Australian Government
The Treasury

Australian
Centre for
Evaluation
ace

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

Evaluation report

February 2025

ISBN: 978-1-923278-07-3

Treasury material used 'as supplied'

Derivative material

Use of the Coat of Arms

Other uses

Enquiries regarding this licence and any other use of this document are welcome at:

Manager
Media Unit
The Treasury
Langton Crescent
Parkes ACT 2600
Email: media@treasury.gov.au

In the spirit of reconciliation, the Treasury acknowledges the Traditional Custodians of country throughout Australia and their connections to land, sea and community. We pay our respect to their Elders past and present and extend that respect to all Aboriginal and Torres Strait Islander peoples.

Acknowledgements

The evaluation of the Copilot Microsoft 365 trial in the Commonwealth Department of the Treasury was completed by the Australian Centre for Evaluation (ACE). The evaluation team comprised: Vera Newman, Ellie McDonald, Brandon Liu, Ethan Slaven and Harry Greenwell from the ACE.

The evaluation was undertaken in collaboration with the Copilot trial project team and the Data and Digital Policy Unit (DDPU). The authors would like to thank all members of the Copilot trial project team, the DDPU and the Treasury Artificial Intelligence Working Group for their contributions to the evaluation.

Lastly, we acknowledge the participants from Treasury's Copilot trial for their contributions to the evaluation and for generously sharing their experiences with the product.

The views expressed in this evaluation report do not necessarily reflect those of the Australian Government.

Contents

Executive summary

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)?

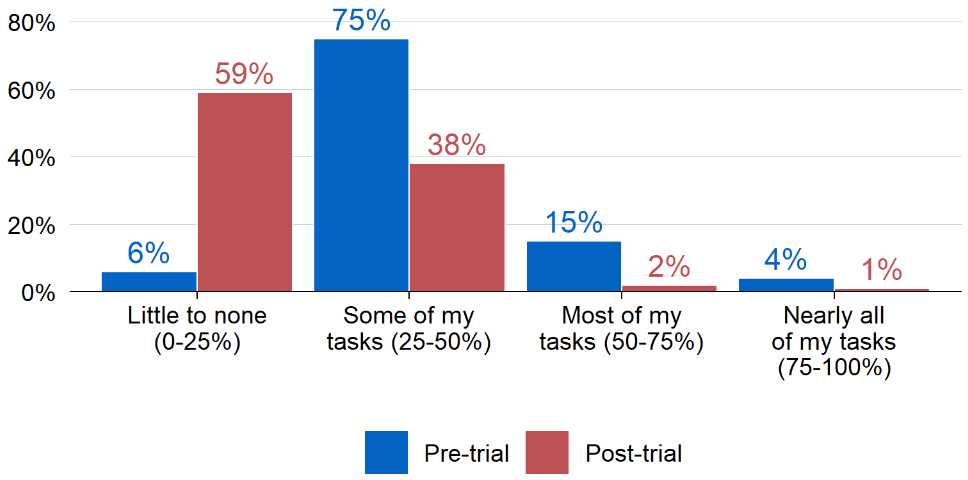
Summary of findings

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

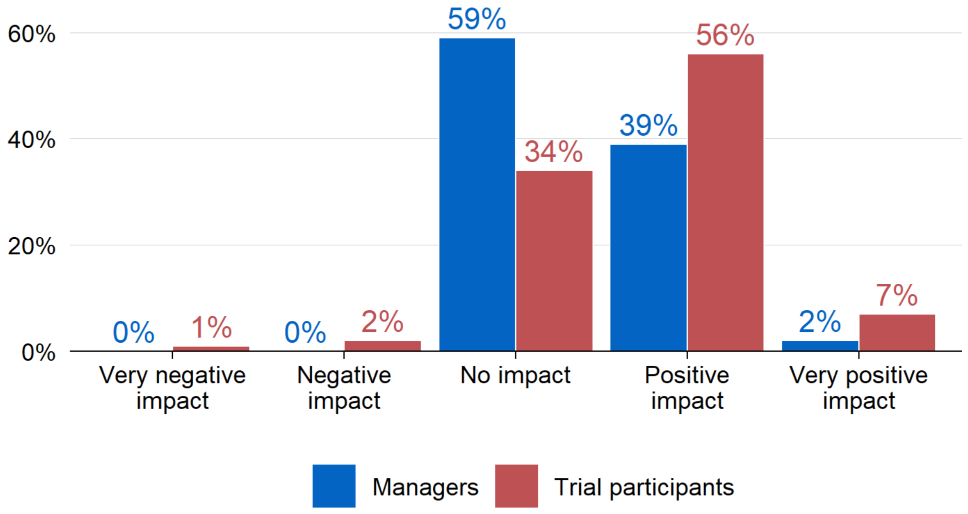


Appropriateness: the initial 'use cases' for Copilot were appropriate for Treasury, however, the product was not suitable for more complex tasks. There were 4 use cases initially proposed for Copilot: generating structured content, supporting knowledge management, synthesising and prioritising information, and undertaking process tasks (see Appendix D 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.

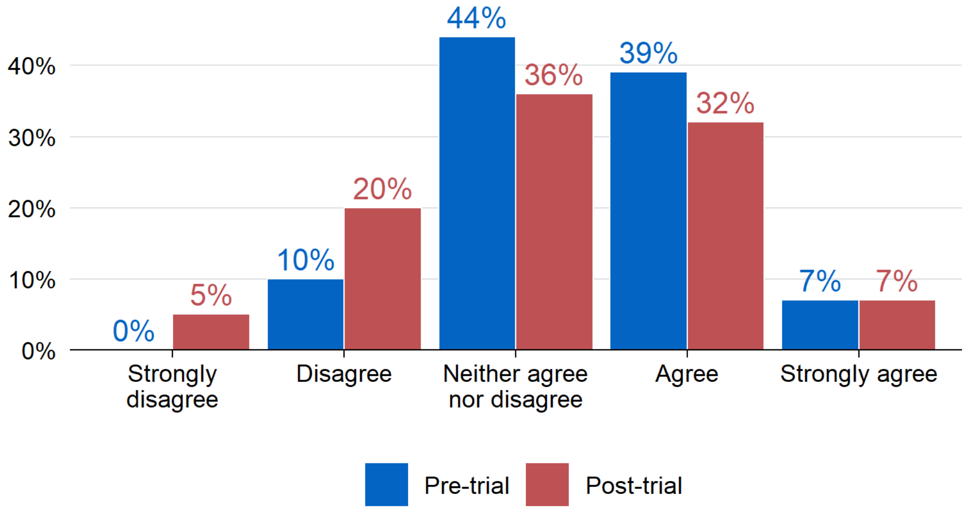
[1] An APS6 staff member is a mid-level position within the Australian Public Service that involves some responsibility and expertise. Further information about the work level standard can be found on the APS Commission's website: https://www.apsc.gov.au/working-aps/aps-employees-and-managers/work-level-standards-aps-level-and-executive-level-classifications

Figure 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.

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: 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.

1. Overview and evaluation approach

Background

The Australian Government ran a trial of Microsoft 365 Copilot (Copilot) between January and June 2024. This whole-of-government trial was coordinated by the Digital Transformation Agency (DTA) with support from the AI in Government Taskforce.

The main objectives of the whole-of-government trial (the DTA trial) were to evaluate:

- Employee-related outcomes: Evaluate APS staff sentiment about the use of Copilot, including:
  a) Staff satisfaction
  b) Innovation opportunities
  c) Confidence in the use of Copilot
  d) Ease of integration into workflow.
- Productivity: Determine if Copilot, as an example of generative AI, benefits APS productivity in terms of
  a) Efficiency
  b) Output quality
  c) Process improvements
  d) Agency ability to delivery on priorities
- Adoption of AI: Determine whether and to what extent Copilot, as an example of generative AI:
  a) Can be implemented in a safe and responsible way across Government
  b) Poses benefits and challenges in the short and longer term
  c) Faces barriers to innovation that may require changes to how the APS delivers on its work.
- Unintended consequences: Identify and understand unintended benefits, consequences, or challenges of implementing Copilot as an example of generative AI and the implications on adoption of generative AI in the APS (Nous Group, 2024)

Treasury participated in the DTA trial but commenced at a later date (due to Budget 2024–25). Instead of starting in January 2024 and ending in June 2024, Treasury started in May 2024 and ended in August 2024. This meant the Treasury Copilot trial overlapped with the DTA trial for approximately 6 weeks, and operated after the DTA trial for 8 weeks.

ACE (the evaluation team) conducted an internal evaluation of Treasury's Copilot trial to capture data regarding the Copilot trial within Treasury, which otherwise would have fallen outside the scope of the DTA's investigation. Treasury applied the same objectives to its trial to inform its program logic (Appendix A: Copilot Trial Program Logic), desired outcomes and key evaluation questions.

Treasury’s implementation of the Copilot trial

Copilot operates within Microsoft Office products, including Word, PowerPoint, Excel, Outlook and Teams. It is currently the only generative AI product deployed within the Microsoft Office suite, which is Treasury’s dominant work platform.

Treasury undertook a 4-phase trial rollout. The first 3 phases focused on onboarding of the Information Services Branch (February–March 2024, n=10) the project team (19 March 2024, n=22) and the Treasury AI Working Group (28 March 2024, n=8). These initial participants were responsible for the implementation, policy and evaluation of the Copilot trial, and therefore received an advance licence to user test the product. The fourth phase focused on onboarding all trial participants (n=218) and 6 senior executives.

