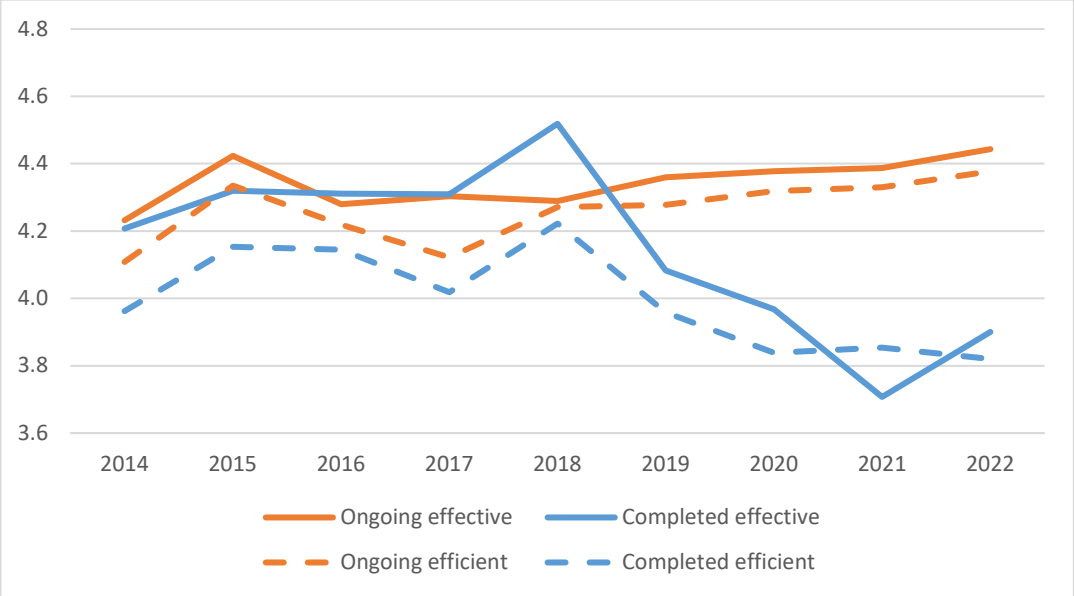


Figure 6 does show a pre-COVID downward trend in the average efficiency and effectiveness scores for completed investments (the blue lines). Such scores are level or even increasing to 2018, but fall rapidly thereafter (Table 3). For the two years before the pandemic (2019-2020), average effectiveness hits the satisfactory threshold (4.03), while average efficiency fails to meet this threshold (3.90).[10] Indeed, most of the downward trend is pre-COVID. This can be seen in Table 3, where the fall in the intermediate period (2019-2020) compared to the first (2014-2018) is greater than the fall in the final period (2021-2022) compared to the intermediate one. (For example, for effectiveness, a fall of 0.3 prior to COVID-19, and a fall of 0.23 during COVID-19.) So, the decline in the performance of completed investments is not just a pandemic story.

Table 3: Average ratings for efficiency and effectiveness for completed and ongoing investments, sub-period averages

	Ongoing		Completed	
Period	**Effective**	**Efficient**	**Effective**	**Efficient**
2014-28	4.31	4.21	4.33	4.10
2019-20	4.37	4.30	4.03	3.90
2021-22	4.42	4.35	3.80	3.84
Average	4.34	4.26	4.15	4.00

Note: These are unweighted annual averages, and thus may differ slightly from the averages shown in Tables 6 and 7, which are weighted.

[10] The second of these two years, 2020 (2019-20) includes the first three months of the pandemic as it affected Australia, that is, April-June 2020. However, it is very unlikely that investments which close in the last three months would also have a completion report written about them in those same three months. More than a few months from closing is needed to meet the demands of a completion report – investment completion reports have to be written, possibly a final evaluation, and then, even after the final rating is done, as discussed later, a validation process needs to be undertaken.

As shown in Table 4, the decline in performance scores for completed ratings is mainly driven by the downgrade of ratings from high (5 or 6) to medium (3 or 4). During the first period (2014-2018), 44% of completed investments were assessed 5 or 6, but in the following two years (2019-2020), only 29% were. The share of investments scoring 5 or 6 in efficiency also fell from 32% to 23%. These trends continued during the pandemic. Very few investments got a score of 1 or 2, either before or during COVID.

Table 4: Rating distributions for efficiency and effectiveness for completed investments, sub-period averages

	Effectiveness			Efficiency		
Period	**1 to 2**	**3 to 4**	**5 to 6**	**1 to 2**	**3 to 4**	**5 to 6**
2014-18	0%	56%	44%	2%	66%	32%
2019-20	4%	67%	29%	6%	71%	23%
2021-22	4%	77%	18%	3%	79%	18%

Note: Unweighted annual averages.

We conclude therefore that, if we look at completed investments, there **is** a downward trend in ratings, the commencement of which precedes the pandemic years.

With a 60% satisfactory or success rate in the most recent year, portfolio performance has reached a level that can only be described as poor. One sign of this is that the average rating for completed investments is less than satisfactory (Figure 6). Others are the sharp downward trend (Figure 5) and the performance against the original 85% target. Comparisons relative to other development agencies that use a rigorous rating methodology are also revealing. The World Bank's 2022 performance report reveals that 84% of completed investments were rated satisfactory in 2021-22, using the same definition and six-point scale.[11] The most recent success rate reported by the ADB is 72% for 2021.[12]

A final observation, one that jumps out of the data, is the growing disconnect between the performance trends for ongoing and for completed investments. Whereas final ratings decline, ongoing ratings improve. In the first two years (2014 and 2015), about four in every five investments were rated satisfactory whether by ongoing or final assessments. However, for the last two years, nine out of ten ongoing investments have been rated satisfactory, but only three out of five completed investments.

Using the satisfactory indicator of Figure 6, the average disconnect for the first five years is only 3 percentage points (80% satisfactory rate for ongoing investments minus 77% for completed investments). In the second four years, the average disconnect is 24 percentage points (90% for ongoing investments

[11] Independent Evaluation Group (2022), *Results and performance of the World Bank Group 2022,* World Bank. A score of 4 or more out of 6 is reported by the Bank to be "moderately satisfactory or higher".
[12] Independent Evaluation, (2022), *2022 Annual Evaluation Review*, Asian Development Bank. See Figure 1. The research of Wood, Otor and Dornan (footnote 3) shows that bilateral donors tend to rate their projects more positively than multilateral ones. See the authors' online appendix (Table A1).

minus 66% for completed investments). The satisfactory rate for ongoing investments goes up by 10 percentage points; that for completed investments goes down by 11. (See Annex Figures 4 and 5 for the rightward shift in the cumulative density function of final ratings, and the leftward shift in that of ongoing ratings.)

Explaining the trends 1: declining final ratings

The previous section highlighted two key trends from the performance data. One is the falling ratings of completed investments. The other is the growing disconnect between the assessment of ongoing and of completed investments. This section and the next attempt to explain these two trends, starting in this section with the former.

Importantly, there has been a change in how completed investments are rated, which we refer to as the introduction of a more independent rating system, or, in short, a methodological change. The drafting of both ongoing and final ratings was, and remains, the responsibility of the investment's DFAT manager and/or that manager's supervisor. However, in 2019, a validation process was introduced for final ratings. This was originally carried out by DFAT's evaluation unit, the Office of Development Effectiveness (ODE); after ODE was abolished, the responsibility was given to DFAT's performance branch. Both entities hired consultants to review completed-investment ratings. Final responsibility for the rating of completed investments was taken out of the hands of investment managers and given to the central unit.[13]

Managers will tend to view the performance of their investments with an upward bias. The more the assessment of these investments is dependent on the views of managers the more positive or optimistic it will be. Therefore, reducing that dependency will cause a fall in recorded performance which is exactly what we see happen in 2019, when the performance of completed investments shows a sharp decrease (from 85% in 2018 to 71% in 2019).

