Australian Government
The Treasury

Australian Centre for Evaluation

Evaluation of a trial of generative artificial intelligence (Copilot) in the Department of the Treasury

Summary report

February 2025

Background

In November 2023 the Prime Minister announced the Australian Government would conduct a 6-month trial of Microsoft 365 Copilot (Prime Minister of Australia, 2023). The Digital Transformation Agency (DTA) coordinated the trial at a whole-of-government level with support from the Artificial Intelligence (AI) in Government Taskforce.

The Department of the Treasury's (Treasury) Copilot trial ran for 14 weeks, from 20 May to 23 August 2024. A total of 218 staff participated. This report summarises the methods, findings and lessons learnt from an internal evaluation of the Treasury Copilot trial, conducted by the Australian Centre for Evaluation (ACE).

Evaluation approach

This evaluation was based on a mixed-methods approach that included:

- surveys of trial participants and their managers
- focus groups with trial participants
- a collation of case studies describing examples of participants' use of Copilot
- a review of a Copilot trial issues log.

This report, its findings and lessons learned, are structured against 5 key evaluation questions:

- To what extent was Copilot **implemented** as intended?
- To what extent is Copilot **appropriate** in Treasury's context?
- To what extent does Copilot support **process improvement**?
- To what extent does Copilot support **quality work outcomes** in Treasury?
- Were there any **unintended outcomes** of using Copilot (positive/negative)?

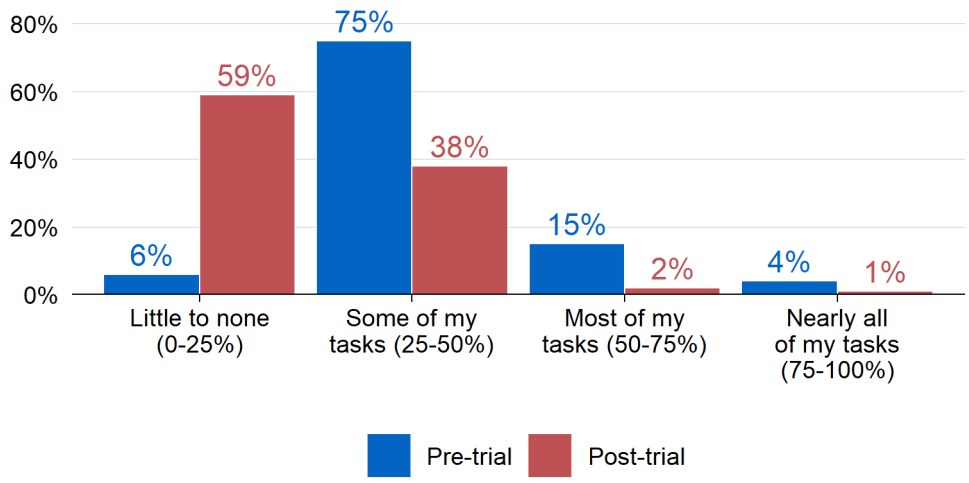
Summary of findings

The findings of this evaluation are organised according to 5 key evaluation questions. These findings relate specifically to Treasury's time-limited trial of Copilot, and do not represent a broader review of generative artificial intelligence (AI) products and their suitability for specific use cases within Treasury.

Implementation: the technical implementation was smooth, however, training and time to learn to use Copilot was limited, and participants had high expectations of the product which were not met. The technical implementation of Copilot was smooth, with relatively few issues encountered during the trial period. However, overall usage of the product during the trial period was lower than expected, and most participants reported using Copilot 2–3 times per week or less. Unrealistically high expectations at the trial outset may have contributed to the problem, as some staff were discouraged by the performance of the product and gave up using it (Figure 1).

There were also more fundamental issues since Copilot did not perform as well as generative AI products that staff had used elsewhere. In part, this was due to restrictions imposed by the Treasury's IT security environment. Staff required time to learn how to use Copilot effectively, which they found challenging to fit into their workload. A common request from participants throughout the trial was for more tailored and targeted education and training to support their use of Copilot.

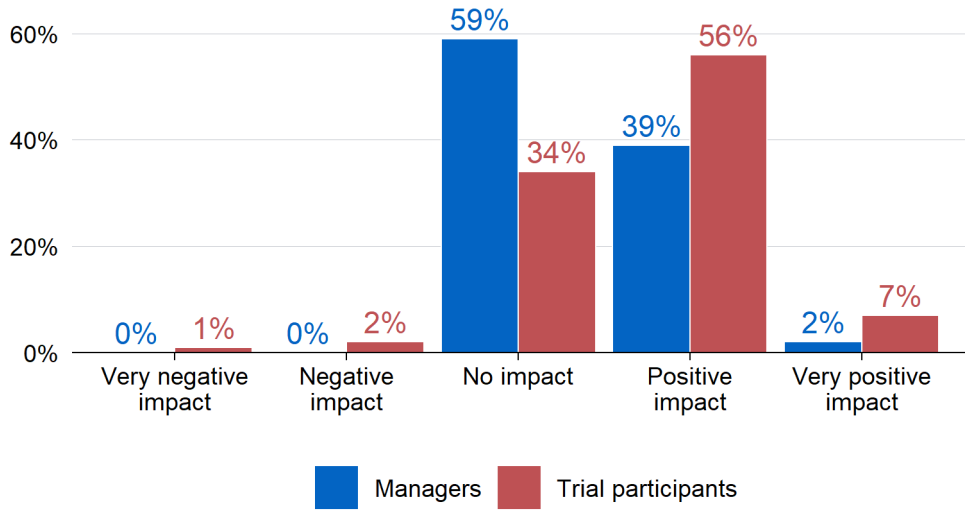
Figure 1 Expected and actual proportion of workload participants felt Copilot could/did support