In total, Treasury procured and deployed 275 licences. The focus of this evaluation explores the experiences of trial participants only and not the broader project team, AI Working Group (a small group of Treasury staff who supported the trial implementation) or executive experience.

The Copilot trial commenced for 218 participants on 20 May 2024. Treasury chose a date after the 2024–25 Budget to ensure staff working on Budget had the opportunity to express their interest in participating. It also ensured the deployment of Copilot did not interrupt Budget-critical work in the Office suite of products.

As part of the trial, participants were required to:

- attend an onboarding information session
- undertake a SharePoint Online permissions audit
- complete a mandatory ‘Pilot Licence’ training module that included passing a knowledge test with 80 per cent or above
- update their Microsoft Office feature release cycle to monthly
- agree to adhere to relevant policies and guidance when using Copilot.

Engagement for learning purposes during the trial primarily occurred via the Microsoft Teams community chat. This was intended to be a place where participants could seek support to use Copilot, and share successes with one another. In addition, 42 trial participants self-nominated as ‘trial champions’, who were specifically tasked with supporting trial participants in their wider team to use Copilot, and engage in the Microsoft Teams chat more regularly. Champions also held fortnightly meetings to share insights and lessons learnt regarding the product.

The formal trial period closed on 23 August 2024 after a period of 14 weeks, however all trial participants retained their licences following the formal trial period (unless they indicated otherwise).

Table 2 Trial sample composition (based on pre-trial survey)

Characteristic	Percentage[2]
Gender	Male: 62% Female: 36% Non-binary or prefer not to say: 3%
Current role level	Contractor: 3% APS3 and below: 9% APS4: 2% APS5: 18% APS6: 23% EL1: 30% EL2: 12% SESB1 and above: 3%
Group	Fiscal Group: 14% International and Foreign Investment Group: 7% Macroeconomic Group: 12% Markets Group: 36% Revenue Group: 14% Small Business, Housing, Corporate and Law Group: 17%

Copilot use cases

To support the expression of interest process and the trial evaluation, the evaluation team developed 4 use cases. The intent of these use cases was to describe potential ways that participants could use Copilot, but they were not intended to be prescriptive or exhaustive. The 4 use cases were:

1. Generating structured content: For example, developing a detailed project plan, generating first pass content for basic documents, rewriting drafted content to improve writing style, or sourcing relevant factual material.
2. Supporting knowledge management: For example, recording and summarising meeting minutes and generating action items based on meeting discussions, and finding relevant or old documentation and guidance materials.
3. Managing personal or process tasks: For example, summarising or drafting email content, calendar appointments, or collating relevant information to prepare for a meeting.
4. Synthesising and prioritising information: For example, preparing first drafts of data analysis or visualisation, generating code snippets or functions, summarising themes within data, and summarising stakeholder feedback from consultations or other sources.

All use cases included an example of usage, and emphasised the importance of reviewing the outputs for any errors and ensuring its accuracy. Further information regarding use cases can be found in Appendix D.

[2] Some totals do not sum to 100 per cent due to rounding.

Evaluation approach and methods

Prior to the trial, the evaluation team worked with stakeholders across Treasury to construct a suitable program logic for the trial. The program logic sets out the inputs, activities, outputs, and short-, medium- and long-term outcomes that are intended from Copilot. Critically, given the short timeframe of the trial, long-term outcomes were considered out of scope for the evaluation, with data collection therefore focusing on short-term and medium-term outcomes only. The program logic is included in Appendix A.

The evaluation team also constructed key evaluation questions which guided the development of the data collection tools. These questions are detailed in the findings section of the report.

This evaluation used a mixed-methods approach, with data collection from trial participants and their managers via the following mechanisms:

Table 3 Data collection methods

Data method	Sample size	Response rate (from 232)
Pre-trial survey of trial participants	n=153	66 per cent
Post-trial survey of trial participants	n=136	59 per cent
Pulse surveys with trial participants during the trial	Pulse 1: n=68 Pulse 2: n=131 Pulse 3: n=79 Pulse 4: n=61 Pulse 5: n=68	Pulse 1: 29.6 per cent Pulse 2: 56.5 per cent Pulse 3: 33.6 per cent Pulse 4: 26.0 per cent Pulse 5: 29.6 per cent
Survey of the managers of trial participants following the trial	n=49	40 per cent
Focus groups discussions (Champions and participants in July and August 2024)	Champions: n=6 (July); n=4 (August) Participants: n=8 (July); n=7 (August)	N/A
Issues log: collection of issues being experienced with the product logged by participants through a central issues register	N/A	N/A
Case studies: sourced directly from participants highlight different use cases of Copilot	N/A	N/A

Following data collection, ACE descriptively analysed surveys to determine any trends in potential benefits or challenges of access to Copilot. Note that these analyses were descriptive in nature (that is, the analysis presented here describes means and frequency changes). This is due to the limited sample size and potential bias in the sample. Focus group and case studies were thematically analysed: transcripts and focus group notes were coded into themes aligned with the key evaluation questions. Further details about these methods are available in Appendices B and C.

Limitations

The major limitation to the trial relates to its overall length: the trial lasted 14 weeks. It is unlikely that participants' usage of the product reached maturity in this period. This meant that any impacts of Copilot may not have materialised during the trial period due to the limited time in which they could be achieved. Further, Copilot itself is still relatively new as a tool, and has not yet reached maturity, with updates and iterations being released throughout the trial: future updates to Copilot may improve its reliability or impact on outcomes.

A limitation of this evaluation is that participants volunteered to participate in the trial of Copilot within Treasury. As such, there is no robust ‘counterfactual’ to determine what would have happened for these participant volunteers in the absence of Copilot to their work processes, work outcomes, and work stress and satisfaction. Any changes in outcomes may be a result of other factors, including participants’ motivation to participate in the trial, participants’ interest in generative AI, or simply the passage of time.

The sample of trial participants is also likely to be skewed towards those who have prior experience with generative AI, and/or hold favourable opinions about its likely utility within their work. The findings from this trial sample may not be generalisable to the whole of Treasury. This selection is discussed in Section 2: Implementation.

Given the trial commenced on 20 May 2024 (the week following the delivery of Budget 2024–25) it is possible that natural fluctuations in sentiment, stress, output productivity and overall workload influenced outcomes described in this report.

Further, the trial relied primarily on self-reported data, with manager-reported data supplementing some elements of data collection. As a result, it is possible that biases in reporting (for example, social desirability bias)[3] influenced the observed outcomes.

Total survey response rates did not reach 100 per cent, and the sample of matched data across the pre- and post-trial surveys was 100 Treasury staff members. This is because survey participation was voluntary, which may have resulted in response bias. We hypothesise that those who found Copilot particularly unhelpful or did not use Copilot during the trial period were less likely to respond, leading to a likely positive skew in survey results. This is supported by input from the trial coordinators and participants that those who did not have a positive experience with Copilot disengaged from the trial and trial-related communications, potentially including invitations to participate in the surveys.

Structure of this report

This report outlines findings relating to each of the 5 evaluation questions, which focus on Implementation, Appropriateness, Process improvement, Quality outcomes and Unintended consequences.

The recommendations are summarised in Section 7, and are based on the lessons learnt from the implementation of Copilot within Treasury. These are therefore elements that any future implementation of generative AI within Treasury should consider. However, ACE recognises that many of these recommendations are contingent on whether Treasury proceeds with a future rollout of generative AI, and the funding and resources available. These recommendations are also product-agnostic, meaning that while they were developed from Treasury’s experience with Copilot, they apply to the rollout of any generative AI products in the future. Finally, ACE recognises that this evaluation is one of multiple inputs into decision-making regarding any future implementation of generative AI.

[3] Social desirability bias is the tendency to provide responses to researchers that will be viewed positively by others, including the researchers (Krumpal, 2013).

2. Implementation

This section addresses the first evaluation question: *To what extent was Copilot implemented as intended?*

Technical implementation of Copilot was relatively smooth, with most trial participants onboarded within 2–3 days. However, despite there being no barriers to access, there was lower use of Copilot than anticipated during the trial period. This appears to be due to several factors including limited technical capability of the product, and a need for more education and training on product use. As a result, a lot of participants did not use the product as much as they initially intended. The issues impacting the implementation of Copilot are further outlined in this section.

A Microsoft representative suggested to the Treasury Copilot project team that it takes up to 12 weeks for an individual to learn how to use the product effectively, which was similar to the length of the trial period. Microsoft highlights in a recent blog that:

> like all new technology, employees will utilise features more effectively as usage and familiarity increases. Ideally, we recommend allowing 2–3 months after adoption, per our research, before conducting analysis of Copilot impact. (Microsoft, 2024).

Therefore, despite the challenges trial participants experienced using Copilot, some issues and learning barriers should be expected while staff transition to using a generative AI product. As such, providing staff with sufficient resources, education and time to learn how to use the tool will reduce any barriers to product use.

The technical implementation of the product was smooth

When considering the technical implementation of the product within Treasury, there was an overarching consensus among trial participants that the implementation was successful. Some participants reported experiencing minor glitches at the beginning of the trial during onboarding, such as uncertainty regarding whether the product had been installed correctly, and the product not being integrated into all Microsoft products (including Outlook in the initial phases of the trial).