In addition, there may have been a decline in performance on account of COVID-19 (as claimed by the DFAT *Annual Report 2020-21*, p .71, and *Annual Report 2021-22*, p. 66). It would certainly not be surprising if the pandemic made effective and efficient implementation more difficult. That said, neither the World Bank nor the ADB rating systems have so far shown a decline in investment performance during the

[13] "[I]f during validation, ratings are found to be inconsistent with the supporting evidence, or do not accurately represent the level of performance described, then the scores assigned by an investment manager may be changed. The ratings from the validation process are the final ratings and are used for external reporting in the DFAT Annual Report."(DFAT January 2022 *Aid Programming Guide*, p. 66). Note that the Aid Programming Guide is updated from time to time, and the hyperlink links to the most recent guide. While the use of final ratings for external reporting commenced in 2021, the more independent rating system for completed investments started in 2019. See *Performance of Australian Aid 2017-18* (p. 76) for 2019 intentions and the *Performance of Australian Aid 2018-19* (pp. 3 and 9) for a report of those reforms as implemented.

pandemic.[14] Nor do the DFAT ongoing ratings show any sign of performance deterioration during that period (Figure 6).

As intimated earlier, there was also a subtle change in the ratings instructions introduced in 2021 (Table 1). Prior to that, the ongoing and completed rating questions were defined either both in relation to outputs and outcomes (for the effectiveness rating) or solely in relation to outcomes (for the efficiency rating). From 2021 onwards, the ongoing ratings questions were both defined in relation to outputs and outcomes and the completed ratings questions were defined only in relation to outcomes. This change, which coincides with the pandemic period, may also have influenced the results, and reduced the scores for completed ratings since outputs are typically regarded as a necessary but not sufficient condition for outcomes.

To explore these issues further, a regression analysis was conducted on completed investment ratings, with a time trend, a dummy for the last four years (for those ratings obtained under the more independent rating system), a dummy for the last two years (to capture the impact of COVID-19 and/or the wording change), and a dummy for the Pacific, since other research has shown that generally aid to the Pacific (from all donors) performs less well on average. [15] Investment sectoral dummies and budgets (in natural log) were also used as explanatory variables.[16] The dependent value was the average value for each investment of the effectiveness and efficiency rating.

The results are shown in Table 5. The introduction of the more independent rating system has the biggest impact: it reduces final ratings by 0.39 (on the 6-point scale) or 9% of the 2014-18 average rating. COVID-19 and/or the wording change also had a substantial impact: it reduced average ratings by 0.26 or 6% of the pre-COVID average. After taking out the methodological change and the pandemic period, there is, if anything, actually a small, insignificant, but positive time trend. This may reflect the same upward trend in ongoing assessment ratings being experienced by final ratings prior to the methodological change (see Figure 6).

The claim that the methodological change caused the decline in final ratings is dependent on the assumption that nothing else happened in or after 2019 that might have influenced final ratings. We can test this more rigorously by comparing ongoing and final ratings before and after 2019, which we do in the next section.

[14] See the references cited in footnotes 11 and 12.
[15] See Wood, Otor and Dornan, (2021), "Why are aid investments less effective in the Pacific?", *Development Policy Review, 40(3),* 121.
[16] The budget variable may change over time, presumably as budget/expenditure estimates are updated.

In other results, as expected, Pacific investments stand out as worse than others: an investment in the Pacific will on average score 0.29 points worse than a non-Pacific one. In terms of sectors, "general" investments perform worse than investments in other sectors: they score on average 0.33 point worse than agricultural investments (the omitted sector with regard to the sectoral dummies), which are similar in performance to most other sectors (resilience, education, governance, and health). Infrastructure and trade investments also do worse, though not as badly as general ones. Larger investments do better than smaller ones: every time an investment doubles in size, its performance can be expected to increase by 0.08 points.

Table 5: Completed investment ratings: regression analysis

Dummy variable for ratings completed under the more independent rating system (2019-2022)	-0.39*** (0.13)
Dummy variable for the last two years	-0.26** (0.13)
Year	0.04 (0.03)
Constant	-82.57 (55.40)
Dummy variable for the Pacific	-0.29*** (0.07)
Budget (natural log)	0.11*** (0.03)
Sectoral dummies	
Resilience	0.00 (0.13)
Education	0.06 (0.14)
Governance	-0.09 (0.12)
General	-0.33* (0.19)
Health	-0.02 (0.14)
Infrastructure & Trade	-0.12 (0.14)
Observations	494
Adjusted R^2	0.12

Notes: Investment clustered standard errors in parentheses. OLS regressions used. The dependent variable is the mean of the efficiency and effectiveness scores. Agriculture is the omitted sector in relation to the sectoral dummies (for more detail on sectoral definitions see Table 7). * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Decomposition analysis can also be used to investigate whether there were any distinct regional or sectoral trends that explain the decline in performance.

Starting with the regional analysis, we compare: Pacific investments, Southeast Asia investments, investments in other regions (rest of the world), and investments that are designed thematically rather

than for particular regions (thematic investments). Given year-to-year volatility, we compare ratings in the first five years with last four (Table 6). All categories, including thematic investments, show a decline in average effectiveness and efficiency ratings. The Pacific is the worst performing region, but its contribution to the decline is about as expected given the size of the portfolio. The same is true for Southeast Asia (see Table 5b). The rest-of-the-world region makes a much smaller than expected contribution to the decline in effectiveness because its ratings drop is only 4%. The thematic investments make a larger than expected contribution, but this is partly because thematic investments first appear in the ratings only in 2018, and so they get a much lower share for the first period than the second. In any case, changes relating to thematic investments only explain about 20% of the overall effectiveness and efficiency changes. In summary, no compelling regional explanation for the decline is evident from the decomposition analysis.

Table 6: Completed investment ratings by region

	Share of projects		**Effectiveness**		**Efficiency**	
Region	**2014-18**	**2019-20**	**2014-18**	**2019-22**	**2014-18**	**2019-22**
Pacific	34%	32%	4.19	3.76	3.87	3.58
Southeast Asia	39%	35%	4.44	3.98	4.22	4.03
Rest of the world	24%	15%	4.37	4.24	4.28	4.16
Thematic	4%	18%	4.17	3.81	4.00	3.84
Total	**100%**	**100%**	**4.33**	**3.92**	**4.11**	**3.87**

	Contribution to the decline in			
	Effectiveness relative to		**Efficiency** relative to	
	Total decline	**Share of projects**	**Total decline**	**Share of projects**
Pacific	33%	100%	38%	115%
Southeast Asia	39%	106%	28%	76%
Rest of the world	6%	27%	14%	69%
Thematic	22%	249%	19%	218%
Total	**100%**	**100%**	**100%**	**100%**

Note: An exact decomposition is used, which decomposes the overall change on the 5-point scale into changes for each region, measured as the change for that region's rating times its original share of investments plus its change in share times the difference between its original effectiveness and the original overall effectiveness. The changes in share make little contribution (8% in aggregate for effectiveness; and 15% for efficiency).

We can also try to explain the ratings decline by decomposing by sector, using the sectoral definitions (or "priorities") defined by DFAT (Table 7). Again, there is no single stand-out sectoral explanation for the decline. The performance of all sectors declines, except for governance which is static. The three sectors with the biggest effectiveness declines (all double digit in percentage terms) are: resilience (disaster risk

reduction and social protection); general; and infrastructure and trade. These three also make the biggest contribution to the ratings declines relative to their prominence in the portfolio, alongside health for efficiency.