Appropriateness: the initial 'use cases' for Copilot were appropriate for Treasury, however, the product was not suitable for more complex tasks. There were 4 use cases initially proposed for Copilot: generating structured content, supporting knowledge management, synthesising and prioritising information, and undertaking process tasks (see Appendix D of the main report for more details). The consensus from participants was that these use cases were appropriate for the Treasury context, but that Copilot was not appropriate for more complex tasks, mostly due to the limitations of the product itself. Participants expressed concerns about functionality relative to other generative AI products on the market. Staff are also particularly sensitive to the need for transparency to ensure public trust in the government is maintained, and guidelines to support the use of generative AI if Treasury adopts Copilot or similar products.

Process improvement: Copilot's clearest benefits related to improvements in basic administrative tasks. These improvements included finding and summarising information, generating meeting minutes, knowledge management and drafting content (Figure 2). Efficiencies in basic tasks meant that trial participants could spend more time on high-value or strategic tasks. Although the evaluation did not explicitly measure time saved for basic administrative tasks, the Copilot licence costs are relatively minor compared to the potential efficiency gains for basic tasks: an APS6[1] would need to redirect approximately 13 minutes of time from low-value to high-value tasks per week to offset the licence cost.

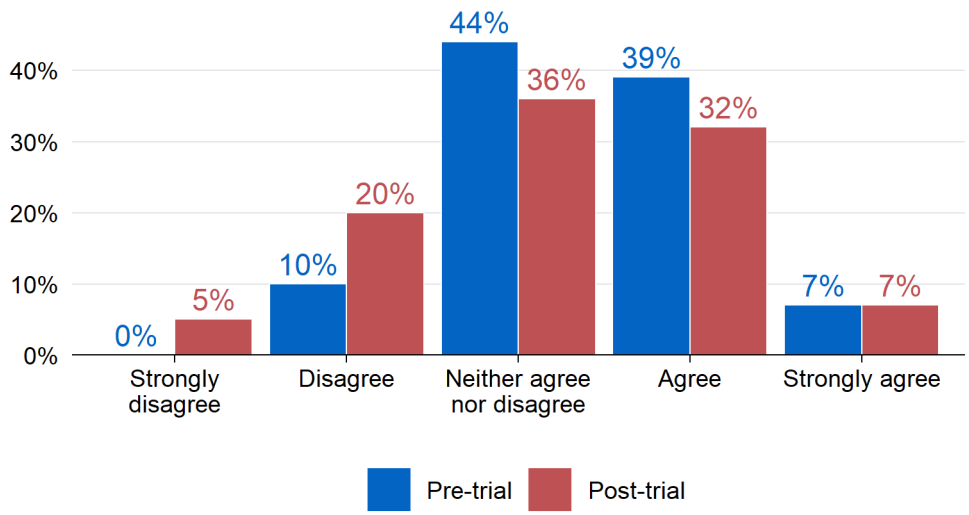
Figure 2 Staff and manager reports on the impact of Copilot on work processes



Quality work outcomes: the evaluation did not find clear evidence that Copilot helped improve work outcomes during the short trial period, but there were promising indicators. This may be due to several factors, including that the trial period was not long enough to provide definitive evidence on the impact of Copilot on work outcomes, or that the effects of Copilot are more difficult to trace because work typically undergoes further revisions prior to finalisation. While some participants were positive about the benefits of Copilot to their work outcomes (Figure 3), many participants and their managers were neutral about Copilot's impact. Further, while there were some slight positive shifts in indices of staff wellbeing and satisfaction, these changes cannot necessarily be attributed to Copilot.

1 An APS6 staff member is a mid-level position within the Australian Public Service that involves some responsibility and expertise. Further information about the work level standard can be found on the APS Commission's website: https://www.apsc.gov.au/working-aps/aps--employees-and-managers/work-level-standards-aps-level-and-executive-level-classifications

Figure 3 Participant ratings of Copilot's impact on work quality



Unintended benefits: Copilot had several unintended benefits relating to accessibility and inclusion, work confidence, and Treasury networks. An unanticipated benefit of Copilot was its ability to contribute to accessibility and inclusion for neurodivergent and part-time staff, or those experiencing medical conditions that require time off work. This occurred via various mechanisms including automatic summaries of missed meetings and support commencing work where staff have previously had issues doing so, and levelling the playing field for those who struggle to navigate workplace norms or culture. This also partially contributed to a small increase in work confidence for some, particularly more junior employees or those newer to Treasury.