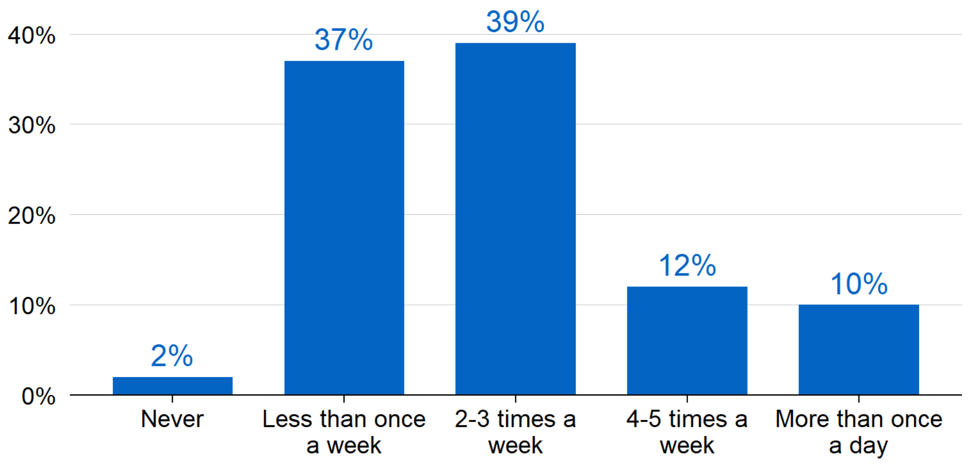
One issue arose for multiple users during a routine Treasury IT Microsoft update whereby multiple users reported losing access to Copilot due to being pushed onto the wrong Microsoft update setting. This issue temporarily restricted access to Copilot for affected trial participants but was quickly resolved via emails and Teams messages to participants, and an additional automatic computer update where required.

Many of the issues that trial participants reported related to limitations of the Copilot product itself (that is, issues with the outputs prompts generated, or usability of the product), rather than major issues with the product's technical implementation within Treasury. We therefore discuss participants' experiences of using the product in further detail throughout the remainder of the report.

One-quarter of participants reported using Copilot frequently by the end of the trial

While trial participants reported using Copilot to some extent across the trial period and for the specified use cases (see further information on Copilot use cases in Section 3: Appropriateness), the product was used less than originally expected. Despite this lower reported use than anticipated, only 2 per cent reported that they did not use Copilot during the trial period (Figure 4).[4]

Figure 4 Participants' self-reported frequency of Copilot use



Source: Post-trial survey: 'How frequently did you use Copilot for work-related tasks throughout the Copilot trial period?'
Note: N = 136.

Almost one-quarter of post-trial survey respondents (22 per cent) reported using Copilot 4–5 times per week or more, with one in ten (10 per cent) using Copilot more than once a day. These proportions are broadly consistent with those seen in the pulse surveys, where 26 to 31 -per cent of pulse respondents indicated they used Copilot 4–5 times per week or more in the preceding fortnight, across all pulse surveys. This suggests that while the average user may have not used Copilot as much as was expected, a small number of 'super users', who used the product approximately once per day or more, persisted with the product throughout the trial period.

Lower use of the product is likely to have emerged due to several factors. This included initial high expectations of the product that diminished when trial participants experienced the reality, technology limitations, and a lack of education and training on how to use the product.

Issues with Copilot limited the overall use of the product

There were several mutually reinforcing issues that limited the usefulness and therefore the use of Copilot among trial participants:

1. Trial participants reported **difficulties and concerns with 'prompt engineering'**, including difficulties finding the correct prompt to use, unhelpful outputs from prompts, and low-quality outputs to complex tasks. That is, participants reported that, even if they could find the correct prompt, the outputs Copilot gave were sometimes inaccurate or inconsistent over time, which reduced the perceived reliability of the product.

[4] Note that frequency of Copilot usage was self-reported via the post-trial survey. It is possible that trial participants who did not use the product did not respond to the post-trial survey; if this is the case, we would expect higher reports of never using the product than is observed in the data.

> *Learning the prompts [is] challenging, like trying to learn a new language.*
>
> ***Trial participant,*** Focus Group discussion

2. Participants reported **concerns about the reliability and accuracy of the responses Copilot** provided in their work, meaning that some participants stopped using Copilot. In particular, participants highlighted limitations in incorporating contextual information into Copilot's responses, and misattributions of statements in documents leading to incorrect summaries. These limitations sometimes led to the generation of fictional content, which suggested to some that it was not as reliable as initially expected.

> *Co-pilot [sic] often created fictional information when asking it to generate output.*
>
> ***Trial participant,*** Post-trial survey

> *After a few early tests, there seemed to be obvious errors which reduced my confidence in using co-pilot [sic] for this purpose.*
>
> ***Trial participant,*** Post-trial survey

3. **Limitations of the product itself** meant that Copilot could not be used when working across multiple Microsoft applications (such as Word and Outlook and PowerPoint), when working in PDFs, and when incorporating information from a large number of files. Given the nature of the work completed within Treasury, which typically involves using multiple applications at once, this was seen as a restriction of Copilot's usefulness in daily work. These limitations are likely to have arisen due to Treasury's security environment.

> *Having Copilot inbuilt into every app was overwhelming, as it was unclear what to do in different apps, as each app required a different approach to use the product.*
>
> ***Trial participant,*** Focus Group discussion

4. For some tasks, using **Copilot was not reliably more efficient compared to completing the task manually**. While many participants in the focus groups and surveys cited improvements in efficiency when using Copilot for administrative and process tasks, there were also experiences of inefficiency. Participants emphasised that some coaching of the product was often required to obtain an adequate response and that responses needed manual checking to verify accuracy and appropriateness for the relevant audience. This process was sometimes more time intensive than creating the output themselves, thereby removing the incentive to use the product as a time saving device.

> *You can sink a lot of time into Copilot and still not get what you need.*
>
> ***Trial participant,*** Focus Group discussion

> *It takes time and effort to produce useful co-pilot [sic] output, so at some point I would stop and just do it on my own – using a tiny bit of what I might have created. I also struggled with the fact that Copilot could not read structured content such as tables or headings.*
>
> ***Trial participant,*** Post-trial survey

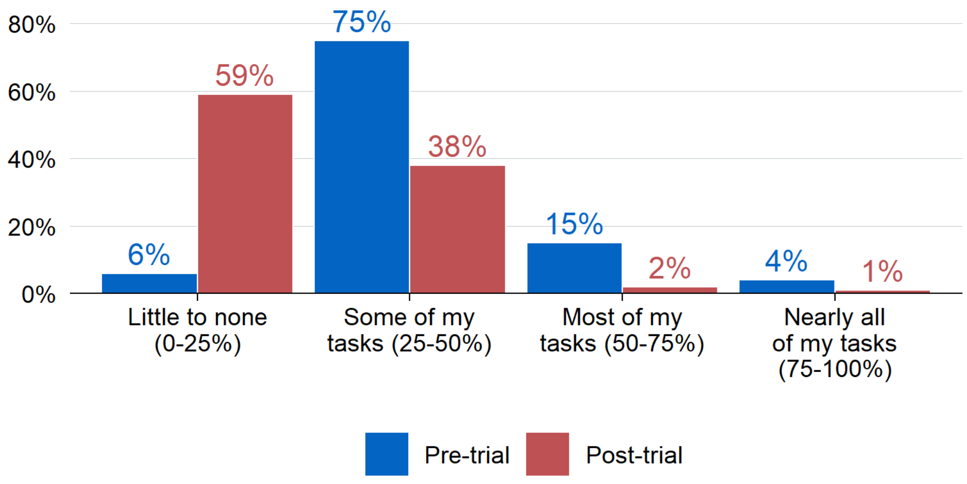
These 4 factors influenced and reinforced one another, which likely contributed to the overall lower-than-expected levels of reported Copilot usage during the trial.

Use of the product was adversely impacted by the gap between expectations and reality

To understand how Copilot could support participants' work, trial participants indicated the percentage of tasks they believed could be (pre-trial) or have been (post-trial) performed more effectively using Copilot. Prior to the trial, most trial participants reported an expectation that Copilot could help perform at least some of their tasks more effectively.

However, these expectations tempered substantially throughout the trial. Following the trial, 59 per cent of post-trial survey respondents reported that Copilot supported them to perform little to none (0 to 25 per cent) of their weekly workload more effectively.

Figure 5 Participants' self-reported percentage of workload Copilot could/did support

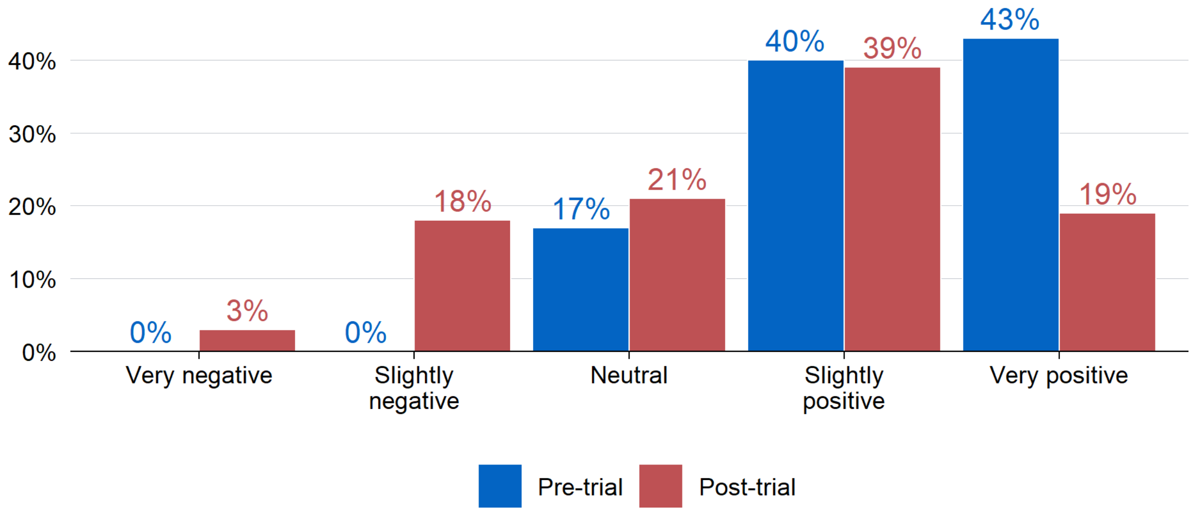


Source: Pre-trial survey: 'How much of your current weekly workload could be performed more effectively with the help of automation?'; Post-trial survey: 'How much of your current weekly workload has been performed more effectively with the help of Copilot's automation?'
Note: N = 100; including only respondents who could be matched across pre- and post-trial surveys by their provided email address.