Table 7: Completed investment ratings by sector

Sector	Share of projects 2014-18	Share of projects 2019-20	Effectiveness 2014-18	Effectiveness 2019-22	Efficiency 2014-18	Efficiency 2019-22
Agriculture	7%	10%	4.27	4.12	4.27	4.06
Resilience	15%	12%	4.53	3.80	4.20	3.75
Education	18%	16%	4.41	4.14	4.17	4.11
Governance	25%	25%	4.15	3.95	4.00	4.00
General	6%	5%	4.10	3.33	3.86	3.22
Health	18%	10%	4.41	3.94	4.22	3.71
Infrastructure and trade	11%	21%	4.36	3.81	3.97	3.75
Total	**100%**	**100%**	**4.33**	**3.92**	**4.11**	**3.87**

	Contribution to the decline in Effectiveness relative to Total decline	Effectiveness relative to Share of projects	Efficiency relative to Total decline	Efficiency relative to Share of projects
Agriculture	4%	53%	7%	86%
Resilience	22%	161%	24%	171%
Education	11%	64%	5%	28%
Governance	12%	49%	0%	1%
General	9%	150%	13%	212%
Health	13%	86%	26%	169%
Infrastructure and trade	28%	190%	26%	175%
Total	**100%**	**100%**	**100%**	**100%**

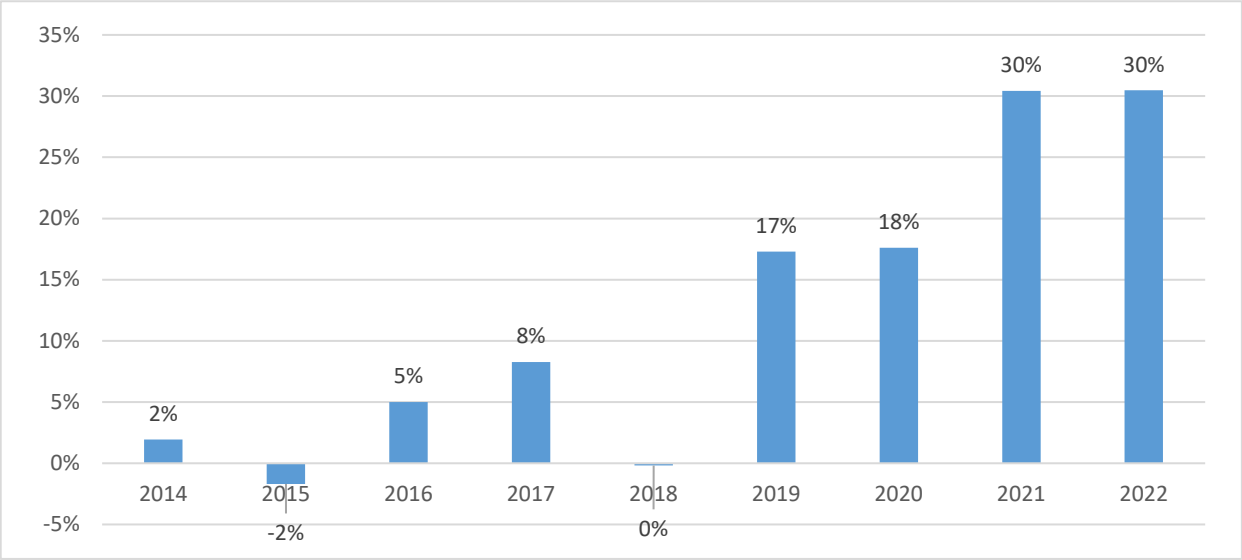
Notes: The same decomposition method is used as in Table 6. See the notes to that table. Again, the changes in share make little contribution (3% in aggregate for effectiveness; and 5% for efficiency). Resilience investments are in disaster risk reduction and social protection as well as humanitarian assistance, though as mentioned in the text humanitarian assessments are not included in this report. 'General' is short for general development support. Agriculture includes fisheries and water investments.

In summary, the shift to a more independent methodology emerges from this analysis as the most important factor behind the decline in the rating of completed investments. There are no prominent regional or sectoral explanations for the decline.

Explaining the trends 2: the growing disconnect

We now turn to the growing disconnect between the ratings of ongoing and completed investments, summarised in Figure 7.

Figure 7: The growing disconnect



Note: Based on Figure 6: satisfactory rate of ongoing investments minus satisfactory rate of completed investments.

In every year except two (2015 and 2018), where there is a marginal difference the other way, ongoing have been higher than final ratings. But the disconnect between ongoing and final ratings has got much higher in recent years. The average disconnect up to 2018 is only 2.7 percentage points. The disconnect since is 24 percentage points. The disconnect in 2021 and 2022 is 30 percentage points. Why this large change? (To view the disconnects in terms of cumulative density functions before and from 2019, see Annex Figures 2 and 3.)

The fact that the disconnect peaks during COVID-19 is particularly surprising. It is not surprising that performance slumped during the pandemic, but it is odd that, whereas the performance of completed investments fell sharply, the performance of ongoing investments was unaffected or, if anything, increased (Table 3). Completed investments would have had on average more pre-COVID years than ongoing investments, so the pandemic should have had a bigger negative impact on the latter than the former.

This section, like the last, uses regression and decomposition analysis, this time to analyse the disconnect. For the regression analysis, a difference-in-differences approach is appropriate, since the trends in ongoing and completed assessment ratings are similar before the change in methodology, and different thereafter

(Figures 5 and 6).[17] Difference-in-differences analysis can be used to quantify the impact of an intervention. In this case, the "intervention" is the change in the shift to a more independent rating methodology, the intervention group are completed investments (to which this change is applied), and the control group are ongoing investments (to which this change is not applied).

To implement this approach, we run a regression using the same dependent variable (the average of the effectiveness and efficiency scores) and the same controls as in the previous section (with dummies for different sectoral investments, for the last four years, for the last two years and for Pacific investments; a time trend; and the log of the investment's budget). We also add a dummy variable for final ratings. Consistent with the difference-in-differences methodology, we include a dummy variable for all final ratings made after the methodological change (that is, from 2019 on). The coefficient on this variable provides an estimate of the "difference-in-differences", that is, the difference made by the shift to the new rating methodology.

In a second regression, we also include a dummy for all final ratings made in the last two years. The interpretation of the second regression is that the "intervention" consisted of not only the introduction of a new rating system but then also a pandemic plus a wording change, two years later. We are interested in whether this combination of factors affected the control and intervention group differently.

The results are shown in Table 8. The most important coefficient is that of the dummy applied only to final ratings conducted after the introduction of the new rating system. The first regression shows that this change reduced ratings by 0.42, an amount that is both large and statistically significant. The coefficient on the dummy variable for final ratings is -0.04 in both models, indicating that, before 2019, final ratings were lower than ongoing ratings, but only marginally. The dummy variable for the last four years is small and positive in both models (0.02 to 0.03); this reflects the upward trend in ongoing ratings.

The estimate of the impact of the methodological change in the second regression is smaller at 0.29. This is because, in this model, COVID-19 and/or the wording change gives rise to another reduction of 0.24 in final ratings. A plausible interpretation is that the shift to a more independent rating system allowed final but not ongoing ratings to respond to the adverse external circumstances posed by the pandemic. However, the change in wording might also have been a factor since it might be easier to get a good rating when outputs and outcomes are considered, rather than when only outcomes are. (Even if aid investments produce outputs, they may not produce outcomes.)

[17] Statistical testing suggests no significant deviation from parallel trends prior to the more independent rating methodology being introduced.

These impacts are large. The average performance of each quartile of investments differs by about 0.5 points (5.06 for the best quartile, 4.54 for the second best, 4.00 for the third, and 3.48 for the fourth). The impact of the methodological shift is roughly equivalent to moving down a quartile. The standard deviation for all ratings is 0.65. In the first regression, the effect size, a measure of the strength of the intervention, is 64%, and in the second it is 81% (including the pandemic/wording effect). These are large effect sizes by any standard.

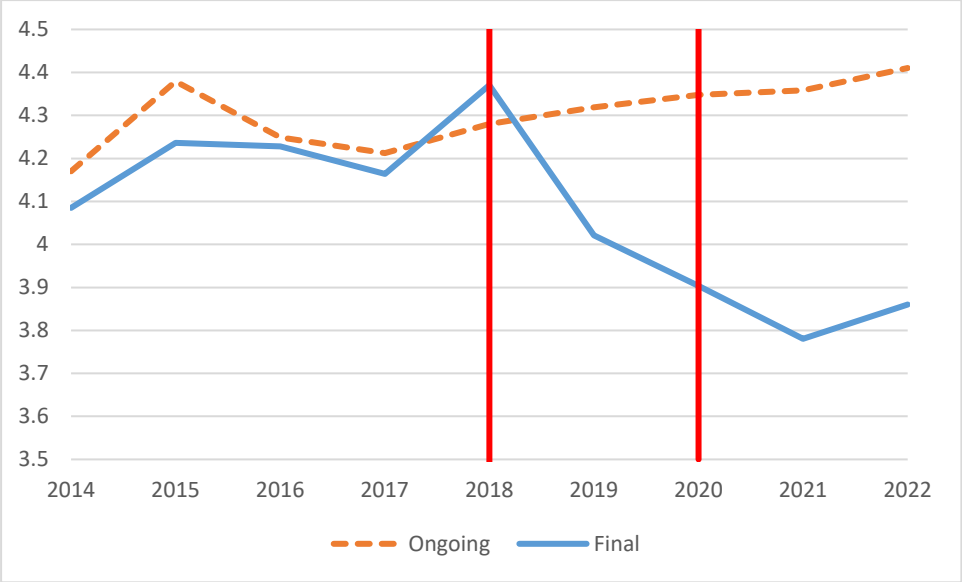
Table 8: All investment ratings: regression (difference-in-differences) analysis