Progress towards outcomes: The evaluation also explored progress towards Copilot's short-term and medium-term outcomes (outlined in Appendix A of the main report: Program Logic). A summary of how the trial is progressing towards these outcomes is documented in Table 1. In accordance with the key findings, Copilot was valuable for the identified use cases, and most beneficial for process improvement and knowledge management. It is expected that the use of Copilot and staff competence will increase with time and experience. Indications of some reductions in workload stress suggest there is potential for generative AI to impact this area in future. There was no evidence from this evaluation that Copilot improved workflows and new approaches to problems in the short-term: with technology enhancements and skill development in this area, this is worth continuing to monitor.

Table 1 Copilot's outcome progress

Type	Outcomes	Rating
Short term	Participants have competence to use and confidence to experiment with the Copilot tool	Moderate progress
Short term	Participants are using Copilot for the identified use cases	Good progress
Short term	Participants indicate an increase in work satisfaction using Copilot	Moderate progress
Short term	Participants indicate an increase in process improvement using Copilot	Good progress
Medium term	Participants are using all relevant functions of the tool to benefit their work	Moderate progress
Medium term	Copilot is supporting improved knowledge management	Good progress
Medium term	Participants indicate reduction in workload stress	Moderate progress
Medium term	Copilot is improving workflows and new approaches to problems	No evidence of progress

Lessons learnt and recommendations

The following recommendations highlight how future implementation of any generative AI product could be improved. These recommendations are both contingent on whether Treasury decides to rollout new generative AI products to staff, and are applicable to any generative AI product (versus Copilot specifically).

1. **In any future rollout of generative AI, provide clear and specific use cases to distribute licences to staff who can demonstrate likely benefits and time savings. And manage expectations about what the generative AI product can offer.** The evidence suggests Copilot has specific benefits for process improvement and basic administrative tasks. Providing specific use cases to staff that outline these benefits will support staff in deciding whether the product is appropriate for them, and how to use it. Priority for licence distribution could be given to those who can demonstrate likely benefits. Communications for staff should also be specific about the intended benefits of any product to avoid inflating expectations, which will mitigate the risk of disengagement with a product if it does not immediately meet expectations. Given evidence that generative AI products could support accessibility and inclusion, priority for access could be given to staff experiencing barriers relating to access and inclusion to support their work.
2. **Any future rollout of new generative AI products should be based on a phased approach.** Future rollouts should commence with a small group of staff and continue rollout to wider groups over time. Such a strategy will require sustained investment and effort to ensure the rollout occurs in line with technology developments.
3. **Any future rollout of generative AI products should include an assessment of the appropriate level of investment in education and training.** Formal training and supports enable staff to make the most of generative AI products. Future training should rely on both structured educational opportunities and dynamic capability building mechanisms. Any future implementation of generative AI will need to account for the cost of training and the associated time commitments for staff.

4. **In any future rollout of generative AI products, develop guidelines to support the transparent use of generative AI.** Guidelines should be used to set expectations around the use and disclosure risks of generative AI, including the requirement to own any outputs created by generative AI. These will mitigate against any potential loss of trust in Treasury's work. Guidelines should be developed in consultation with relevant parties and should be consistent with legislative and other APS requirements.
5. **The implementation and impact of new generative AI products takes time and should be monitored over the longer-term to determine potential impacts on quality and timeliness of work.** Regular reviews of work outputs should include subjective and objective data where possible. Once the product has reached maturity, the impact of generative AI could also be tested in an experimental setting. This will contribute to the nascent evidence base on the benefits and appropriateness of generative AI in Treasury, and more generally.
6. **Staff outcomes, including staff wellbeing, job satisfaction, and workload-related stress should be considered as important secondary outcomes of any generative AI product implementation.** Improvements in staff-related outcomes are a foreseeable secondary benefit of generative AI. Monitoring of such outcomes and any unintended benefits should continue throughout the rollout of any future generative AI product.
7. **Conduct periodic assessments of whether emerging generative AI products may be better suited to Treasury's security requirements and existing IT infrastructure.** Treasury should continue to review the suitability of emerging generative AI products for implementation within Treasury's IT environment.

Limitations

There are several limitations of this evaluation. These include:

- The trial was conducted for a total of 14 weeks, which was only long enough for an initial pilot of the product. This meant that neither Copilot as a product (which is relatively new) nor the participant's usage of the product had the opportunity to reach maturity of implementation.
- The evaluation relied on voluntary, self-reported data, meaning that biases in reporting or response bias may influence observed outcomes.
- Participants applied to be part of Copilot pilot trial. It is likely that at least some members of the participant group were already familiar with, or motivated to learn about, generative AI. Consequently, the findings for this group may not apply to all other Treasury staff.
- The lack of a robust 'counterfactual', against which any changes in work processes and outcomes could be assessed and attributed to Copilot. It is plausible that outcomes described in this report may be due to external factors unrelated to Copilot access, including motivation to participate in the trial or the passage of time.