Positive sentiment towards Copilot use at Treasury decreased from pre-trial to post-trial. Prior to the trial, most respondents in the pre-trial survey were positive about Copilot use at Treasury, and no trial participants reported negative sentiments toward Copilot use. This changed following the trial: one in 5 (21 per cent) post-trial survey respondents reported negative sentiment toward using Copilot at Treasury (Figure 6). Similar sentiments were also shared by both trial Champions and trial participants in the focus groups.

> *By comparison to the use of GAI [generative AI] on public domain information, the Co-Pilot [sic] implementation did not produce outputs that were good enough to improve efficiency or effectiveness. So much editing and work was required that initial enthusiasm faded quickly.*
>
> ***Trial participant,*** Post-trial survey

Figure 6 Participants' self-reported sentiment regarding Copilot



Source: Pre-trial survey: 'Which of the following best describes your sentiment about using Copilot at Treasury?'; Post-trial survey: 'Which of the following best describes your sentiment about using Copilot at Treasury'
Note: N = 100; including only respondents who could be matched across pre- and post-trial surveys by their provided email address.

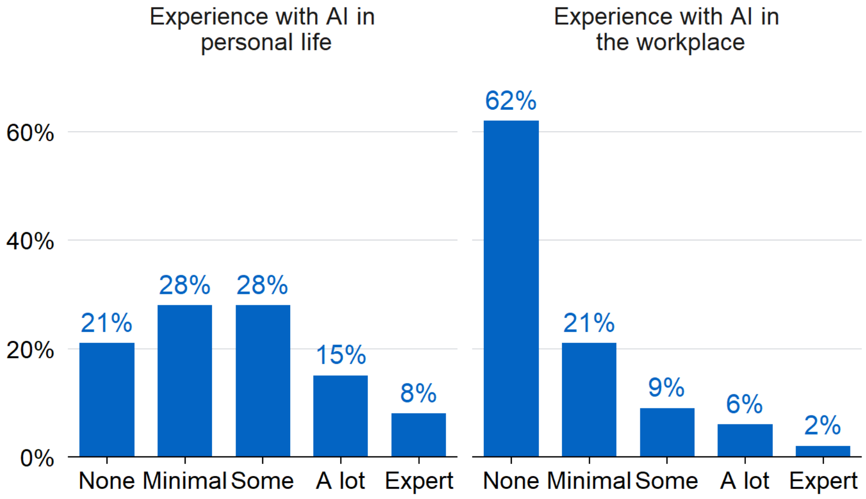
This gap between initial high expectations for Copilot and its reality is likely to have influenced use as participants discovered the product's limitations throughout the trial. For example, although 61 per cent of trial participants reported using the product at least twice per week (Figure 4), the number of pulse survey respondents indicating they used Copilot either never or less than once per week increased as a share of the total survey respondents over time: from 24 per cent in pulse survey 2, to 29 per cent, 38 per cent, and 40 per cent in the subsequent pulse surveys. Although these pulse surveys suffered from a low response rate and therefore potential selection issues, they can be taken as an indication of decreasing rates of usage of Copilot for non-super users over the trial period.

> *Noticed disengagement when attempts to produce content were inaccurate or unhelpful.*
>
> ***Trial participant,*** Post-trial survey

Part of this effect is likely due to a bias in the sample of trial participants. Employees across Treasury were offered the opportunity to sign-up to participate in the Copilot trial via an expression of interest process. As part of this process, potential trial participants were asked to provide an indication of how they intended to use Copilot. Further, trial recruitment materials such as the Treasury SharePoint page indicated that the product is powered by the same technology as ChatGPT, and may deliver productivity benefits.

This approach for recruiting trial participants is likely to have led to a selection bias for a variety of reasons. It is plausible that employees with previous or positive experiences in using generative AI – and particularly ChatGPT – were more likely to express interest in participating, and therefore end up in the trial. This was partially borne out in the data (Figure 7). Seventy-nine per cent of pre-trial survey respondents indicated they had at least minimal experience with generative AI in their personal lives. It is also possible that employees seeking productivity enhancements also expressed interest in participating, regardless of their level of experience using generative AI; this cannot be tested using available data. Both mechanisms would result in a sample of trial participants with relatively high expectations regarding the potential benefits of Copilot.

Figure 7 Participants' reported experience with AI prior to the trial



Source: Pre-trial survey: 'How much experience do you have with using generative AI tools: a) in the workplace? b) in your personal life'
Note: N = 153.

There was scope for more education and training to support participant onboarding

One of the most common needs participants expressed during the focus groups and in the pulse surveys was greater levels of support during the trial, particularly through targeted and tailored education and training. This request for further education was persistent throughout the trial period.

> *More sustained, targeted training about use cases would have been great. It was hard to know exactly how I could have solved different problems using AI – by the time I got through working out how I could save time, I had run out of time to actually do the work.*
>
> ***Trial participant,*** Post-trial survey

> *I think it would be good to continue using Co-pilot [sic], but investment in capability and education (similar to what Treasury does for knowledge management) is needed.*
>
> ***Trial participant,*** Post-trial survey

Due to limited trial resources, Treasury trial onboarding was limited to providing an onboarding session to participants and the basic Digital Transformation Agency generative AI training via Treasury's online learning platform. The onboarding session included information about Treasury's trial of Copilot, trial requirements, and an indication and explanation of the trial's specified use cases. However, providing further training or individualised support to trial participants was outside of scope and available resourcing.

Strategies to support participants to learn and use Copilot included:

- a Champion's Network
- a Copilot Community of Excellence
- a Copilot Trial Community chat on Microsoft Teams
- Copilot Lab and Microsoft training sessions
- Examples of general prompts.

Although the Treasury Project Team put several strategies in place to provide information, resources and guidance to trial participants, many of these resources were underutilised. The only consistently active support network was the Copilot Trial Community chat.

The Champions Network, for example, was intended to be a resource all trial participants could draw on for support in undertaking specific tasks. Champions were expected to engage with trial participants within their branches, via the Teams platform and share their experiences and tips on how to use Copilot more effectively or innovatively. In practice, the overall level of Champion engagement was low. While there were some enthusiastic Champions who engaged with trial participants via the Teams channels, many Champions demonstrated limited engagement in these channels and in Champion's meetings. This ultimately resulted in limited support to trial participants.

In addition to the informal training supports employed within Treasury, trial participants were offered access to external supports such as Copilot Lab and Microsoft training sessions. Since these training offerings were external to Treasury's systems, there is no information about their uptake. However, given participant requests for additional training, it is likely the perceived value of these external supports was limited. It is possible that such external training offerings did not provide the Treasury- or role-specific guidance participants needed.

The experiences of Treasury staff are consistent with the DTA's whole-of-government evaluation of the Copilot trial, which found that:

> There was a positive relationship between the provision of training and capability to use Copilot. Copilot training was most effective when tailored to the APS, the users' role and the agency context. (Nous Group, 2024)

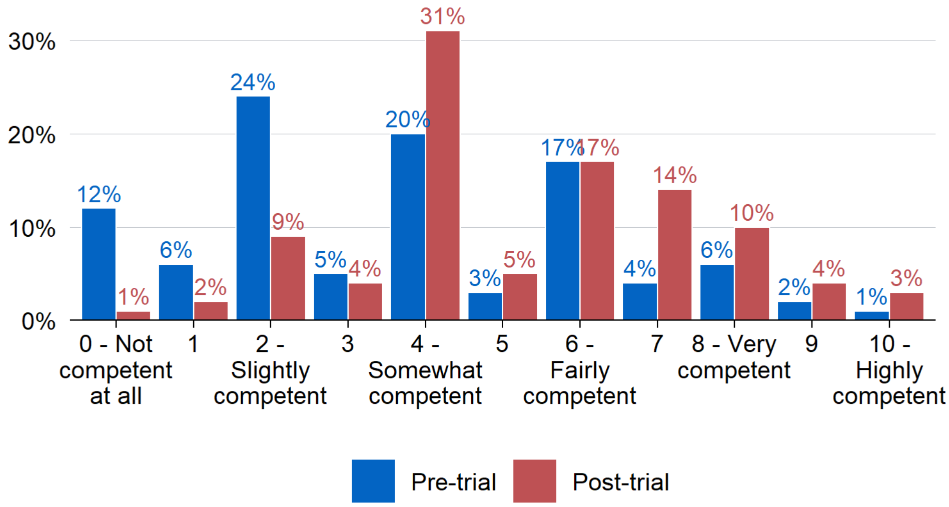
Further, online reports indicate that while many workplaces are adopting generative AI products, specific and effective training remains limited.