	(1)	(2)
Difference-in-differences estimate of impact of more independent final rating system	-0.42***	-0.29***
	(0.01)	(0.00)
Difference-in-differences estimate of impact of last two years on final ratings		-0.24***
		(0.00)
Dummy variable for final assessments	-0.04**	-0.04**
	(0.00)	(0.00)
Dummy variable for last four years	0.03*	0.02
	(0.00)	(0.03)
Dummy variable for last two years	0.01	0.04
	(0.07)	(0.01)
Year	0.01	0.01
	(0.01)	(0.01)
Observations	3080	3080
Adjusted R^2	0.06	0.07

Note: Other controls used, not reported above, are: natural log of expenditure; sectoral dummies; Pacific dummy. Standard errors, clustered at report type (final/ongoing), in parentheses. OLS regressions used. The dependent variable is the mean of the efficiency and effectiveness scores. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

These results are striking but not surprising. Figure 8 plots the dependent variable for the two types of ratings. The disconnect clearly starts in 2019 (when the new methodology was introduced for final ratings) and then is exacerbated over the last two years. The disconnect is also exacerbated by the ever-growing optimism attached to ongoing ratings.

Figure 8: Average efficiency and effectiveness final and ongoing ratings



We also undertake a decomposition analysis. We can decompose the disconnect between the satisfactory rate for ongoing and completed investments in any year (shown in Figure 7) into four distinct components: a "sample" effect, a "rating" effect, a "timing" effect, and a residual. A rating effect will arise if investments are rated differently when closed than when they are ongoing. A sample effect will arise if the average completed investment was, when it was an ongoing investment, on average worse or better than the average ongoing investment in that year. A timing effect will arise if ongoing investments, as a whole, perform differently from year to year. Finally, the matching required for this decomposition may not be complete as there may be some investments we cannot match. To the extent that it is not, there will also be a residual.

Mathematically, this decomposition can be expressed as follows.

$O_t - C_t =$	[The disconnect equals
$O_t - O_{t-1}$	the timing effect
$+ O_{t-1} - O_{m,t-1}$	plus the sample effect
$+ O_{m,t-1} - C_{m,t}$	plus the rating effect
$+ C_{m,t} - C_t$	plus the residual]

where

O_t is the satisfactory rate for ongoing investments in year t

C_t is the satisfactory rate for completed investments in year t

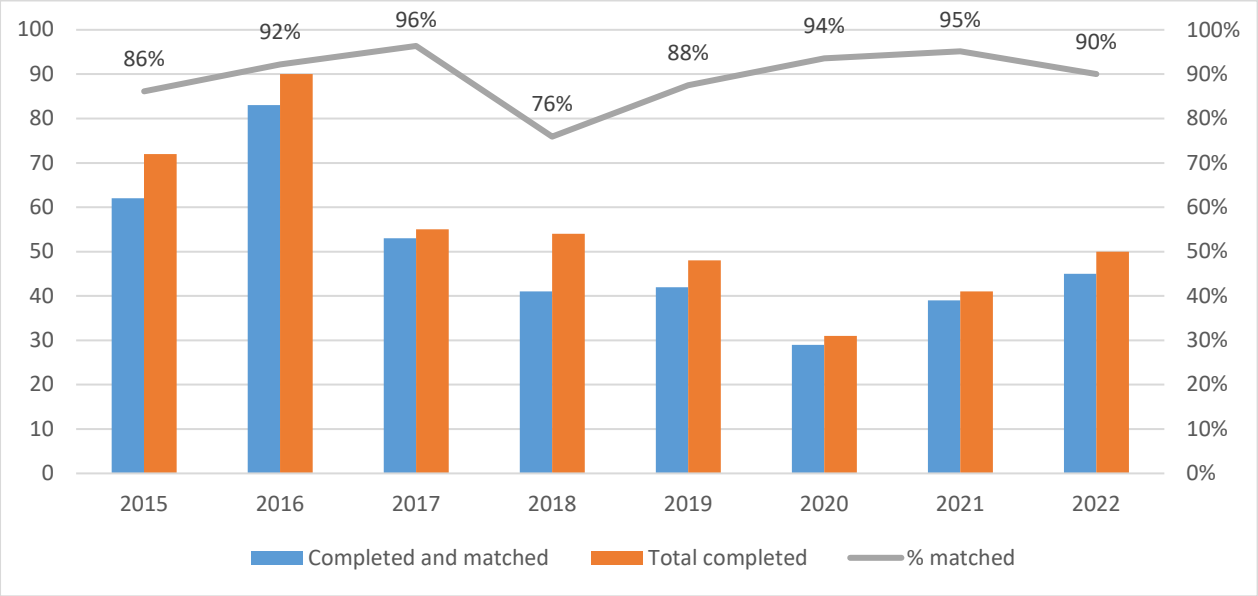
O_{t-1} is the satisfactory rate for ongoing investments in year t-1

$O_{m,t-1}$ is the satisfactory rate for ongoing investments in year t-1 that can be matched to a completed investment rated in year t.

$C_{m,t}$ is the satisfactory rate for completed investments in year t that can be matched to an ongoing investment rated in year t-1

On average, we would expect the sample effect to be zero: every investment has to finish at some point. But it will not be zero every year. We already know that ongoing ratings have been improving, so we know there will be a positive timing effect. Other agencies have experienced a divergence between more and less independent ratings, so we might expect the rating effect to be important following the 2019 reform which made final ratings more independent, but not necessarily before then.[18] The residual, like the sample effect, should be zero on average. Its size in any one year will depend on the share of completed investments we can match. Fortunately, we can match 86% or more of the ratings of completed investments with their ongoing rating in the previous year in every year except in 2018 where this ratio falls to 76% (Figure 9).[19]

Figure 9: All completed investments rated, and those that can be matched



Note: The matching starts in 2015 because completed investments in that year need to be matched with ongoing investments the year before, ie 2014, the first year of the data set.

Based on this methodology, Table 9 below contains the key results using averages over several years to reduce volatility.

[18] For example, the World Bank reports (Box 2.1) a 15-18 percentage point disconnect (for FY20 and FY19, respectively) between the final ratings of management and those of its Independent Evaluation Group.

[19] 22 investments have more than one final rating and 10 of these have final ratings in two consecutive years. This happens when the replacement investment is delayed and the original investment get extended, or investments with multiple phases fail to have a smooth transition. In our disconnect analysis, an investment is considered matched only if it is ongoing in year t-1 and completed in year t.

Table 9: Decomposing the disconnect

	Timing effect	Sample effect	Rating effect	Residual	Disconnect
Averages (percentage points)					
2015-2018 (4 years)	2.37	-3.21	4.37	-0.68	**2.85**
2019-2022 (4 years)	1.37	2.75	17.73	2.11	**23.96**
Entire period	1.87	-0.23	11.05	0.71	**13.40**
Averages (percentages adding to 100%)					
2015-2018 (4 years)	83%	-113%	153%	-24%	**100%**
2019-2022 (4 years)	6%	11%	74%	9%	**100%**
Entire period	14%	-2%	82%	5%	**100%**

Note: The three individual effects (timing, sample, rating) and the residual add up to the disconnect; see the equation on p.17. The analysis starts in 2015 because the decomposition involves a lag of one year: again, see the equation.

Beginning with the entire period, as expected, the sample and residual effects are the smallest. While not zero, they explain only -2% (for the sample effect) and 5% (for the residual) of the average disconnect over the entire period.

The timing effect is larger and explains 14% of the average disconnect for the entire period. This reflects the fact that the satisfactory rate of ongoing investments has only fallen in two of the eight years covered, and has risen over this period from 76% to 90% (Figure 5).

However, it is the rating effect that is by far the largest, explaining 82% of the average disconnect for the period as a whole.

We can also use this framework to explain why the disconnect has increased over time. There is considerable year-to-year volatility, so we focus on the four-year results shown in Table 9: the first four years from 2015 to 2018 and the last four from 2019 to 2022. Results for these two sub-periods are also shown in the table. The disconnect has grown massively between these two periods (from 2.9 percentage points to 24 percentage points). This is for two main reasons. First, and most importantly, the rating effect has quadrupled, from 4.4 to 17.7 percentage points. Second, the sample and residual effects have both changed from negative (so reducing the total disconnect) to positive (adding to it). The change in the sample effect is particularly large: it adds 6 percentage points to the disconnect. In the earlier period, on average, better investments tended to be completed and assessed; in the later period, on average worse investments did.[20]

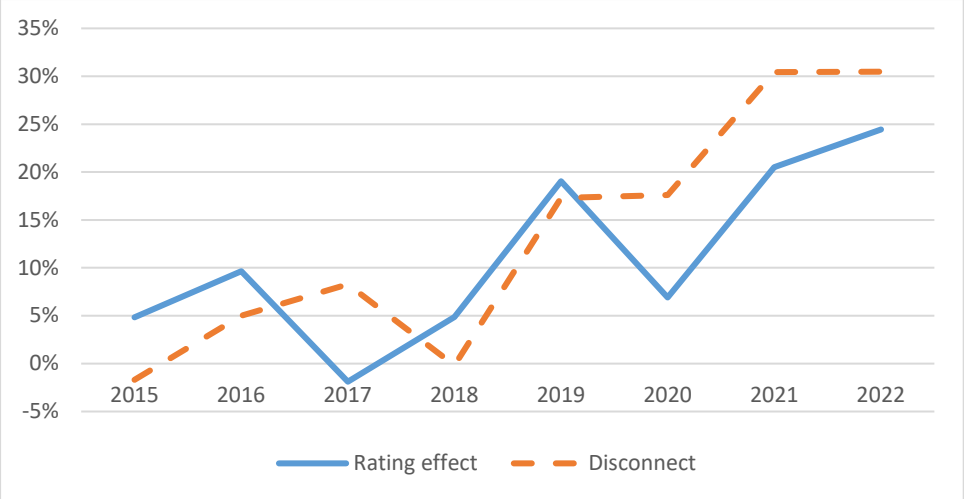
[20] The DFAT *Annual Report 2020-21* (p.71), without using this terminology, notes this sample effect at work when it writes that one reason for only 61% of completed projects being rated satisfactory that year is that "[m]any of the completed investments were located in challenging development contexts (including in PNG, Solomon Islands, Afghanistan and Pakistan)".

To summarise, the main reason for both the size of and the increase in the disconnect is the rating effect. It is increasingly likely that an investment will be downgraded when going from its final ongoing assessment to its completed assessment. In the last two years, about one-fifth to one-quarter of all final ratings have been downgraded from satisfactory to unsatisfactory. And, for the same two years, more than half of all the unsatisfactory investments were rated as satisfactory the year before (see Tables 11 and 12).

To understand this result, it is important to appreciate that the disconnect and the rating effect are **not** the same. The disconnect is the difference in a year between the satisfactory rate of ongoing and completed investments. The rating effect is the difference between the satisfactory rate of completed investments and the satisfactory rate of those same investments in the previous year when these investments were ongoing, and it is calculated (and can only be calculated) for those completed investments that can be matched. As the decomposition analysis shows, there could be several explanations for the disconnect, and it needed to be shown (rather than assumed) that the major explanation for it was in fact the rating effect. As it turns out, the rating effect is on average smaller than the disconnect but it is still significant and has also grown over time (Figure 10).

If it is helpful, one can think of the rating effect as the "pure" disconnect. Once we remove all the noise, we still see a big and growing gap between the way investments are rated in their last ongoing rating and in the way they are then rated once closed.

Figure 10: The rating effect and the disconnect



Note: As defined in the text. The satisfactory rate for completed investments minus the satisfactory rate for ongoing investments from the previous year, for those investments that can be matched.

The quadrupling of the rating effect confirms the importance of the change in methodology for final ratings in 2019. The rating effect was not zero prior to 2019 but it was very small, just 4.4 percentage points on average between 2015 and 2018; there was only a small tendency for investment managers to be more critical on completion than during implementation. The real change came when the rating of completed

investments was taken out of the hands of their managers. This increased the rating effect to 17.7 on average between 2019 and 2022. The change in the rating effect after that change in methodology (13.4) is slightly below the range of the 15-18 percentage point rating effect noted by the World Bank (comparing final ratings of managers with those of the Bank's Independent Evaluation Group in relation to the same investments; see footnote 18.)[21] Given the greater independence of the Bank's evaluation function, that makes sense.

The second most important factor after the rating effect is the timing effect, which explains 14% of the disconnect for the period as a whole and which arises because of the increasing optimism of the ongoing rating system. The 2014 target of 85% may have had an influence by increasing institutional pressure on investment managers to assess their investments' performance as positive to help meet that DFAT-wide goal. The Office of Development Effectiveness, before it was abolished in 2020, reviewed a sample of implementation ratings for robustness. There is also a moderation process for ongoing ratings, in which staff who are not managing the intervention can present their views, but the stringency of this process is variable (e.g., it may just involve an email exchange). These mechanisms would have provided a countervailing force against the pressure to inflate ratings to achieve the DFAT-wide goal, but perhaps not one that was strong enough to counter the upward pressure on ratings from the 85% target.

The 85% target was achieved in 2018, and dropped in 2020. Nevertheless, average ongoing ratings have continued to increase. (The average ongoing rating, measured by the mean effectiveness-efficiency score, increases in six out of eight years, including every year from 2018 onwards, from 4.17 in 2014 to 4.41 in 2022, an increase of 36% of the standard deviation.) Why? In addition to the 85% target, the 2014 framework also required that "[i]nvestments not delivering value-for-money will be required to be improved within one year or be cancelled."[22] This arrangement remains in place to this day, with investments rated three or less for effectiveness and efficiency designated as "investments requiring improvement" or IRIs. Such investments need remediation plans and high-level management attention, must be cancelled within a year or a written reason provided for why not, and IRI numbers are included in the annual report.[23] All this must result in an incentive not to rate an ongoing investment's performance as unsatisfactory due to the extra work and pressure involved. This incentive was not in play in relation to final assessments since, by definition, these ratings were in relation to investments that had already been finished and thus could not be improved. And it is plausible that over time, as staff and their managers experienced the costs of reporting an investment to be an IRI, the incentive not to so report has increased.

[21] 13.4 is 17.73 minus 4.47 in Table 9. We need to subtract the initial amount to compare final ratings before and after the methodological change. In other words, we assume that, absent the methodological change, the same gap between final and ongoing ratings would have prevailed in the two subperiods.
[22] From *Australia's new development policy and performance framework: a summary* (2014).
[23] DFAT January 2022 *Aid Programming Guide*, p. 67. Note that hyperlink may link to more recent edition.

Hence the upward trend in ongoing ratings, even after the 85% target was achieved, and even after it was abandoned. Evidence for this is provided by Annex Figure 4, which shows that for ongoing ratings the biggest improvement before and after 2019 was the fall in investments rated 3.5 or less, the share of which fell by about two from about 20% to about 10%. One plausible explanation for this is that DFAT's aid managers became increasingly reluctant to rate their investments as unsatisfactory because of the consequences of doing so.

Finally, final assessments have been the focus of performance reporting since 2020. However, it can be argued that the same pressure to inflate them has not arisen. Whereas the earlier framework had a target of 85% for all investments, the current framework has no target, and DFAT has even stated that a result of than 61% of completed investments being rated satisfactory is "on track".[24] The earlier 85% target was a high-profile one. It was one of only "ten key targets" DFAT reported against annually in special, short and self-standing *Performance of Australian Aid* reports. By contrast, the new indicator is only one of a number of "tier three" indicators (which follow numerous higher profile "tier one" and "tier two" indicators), and is only reported on in the annual report, a document of several hundred pages. Performance against the 85% target was assessed by showing trend data over several years, to see whether performance was improving or worsening. By contrast, only the current and previous year's performance is shown under the current framework. Lacking the impetus of either a target or a series of data, no assessment of performance is possible, and this reduces the pressure to improve final ratings. Moreover, from 2019 on, the ratings of completed investments were reviewed by external consultants, who are less likely to be influenced by corporate imperatives.