Despite the limited opportunities for engagement with formal training, survey respondents told us their competence to use Copilot increased. As is shown in Figure 8, the distribution of self-reported competence ratings moved toward higher competence following the trial, compared to prior to the trial. Prior to the trial, 30 per cent of survey respondents indicated they felt at least 'fairly competent' in their ability to use Copilot. This increased by 18 percentage points to 48 per cent of respondents post-trial.

This suggests that, at least for some participants, their competence increased with exposure to Copilot. It also indicates that on average, the trial progressed towards the short-term outcome: 'Participants have competence to use and confidence to experiment with the Copilot tool' (Table 1).

However, it is not clear how much more self-reported competence may have increased if additional training or supports were available.

Figure 8 Participants' self-reported competence using Copilot in their role



Source: Pre-trial survey: 'How competent do you currently feel using Copilot in your role?'; Post-trial survey: 'How competent do you currently feel using Copilot in your role?'
Note: N = 100; including only respondents who could be matched across pre- and post-trial surveys by their provided email address.

While many participants experienced issues with Copilot, these teething issues in navigating and using Copilot were expected. This is because it is common that individuals require time to learn how to use new products in their work, and this extends to generative AI products. In addition, prompt engineering requires a specific skill set that may require investing in new capabilities specifically for using generative AI in future.

> *I have noticed a lot of Treasury employees don't know how to use Co-Pilot [sic] properly. Ergo, it is essential to consider giving more guidance and virtual tutorials to staff.*
>
> ***Trial participant,*** Post-trial survey

3. Appropriateness

This section addresses the second evaluation question: *To what extent is Copilot appropriate in Treasury's context?*

Evidence gathered during the trial period suggests that Copilot was an appropriate tool for elements of the identified use cases – those which were basic administrative tasks and processes. However, data from focus group discussions, pulse and post-trial surveys indicated that the product's technology is still new, relative to other available products in the market. This limited the product's reliability for more complex tasks. In addition, the sector will need to undertake further work to guide how staff reference the use of generative AI in their work to ensure transparency and public trust.

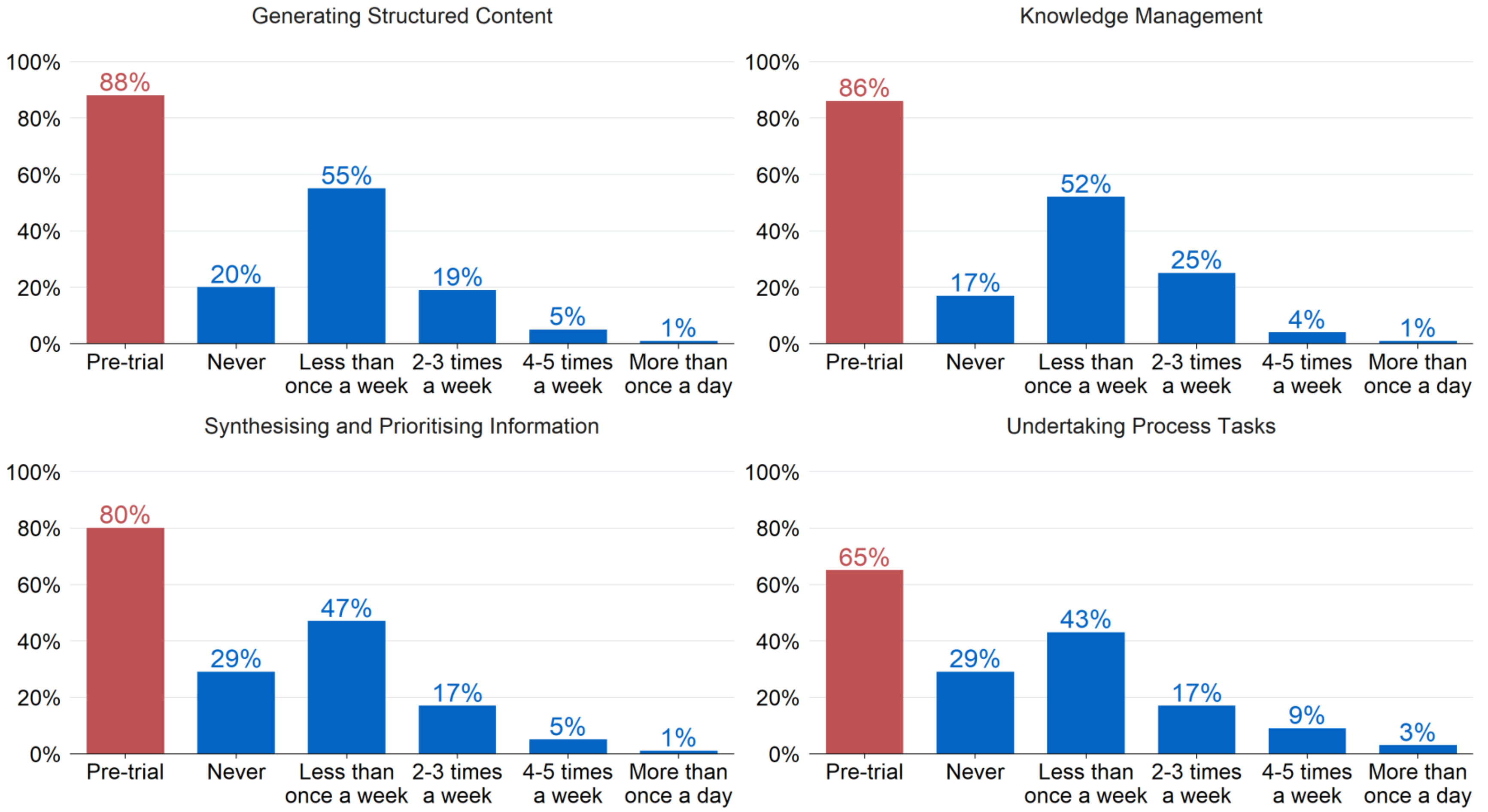
Copilot was appropriate to use for the identified use cases

Respondents in the post-trial survey indicated that they used Copilot for the identified use cases (outlined in Figure 9, with use cases described in detail in Appendix D). Participants in focus groups outlined broad areas where Copilot was useful for their work, including mostly process-related tasks such as those outlined in the use cases. In combination, this evidence suggests participants perceived that, for the most part, Copilot was an appropriate tool for Treasury staff to use for the identified use cases.

> *CoPilot [sic] proves to be consistently accurate in identifying clear deliverables and action items, making it a valuable tool for administrative tasks like meeting minutes.*
>
> ***Trial participant,*** Focus Group discussion

While there were higher expectations of use pre-trial, 70–80 per cent of trial participants reported using Copilot for all use cases; and approximately 25 to 30 per cent reported using Copilot for each use case more than twice per week during the trial (Figure 9). 'Knowledge management' and 'Generating structured content' were reported as the most common use cases participants used during the trial, with 82 per cent and 80 per cent of participants endorsing use for these use cases respectively. This highlights that the intended Copilot use cases aligned with how Treasury staff reported using the product. It also indicated that the trial progressed toward the short-term outcome: 'Participants are using Copilot for the identified use cases' (Table 1).

Figure 9 Trial participant self-reported intended and actual usage of Copilot for nominated use cases



Source: Pre-trial data is taken from the Expression of Interest process: respondents were asked to indicate which use cases they intended to use Copilot for. Post-trial survey: 'How frequently did you use Copilot for the following types of tasks?'
Note: N = 202 for pre-trial expression of interest data, and N = 136 for post-trial data.

In contrast, trial participants indicated that Copilot was not suitable for more complex tasks, and in many cases attempting to use Copilot for these tasks reduced the efficiency of their work. Reasons included Copilot's limited technical functionality and the time it took to learn how to use the product (these issues are elaborated on further in Section 2: Implementation).

Overall functionality is limited compared to other products on the market

Some participants highlighted that there are other generative AI tools that have considerably more functionality (including for analysis or coding purposes), and could be more beneficial to Treasury's work, such as products including ChatGPT and Claude.

Feedback from Copilot Champions highlighted that other generative AI products are more intuitive and provide higher quality responses. As such, while generative AI technology is relatively new and the products will only improve over time, the general perception among trial participants was that Copilot did not compare to other products on the market at this time.

> *It's not as good as other LLM [Large Language Models] out there, when doing coding I would use GPT4.*
>
> ***Trial participant,*** Focus Group discussion

> *I have found in my personal use that ChatGPT does a much better job at performing similar tasks and is better able to pull useful information from the internet.*
>
> ***Trial participant,*** Post-trial survey

One reason for this was the alternative products' ability to draw on data and information outside of Treasury's systems. Consistent with this, some participants identified that the unrestricted version of Copilot performed better than Copilot limited to Treasury's internal systems, indicating that Treasury's necessary privacy and security restrictions limited the product's quality.

> *… given the inability (or limited ability) of Treasury Copilot to connect to the internet, I have had less optimal outcomes compared to my personal (unrestricted) Copilot license.*
>
> ***Trial participant,*** Focus Group discussion

Given Treasury's privacy requirements to ensure protected data and information is not externally accessible, it is unclear whether alternative products would meet Treasury's needs. It is also unclear if further advances in generative AI products will enable Treasury to use a tool that draws on external information in the future.