Discussion

The previous three sections have documented two important trends from recent Australian performance data. One, the ratings given to completed investments have declined to very low levels: the average rating is now less than satisfactory, and two-in-five completed investments are rated as such. Two, there is a growing and now large disconnect between the final and ongoing ratings of investments.

The report has also tried to explain these two trends. One main factor emerges as explaining both, that is a shift to a more independent rating system in 2019 for completed investments. This shift has led to a decline in the rating of completed investments but has not affected the rating of ongoing investments, resulting in a quadrupling of the gap between them (the rating effect).

[24] For example, see the DFAT *Annual Report 2020-21*, p. 70.

We have also argued that aspects of DFAT's performance management system have put upward pressure on ratings, especially of ongoing investments, and that this also helps explain the growing disconnect. In particular, the additional management effort required if an investment is rated as unsatisfactory may help explain why ongoing ratings have continued to increase even after the target for them was achieved and then abandoned.

It may be that COVID-19 led to a decline in the performance of investments, though this is evident only in the final not the ongoing ratings, thereby adding to the disconnect. The pandemic period overlaps with the subtle recent changes in instructions for ongoing and final ratings, which might have also had an impact.

There are no prominent regional or sectoral explanations for the decline in final ratings.

Other criticisms have been made of DFAT's management of the aid program, such as declining in-house development expertise and a growing weight given to geopolitical objectives. We cannot be absolutely sure that these are not driving the post-2018 decline, or even the decline during the pandemic. However, it is not clear why these factors would only influence performance from 2019 onwards, and why they would only influence final and not ongoing ratings.

If the pandemic has caused performance to fall, we would expect final ratings to bounce back with the waning of the pandemic, but the two-year pre-COVID average rating for effectiveness was a marginal pass (4.03) and for efficiency a marginal fail (3.90) (Table 3). What needs to be done to lift aid performance back to and above its pre-COVID level will require a lot more analysis, including of information (the actual ratings) that is not in the public domain. Our own recommendations follow below.

Recommendations

The shift to a more independent rating methodology in 2019 for completed investments is a good thing. The use of independent rating methodologies is common for multilaterals, but not for bilateral aid agencies. DFAT is to be commended for moving in this direction, and for shifting the focus of its portfolio performance assessment from all to only final ratings. At the same time, it can be seen from the analysis presented in this report that these positive reforms have had some adverse and likely unanticipated consequences. This section contains six recommendations that flow from our analysis.

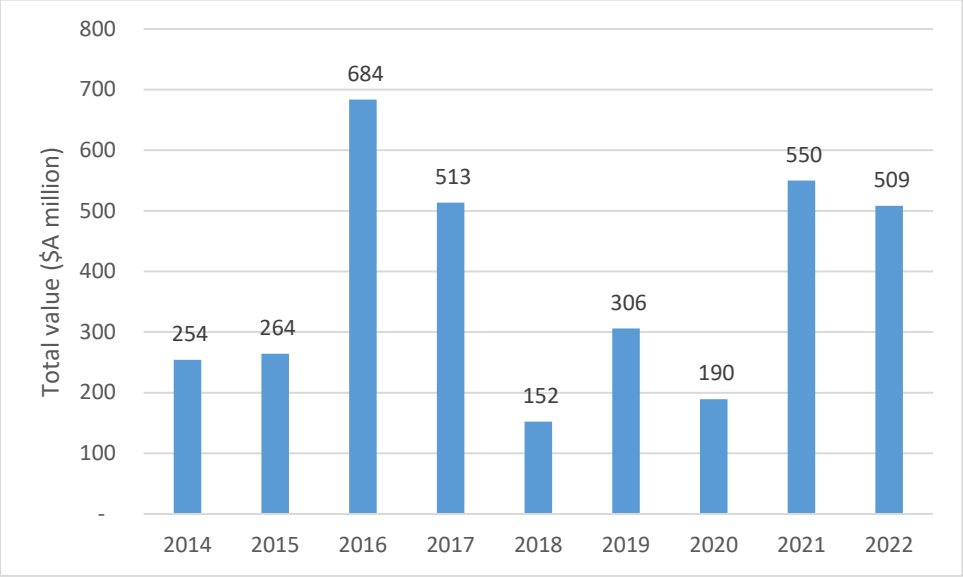
First, DFAT should continue with its more independent rating system introduced in 2019. As noted, this is a positive reform, one that has likely resulted in more frank and insightful assessments, and one that puts DFAT at the front of the pack relative to other bilateral donors. We argue in favour of retaining consultant review, as part of a central validation process that keeps final ratings removed from the control of investment managers.

Second, the Office of Development Effectiveness (ODE) and Independent Evaluation Committee (IEC) should be re-established. The problems identified in this report have largely occurred largely since the abolition of ODE, DFAT's evaluation body, and its oversight body, the IEC. One of the roles of ODE, prior to its abolition in 2020, was to assess "DFAT's internal performance management system".[25] This was a role distinct from that of the performance branch, which is responsible for managing the performance system, as against assessing it. This gap has allowed the current situation to develop. In short, the trends revealed in this report, and the lack of any response to them to date by DFAT, make a compelling case for the reintroduction of ODE and the IEC as aid effectiveness champions and watchdogs within the department.

Third, DFAT needs to take its own performance assessments more seriously, and reverse some of the recent decline. DFAT appears unperturbed by its low performance ratings of completed investments, declaring it be "on track".[26] However, a three-in-five satisfactory rate and an average rating of unsatisfactory seems too low to be on track, whether assessed against the target that applied until just before 2020 (85%) or against earlier performance (see Figure 6) or against the performance standards of other aid agencies.

Of course, one would not expect all investments to succeed, but, in each of the last two years, the annual value of unsatisfactory investments has been half a billion dollars (Figure 11). This is, by any standard, of a size significant enough to warrant corrective measures.

Figure 11: Value of investments completed each year rated unsatisfactory



[25] DFAT, "The Office of Development Effectiveness".
[26] For example, see the DFAT *Annual Report 2021-22*, p. 65.

DFAT should carry out its own investigation into the fall in investment ratings. A review of unsatisfactory investments should be undertaken to discover lessons learnt. More staff with development expertise might be needed, or better preparation, or different types of investments, or other changes may need to be made.

Fourth, DFAT should guard against solving the problem by grade inflation. The easiest way to lift the pass rate would be to mark more easily. While the involvement of consultants in reviewing completion grades is useful, the fact that responsibility for completion grades lies with a central unit means that the grades could still be very responsive to an effort to lift them. There is no easy solution, but we recommend against adopting a performance target for final ratings, given the deleterious effect the earlier target for all ratings seems to have had.

Fifth, a communication effort is needed. There may be a fear that a higher profile for the low satisfactory rate of completed aid investments could lead to a backlash and a loss of support for Australian aid**.** However, with the change in government and new aid strategy there is an opportunity to draw a line in the sand, and to get on the offensive. Reporting a low satisfactory rate without explanation is a great vulnerability. DFAT is to be commended for reporting publicly on its investment ratings. We found that about 70% of ratings were published from 2018 to 2022 (Table 10). Most of the unpublished ratings belong to global or sectoral program portfolios: performance reports are not written for these portfolios (unlike for country and regional portfolios) but should be.

Table 10: Publicly released investment ratings as a proportion of the total

	Number of projects in the list provided by DFAT		**Number of projects published on the aid website**		**% of published**
Year	Ongoing	Completed	Ongoing	Completed	
2018	280	54	196	37	70%
2019	295	48	199	33	685
2020	289	31	212	20	73%
2021	279	41	191	34	70%
2022	273	50	189	33	695
Total	**1416**	**224**	**987**	**157**	**70%**

While it is good that most ratings are published, because no reasons are provided for any of the ratings, once one becomes aware of them, the effect is discombobulating.