The use of generative AI requires transparency so as not to erode public trust in institutions

While the product was generally perceived as appropriate for use in a Treasury context, quite a few participants indicated that they were unsure how to communicate the use of the product in their work to their managers, teams and senior executives. Trial participants highlighted that they wanted to be transparent in their use of Copilot but that there were not clear guidelines on when and how to disclose the use of Copilot in Treasury's work. This was particularly important when using Copilot for more complex tasks, such as drafting briefs and reports.

> *It was hard to find the appropriate level of transparency to explain whether you used Copilot to your manager or SES, credit to Copilot for work.*
>
> ***Trial participant,*** Focus Group discussion

The DTA has drafted a Policy for the Responsible Use of AI in Government and standards for accountable officials. While these documents provide high-level guidance on how to use generative AI responsibly in government, there is no direct reference to how to disclose the use of generative AI in staff work. These documents were also not available to trial participants during the trial. Development and dissemination of transparency and disclosure guidelines in the future could further support Treasury staff to feel confident in their use of generative AI for official work purposes.

> *It would be great to better understand how to give credit to Copilot in our work.*
>
> ***Trial participant,*** Focus Group discussion

Some trial participants were also conscious of the need to ensure that the use of generative AI did not reduce or disrupt public trust in institutions. This will require a cultural shift, which will take time. Further, it requires a recognition of the different ways that generative AI can be used within government – for example, there are likely different implications of the use of generative AI for frontline service delivery, versus in the preparation of specific internal documentation or basic administrative tasks. As a result, transparency and accountability in the use of generative AI products such as Copilot is even more important to uphold trust in government.

4. Process improvement

This section addresses the third evaluation question: *To what extent does Copilot support process improvement?*

Overall, benefits to work processes were the most well-evidenced outcomes from the Copilot trial. Trial participants indicated through all the evaluation's data sources that Copilot provided the most value to their work by improving basic administrative processes, contributing to idea development and supporting knowledge management. There were also reports that Copilot saved participants time on basic administrative tasks, although time saved was not explicitly measured. Examples of this included using Copilot to review multiple procurement documents and to support coding process to save up to 4–6 hours of work. Despite this, some participants did experience issues with low quality responses, especially with more complex tasks.

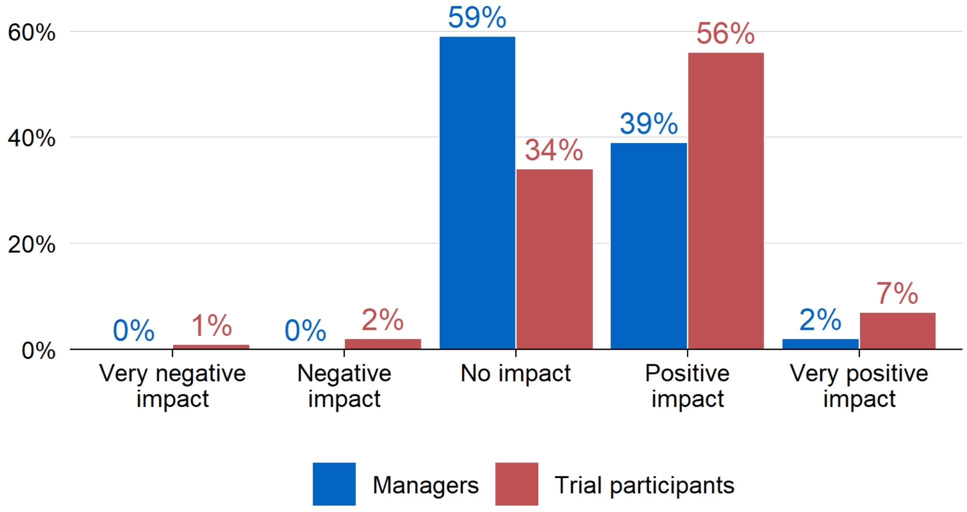
Copilot supported improved work processes

Trial participants found that Copilot was most useful at supporting work processes including summarising meeting minutes, finding files and information on SharePoint online, developing draft plans and documents, and adapting the tone of writing.

Overall, 63 per cent of respondents in the post-trial survey indicated that Copilot had some form of positive impact on work processes with just less than one in 10 (7 per cent) reporting a very positive impact (Figure 10). This is also reinforced by reports from managers of trial participants: 41 per cent of responding managers reported that Copilot had some positive impact on their staff's work processes (and the remainder of respondents reported no impact).

This suggests that through the period of the trial, the implementation of Copilot did support achievement towards the short-term outcome: 'Participants indicate an increase in process improvement using Copilot' (Table 1).

Figure 10 Manager and staff ratings of the influence of Copilot on work processes



Source: Manager survey: 'For the staff you manage who are participating in the Copilot trial, based on your experience, what average impact has Copilot had on the team members' work processes (i.e., how they go about creating outputs)?'
Post-trial survey: 'What impact has Copilot had on your work processes to date? (e.g., work processes could include searching for emails or docs in SharePoint, process of brainstorming or developing meeting minutes)'
Note: Manager survey N = 49; Post-trial participant survey N = 134.

In the pulse surveys, 45 per cent (in the first pulse survey) to 69 per cent reported some form of positive impact of Copilot on work processes. The positive impact hovered around this level, and trended upwards over the trial.

Reflections from focus group discussions also provided evidence Copilot supported process improvement, identifying that Copilot was useful for organising meeting times, developing tables in Microsoft Word, summarising notes and using it as a thesaurus tool.

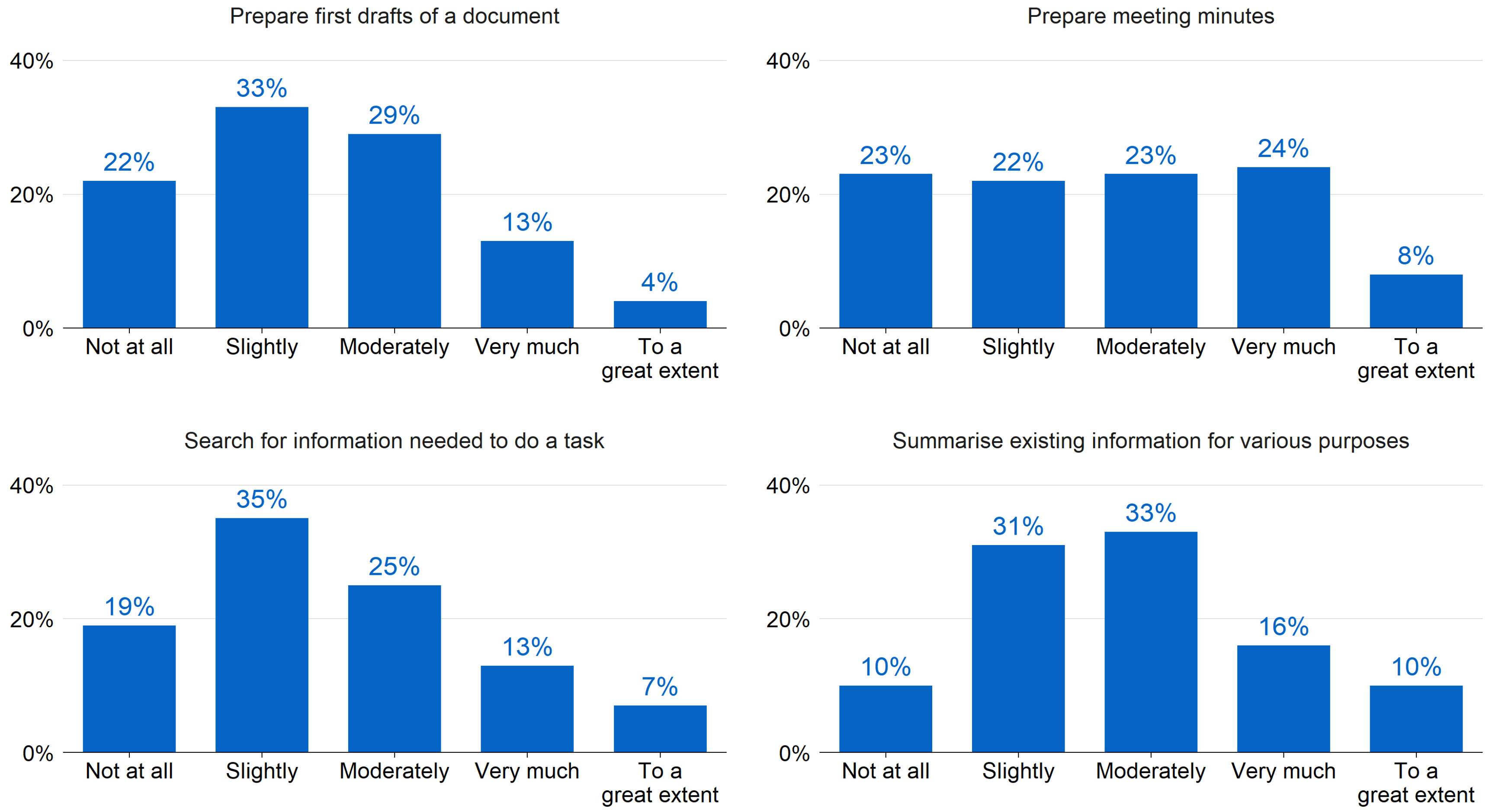
> *It was useful for idea generation – putting in the questions someone else had provided into Copilot to expand your mind.*
>
> ***Trial participant,*** Focus Group discussion

> *One way it helped processes were to organise meeting times with stakeholders – Copilot told you who would miss out on the meeting and who was going to be there.*
>
> ***Trial participant,*** Focus Group discussion

When considering uses of Copilot outside of the specified use cases, respondents to the post-trial survey indicated that Copilot contributed to some improvements in process tasks, including processes to:

- prepare first drafts: 46 per cent reported at least moderate improvement
- prepare meeting minutes: 55 per cent reported at least moderate improvement
- search for information needed to do a task: 45 per cent reported at least moderate improvement
- summarise existing information for various purposes: 59 per cent reported at least moderate improvement.

Figure 11 Trial participant self-reported usage for further process tasks

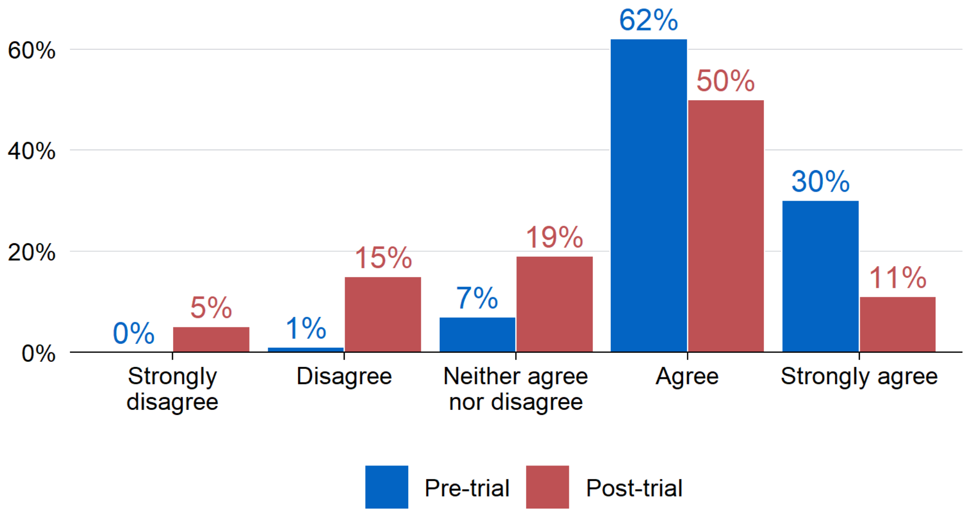


Source: Post-trial survey: ‘To what extent has Copilot improved processes to’ a) Prepare first drafts of a document; b) Prepare meeting minutes; c) Search for information needed to do a task; d) Summarise existing information for various purposes’. Note: N = 134.

Many trial participants reported reduced time on basic work processes and tasks

Most trial participants highlighted that they experienced time savings in their work when using Copilot for basic administrative work and processes. Sixty-one per cent of post-trial survey respondents indicated that Copilot reduced the amount of time they spent on low-value or low-priority tasks.

Figure 12 Participants' perceived value of Copilot for low-priority tasks



Source: Pre-trial survey: 'To what extent do you agree or disagree with the following statements? Using Copilot will: Reduce the amount of time I spend on low-value or low-priority tasks'; Post-trial survey: 'To what extent do you agree or disagree with the following statements? Using Copilot has: Reduced the amount of time I spend on low-value or low-priority tasks'
Note: N = 100; including only respondents who could be matched across pre- and post-trial surveys by their provided email address.

Case study: Procurement

What: Review multiple quotes from Service Providers in an efficient manner against an evaluation criteria.

Prompt: Hey Copilot! Please review/*(service provider)documents*, and provide an assessment against */Service Provider Evaluation Report*. Please provide a list of pros and cons against each criteria.

Value: Service Providers can often provide more information than required when responding to a Request for Quote.

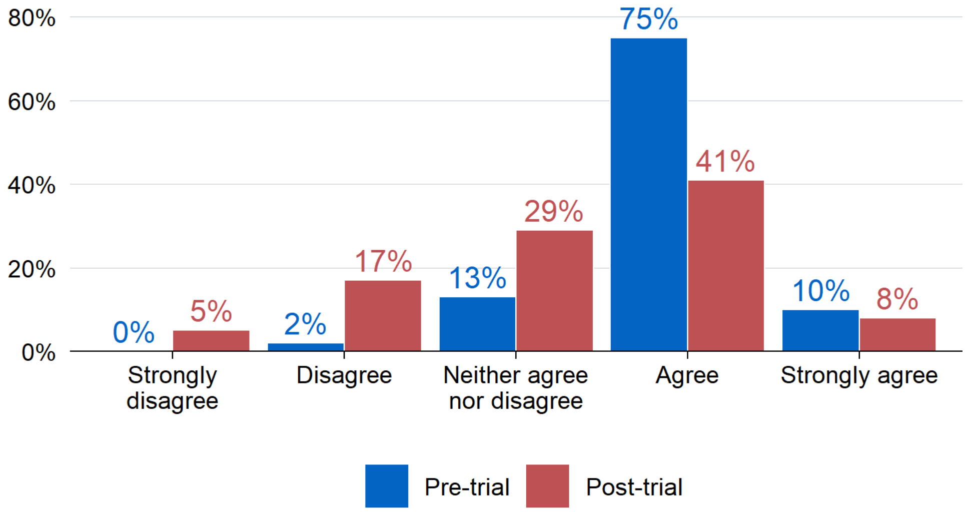
This process has historically taken me around 6–7 hours to complete based on three quotes received. This is to collate the information, review the packages, provide an individual summary on the template against each evaluation criteria and make an informed decision. I am experienced in this, so would say that 6–7 hours is efficient. Less experienced staff would take much longer.

Copilot assisted me to complete this task in less than an hour, saving approx. 6 hours, even with cross checking Copilots results for accuracy and making tweaks where necessary.

Post-trial survey results show that 49 per cent of respondents agreed or strongly agreed that Copilot improved the speed at which they completed tasks (with 22 per cent disagreeing or strongly disagreeing to this statement). While this is lower than the initial expectations reported in the pre-trial survey (85 per cent reported expecting that Copilot would improve the speed at which they complete tasks), it does demonstrate that some users have experienced efficiencies in their work. ACE estimates that a current APS6[5] staff member would need to redirect approximately 13 minutes of time per week to higher value tasks for the licence cost to be offset. Although the evaluation did not explicitly measure time savings for specific tasks, the benefits reported by participants via case studies and in the post-trial survey provide some evidence that Copilot has, for the most part, achieved these time savings.

[5] An APS6 is the classification for a mid-level position. in the APS. Further information about APS level and Executive Level classifications can be found on the Australian Public Service Commission’s website: https://www.apsc.gov.au/working-aps/aps-employees-and-managers/work-level-standards-aps-level-and-executive-level-classifications

Figure 13 Participants' perceived value of Copilot for improving speed of task completion



Source: Pre-trial survey: 'To what extent do you agree or disagree with the following statements? Using Copilot will: Improve the speed at which I complete tasks'; Post-trial survey: 'To what extent do you agree or disagree with the following statements? Using Copilot has: Improved the speed at which I complete tasks'
Note: N = 100; including only respondents who could be matched across pre- and post-trial surveys by their provided email address.

> *[Copilot]… has enabled me to step back from time consuming, process-driven tasks and focus my energy on higher-value strategic tasks.*
>
> ***Trial participant,*** Post-trial survey

> *Helpful to get a scaffold of a document, draw key points from specific docs, Copilot was effective for doing this – took away the time to format the document.*
>
> ***Trial participant,*** Focus Group discussion

Case study: Creating a SAS Macro

What: Shorten my SAS program making it easier to read and faster to run.

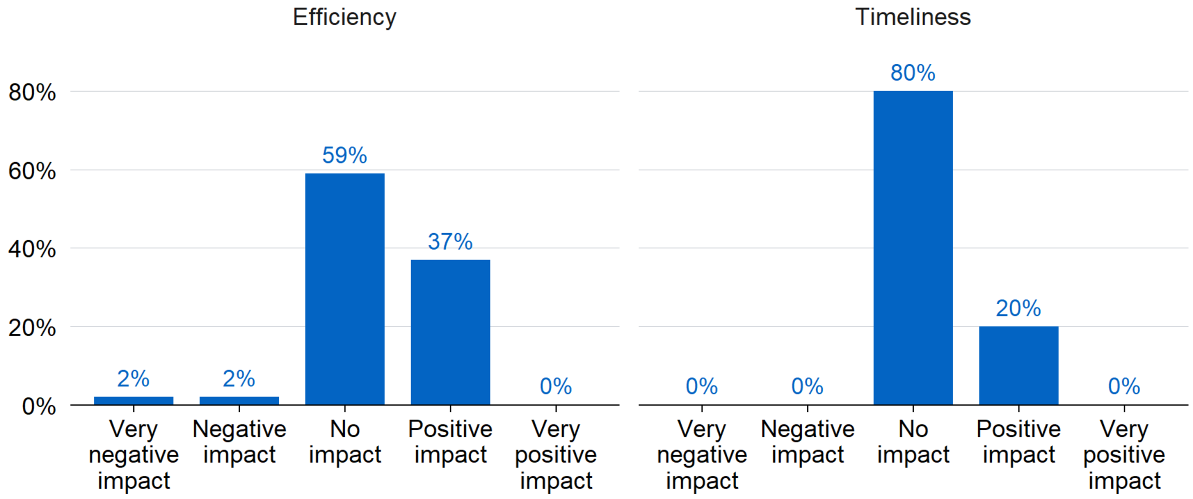
Prompt: Create a SAS macro that merges datasets of different years with year being the input.