Table 11 is a list of the completed investments rated as unsatisfactory in 2021 and Table 12 the corresponding list for 2022. There were 16 unsatisfactory, completed investments worth $550 million

reported on in 2021. These are listed in Table 9, arranged in the order of their average effectiveness and efficiency rating. Nine of the 16 (the ones whose names are bolded) went from being assessed as satisfactory in their final ongoing rating to being assessed as unsatisfactory during their completion rating. Only two of these investments were rated (prior to closure) as medium risk, and only one as high risk; the other investments, despite being regarded as low risk, nevertheless ended up as unsatisfactory.

Table 11: Unsatisfactory investments completed in 2021

Country	Project name	Amount ($AUD)	Period	Eff'ness	Eff'ncy	Risk
Thematic	**Sexual and Reproductive Health Supplies Prepositioning for Disaster Response**	5,188,569	2016-2020	3	5	Low
Timor-Leste	M&E	8,504,529	2016-2021	4	3	Low
Kiribati	***Growth and Economic Management***	8,523,308	2010-2021	3	4	Low
Laos	**Development Learning Facility**	7,091,807	2015-2021	3	4	Low
Afghanistan	**Elections Assistance**	7,163,293	2017-2020	3	4	Low
Solomon Islands	**Program Enabling**	41,507,388	2015-2021	4	3	Low
South and West Asia Regional	Sustainable Development	47,937,021	2016-2021	3	3	Low
Samoa	*Education*	17,256,939	2011-2021	3	3	Low
Solomon Islands	Transport	20,639,725	2016-2021	3	3	Low
PNG	**Education**	268,422,538	2011-2020	3	3	Medium
Thematic	**Children with Disabilities**	4,993,000	2015-2020	3	3	Low
Solomon Islands	**Growth and Economic Management**	53,600,626	2016-2022	3	3	Low
Mekong	**IFC Hydropower**	8,000,000	2013-2020	3	3	Low
PNG	*Pacific Leadership and Governance Precinct*	12,038,993	2014-2020	3	3	High
Pakistan	WB Trust Fund for Nutrition	26,125,355	2013-2020	2	2	NA
Indonesia	Indobeef Project	13,005,776	2012-2021	1	2	Medium
TOTAL		**549,998,866**				

Note: Actual investment names have been replaced by simplified/shortened ones. Bold indicates the investment went from satisfactory in its last ongoing rating to unsatisfactory in their final rating. Plain indicates no change in status from satisfactory to unsatisfactory or vice versa. Italics means no rating in the previous year.

There were 20 unsatisfactory, completed investments worth $508 million reported on in 2022. These are listed in Table 12, again arranged in the order of their combined effectiveness and efficiency rating. 11 of the 20 (the ones whose names are bolded) went from being rated as satisfactory in the course of their final ongoing rating to being rated as unsatisfactory in their course of their completion rating. Only four of these investments were rated (prior to closure) as medium risk, and none as high risk.

Table 12: Unsatisfactory investments completed in 2022

Country	Project name	Amount ($AUD)	Period	Eff'ness	Eff'ncy	Risk
Mongolia	**Australia Mongolia Extractives Program**	16,749,525	2012-2022	3	4	Low
East Asia Regional	**Open, Inclusive and Accountable Governance in SEA**	3,272,171	2018-2022	3	4	Low
Thematic	*Innovation Resource Facility*	35,500,000	2015-2022	3	4	-
Tonga	**Health Systems Support Program II**	19,934,597	2015-2022	4	3	Low
Thematic	**LAUNCH Food**	13,500,000	2015-2021	4	3	Low
PNG	**Partnerships Fund**	194,094,653	2017-2022	4	3	Low
Solomon Islands	**Democratic Governance Phase 2**	14,773,332	2017-2021	4	3	Low
PNG	*Education Support Services*	5,995,719	2020-2022	4	3	Low
PNG	Combating Corruption	6,310,584	2014-2022	3	3	Low
Thematic	MIT Solve Challenge on the Workforce of the Future	4,013,185	2017-2021	3	3	Low
Samoa	Governance Support Program	7,728,333	2018-2022	3	3	Low
Nauru	**Improved Education**	34,262,947	2009-2022	3	3	Low
Samoa	Parliament Complex Redevelopment	13,208,545	2012-2022	3	3	Medium
Samoa	Economic Infrastructure	18,740,951	2012-2022	3	3	Medium
Thematic	**Climate Change projects (Dept of Climate Change etc)**	14,750,000	2013-2021	3	3	Low
Indonesia	**Environmental Governance**	10,096,238	2015-2024	3	3	Low
Fiji	**Supporting Private Sector Development**	17,064,000	2016-2021	3	3	Low
PNG	Institutional Partnership Program	65,719,068	2017-2022	3	3	Medium
Pacific Regional	Innovations in Financing	9,558,061	2015-2021	2	3	Low
Laos	**Microfinance V - Access to Finance for the Poor**	3,230,000	2018-2022	2	2	Medium
TOTAL		**508,501,908**				

Note: Actual investment names have been replaced by simplified/shortened ones. Bold indicates the investment went from satisfactory in its last ongoing rating to unsatisfactory in their final rating. Plain indicates no change in status from satisfactory to unsatisfactory or vice versa. Italics means no rating in the previous year.

Several of the investments in the two tables, despite being judged unsatisfactory, have slick websites extolling their achievements. Some of the investments that are rated as unsatisfactory are very important ones. In Papua New Guinea alone, these include: a ten-year, $268 million investment in school education; a five-year $194 million investment that funded NGO programs in education and health; and a $66 million investment to foster links between Australian and Papua New Guinean government departments.

Intriguingly, two ministerial flagships are recorded as unsatisfactory. The Pacific Leadership and Governance Precinct was Foreign Ministers Julie Bishop's flagship PNG program. The Innovation Resource Facility, LAUNCH Food and the MIT Solve Challenge were all part of her aid program-wide flagship InnovationXchange. All four of these interventions were rated as unsatisfactory after closure.

There is no public explanation at all of why DFAT regards any of the 36 investments reported on in 2021 and 2022, with a total value in excess of $1 billion, to be unsatisfactory.

DFAT should publish all its completion rating reports. In addition, any verdict of unsatisfactory performance of a completed investment given in an annual country performance report (which is where the ratings are currently provided) should be accompanied by a statement explaining the reason for the verdict. At the portfolio level, a detailed public performance analysis is needed.

Sixth, ongoing ratings are meaningless due to the large disconnect with final ratings, and the performance assessment system for ongoing investments should be overhauled. More than half the unsatisfactory completed investments rated in 2021 and 2022 were rated as satisfactory in their previous ongoing investment rating (9 out of 16, and 11 out of 20, respectively). The improvement in ongoing ratings even during the pandemic further stretches credulity.

A review is needed to work out what is going wrong, and a more realistic system of review of ongoing investments needs to be instituted. This could include subjecting a sample of self-assessments to independent review every year. To minimise potential disconnect, the questions used for ongoing and final ratings should be identical except for the tense (e.g. "Are we achieving …?" for ongoing investments, and "Have we achieved …?" for completed ones.) It would also be useful, if this is not already done, to feed final ratings back to the managers of the relevant investments (when they were ongoing) for their reflection and response. The practice of putting pressure on managers to prematurely close unsatisfactory projects should be reconsidered given the evidence that this is leading to a reluctance to rate projects critically.

Conclusion

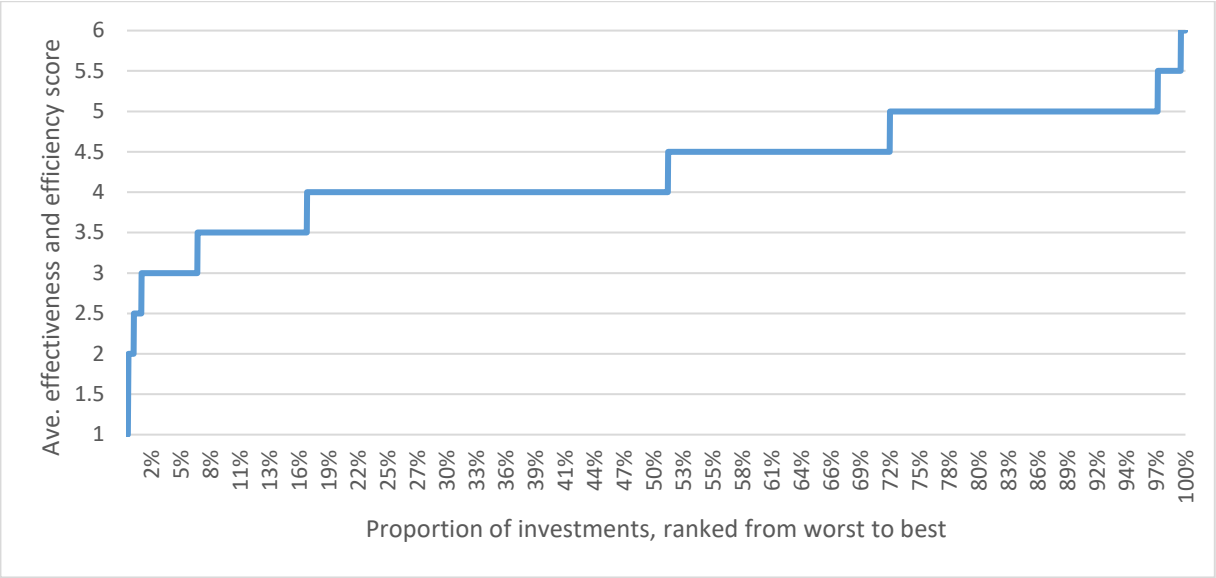
This report began with a question in its title: Why are two-in-five Australian aid investments rated unsatisfactory on completion? Despite our best efforts, we have been able to give a partial rather than complete answer to this question. Only DFAT has the information, in particular the access to the actual ratings, to answer this question fully.

What we can say is that an important part of the story of the decline in reported performance to its current low level is methodological – which explains why we refrain from making any comments about underlying performance trends. DFAT is to be commended for having moved to a more independent system of investment ratings of completed investments, and to have shifted from ongoing to final ratings when assessing the health of its aid portfolio. However, if it retains these basic elements of its performance system, as we hope it does, additional reforms are needed. It is not compatible to introduce a frank-and-fearless, publicly reported investment rating system, and not adjust other parts of the reporting and review

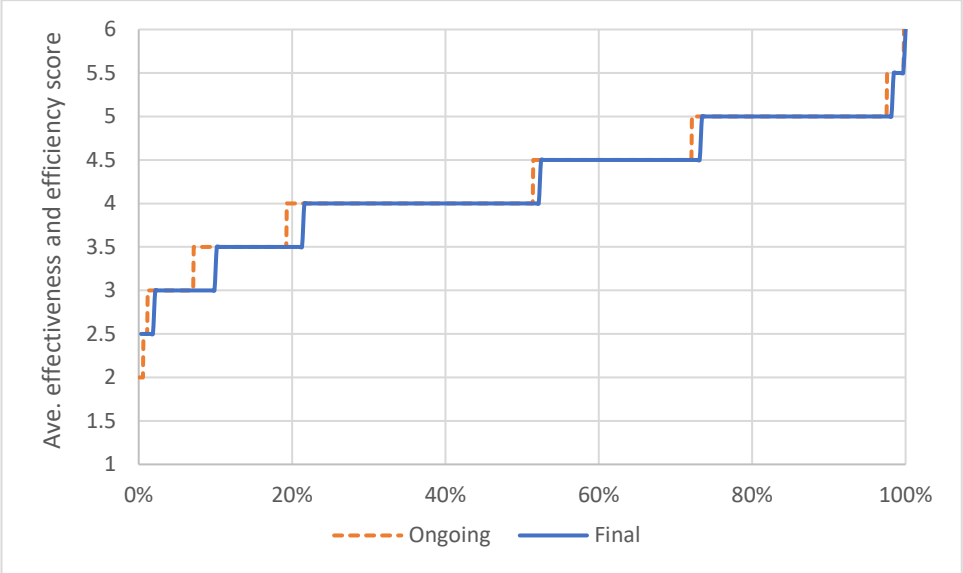
system. A rating system that throws up high and increasing investment ratings is one thing; much more difficult to manage is one that has resulted in significantly lower ratings. Our recommendation, and hope, is that DFAT retains its rating system, re-establishes its performance-oversight architecture – in the form of the Office of Development Effectiveness and the Independent Evaluation Committee – and improves other parts of its performance review and reporting system.

Annex

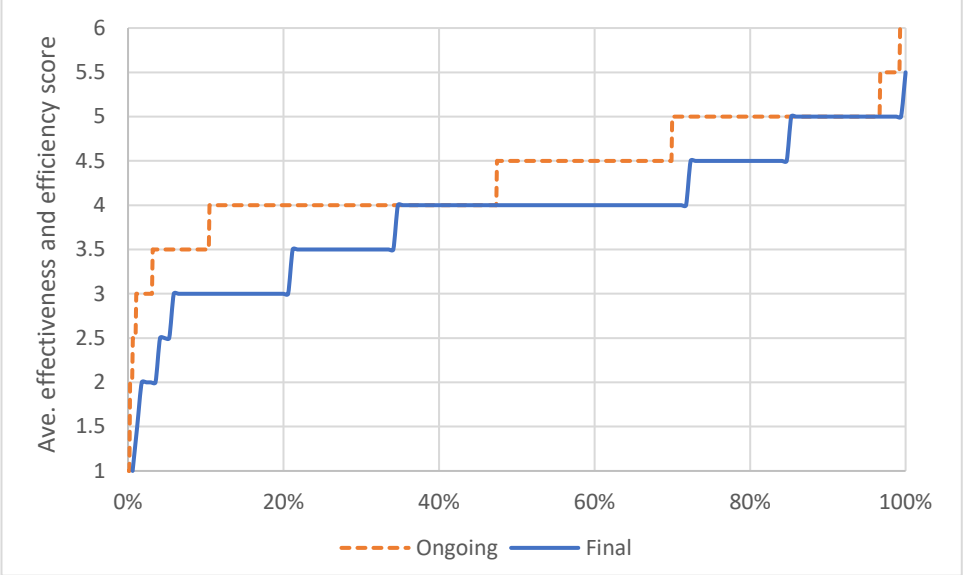
Annex Figure 1: Cumulative distribution function for all ratings



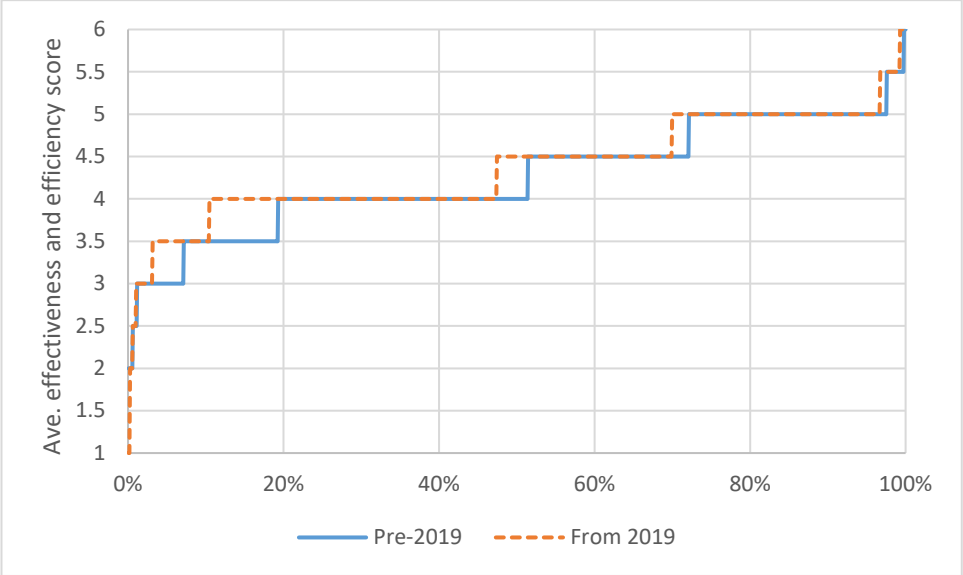
Annex Figure 2: Cumulative distribution function for ongoing and final ratings, pre-2019



Annex Figure 3: Cumulative distribution function for ongoing and final ratings, from 2019



Annex Figure 4: Cumulative distribution function for ongoing ratings, pre-2019 and from 2019



Annex Figure 5: Cumulative distribution function for final ratings, pre-2019 and from 2019