Value: I don't have an advanced level of SAS programming knowledge but enough to know what I want to do and to be able to read most code. That meant I was able to prompt Copilot with what I wanted it to do and check if it responded correctly.

I was able to save a lot of time by needing to develop the macro myself. Using macros is good coding practice as it reduces the chance of errors and improves readability.

Reports from managers were more ambivalent about Copilot's effects on their team's work: 80 per cent of managers reported no impact in the timeliness of their team's work, and 59 per cent reported no impact on the efficiency of their team's work. This suggests that despite Copilot contributing to an individual's personal experiences of improved efficiencies on basic administrative tasks, the product has not gone as far as improving noticeable work efficiencies across entire teams.

Figure 14 Manager ratings of the influence of Copilot on timeliness and efficiency of staff outputs



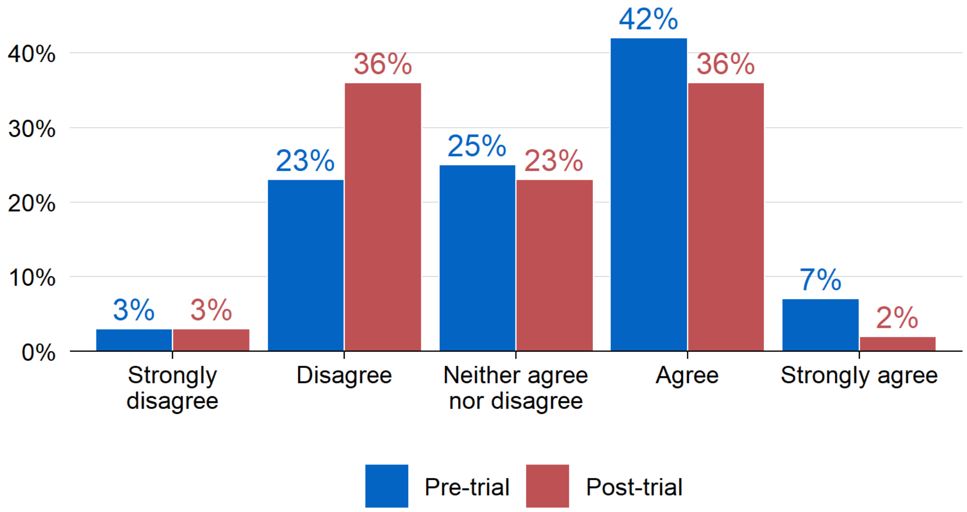
Source: Manager survey: 'For the staff you manage who are participating in the Copilot trial, based on your experience, what average impact has Copilot had on: a) the efficiency of your team members' work, and b) the timeliness of your team members' work?'
Note: N = 49.

Copilot has supported knowledge management in some circumstances

There is some evidence that Copilot supported improved knowledge management by facilitating identification of relevant documents and encouraging knowledge management protocols. This indicates progress toward the medium-term outcome: 'Copilot is supporting improved knowledge management' (Table 1). Across the trial period, 82 per cent of respondents reported they used Copilot for knowledge management, but half of respondents (52 per cent) reported doing so fewer than once a week (Figure 9). This aligns with the results in Figure 11, demonstrating that almost half of respondents indicated that Copilot led to at least a moderate improvement in processes to search for information to complete a task.

During the trial period, there was a slight (descriptive) decrease of 11 percentage points in responses from participants who agreed that they struggled to find the information or documents they needed to complete their job. Further, qualitative responses indicated that participants felt Copilot assisted with record keeping and recovering lost corporate knowledge in some areas.

Figure 15 Participants' self-report of whether they find it difficult to find information for their role



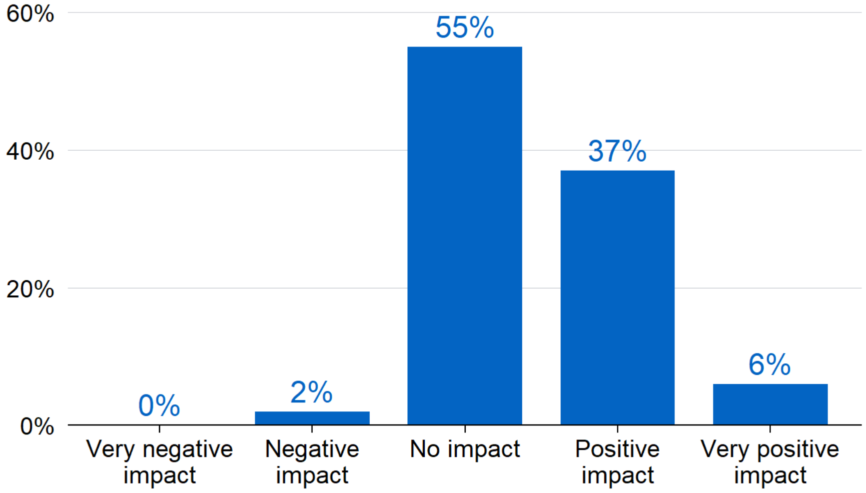
Source: Pre-trial survey: 'To what extent do you agree or disagree with the following statements? I struggle to find the information or documents I need to complete my job'; Post-trial survey: 'To what extent do you agree or disagree with the following statements? I struggle to find the information or documents I need to complete my job'
Note: N = 100; including only respondents who could be matched across pre- and post-trial surveys by their provided email address.

> *I was more diligent about saving things to SharePoint because Copilot only lets you reference documents there!*
>
> ***Trial participant,*** Post-trial survey

> *Super useful pointing to files that existed from last year as an example that you did not know existed.*
>
> ***Trial participant,*** Focus Group discussion

Similar to the findings reported in Figure 10, managers were more ambivalent about the positive impacts of Copilot on record management; 55 per cent of managers indicated that Copilot had no impact on their staff's knowledge management processes (Figure 16). These results are likely in part due to the lack of visibility of this work.

Figure 16 Manager ratings of the influence of Copilot on knowledge management processes



Source: Manager survey: ‘For the staff you manage who are participating in the Copilot trial, based on your experience, what average impact has Copilot had on: the team’s knowledge management processes? (i.e., generating meeting minutes and action items, or finding/summarising existing content?’
Note: N = 49.

This suggests that Copilot has some value in strengthening knowledge management through its ability to filter and find relevant information in Treasury’s systems, and to ensure that knowledge management protocols are followed. This is an important benefit given the frequency of staff movement across different areas of Treasury and the APS. However, this benefit may only extend to individuals, and not yet across entire teams.

5. Quality work outcomes

This section addresses the fourth evaluation question: *To what extent does Copilot support quality work outcomes in Treasury?*

Contrasting with the broadly positive impact of Copilot in supporting basic work processes, there was more ambiguity in the reported benefits of Copilot on work outcomes during the trial period. While some of the quantitative self-report data demonstrates some benefits of Copilot on work outcomes, the qualitative data collected from focus group discussions and the surveys provides a more nuanced picture. For example, qualitative data suggests that although Copilot can achieve work outcomes more quickly, the work outcomes are not necessarily better compared to human-generated work outputs.

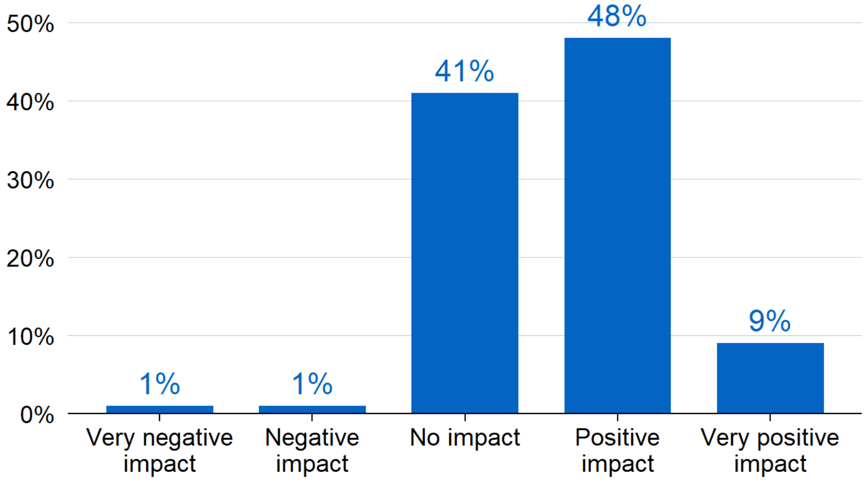
It is likely that the trial period was not long enough for these benefits to materialise. Just over half of post-trial survey respondents indicated that they believed Copilot had a positive impact on work outputs, but most of the remainder of participants reported no impact of Copilot on outputs. Most managers were also neutral about the average impact of Copilot on the quality of their team members' work. While there were some early case study examples of Copilot contributing to quality work outcomes, these were limited.

When considering Copilot's support of participant wellbeing and workload stress, there were some indicative positive findings, but these were not reported for most participants. Most trial participants indicated that Copilot had no impact on their stress levels or wellbeing, with just a few reporting some level of positive impact. While there were some descriptive increases in staff role satisfaction compared to the pre-trial survey, this change cannot be attributed to Copilot given the passage of time and the absence of a counterfactual.

There is some evidence that Copilot influenced work outcomes in the short-term

There were mixed views on the extent to which Copilot impacted the quality of work outcomes. The post-trial survey indicated that 57 per cent of respondents felt Copilot had a positive or very positive impact on work outputs. However, 41 per cent of respondents reported no impact of Copilot on work outputs. Evidence from the pulse surveys also showed similar results, with ratings of positive impact on work outputs hovering between 42 to 61 per cent during the trial.

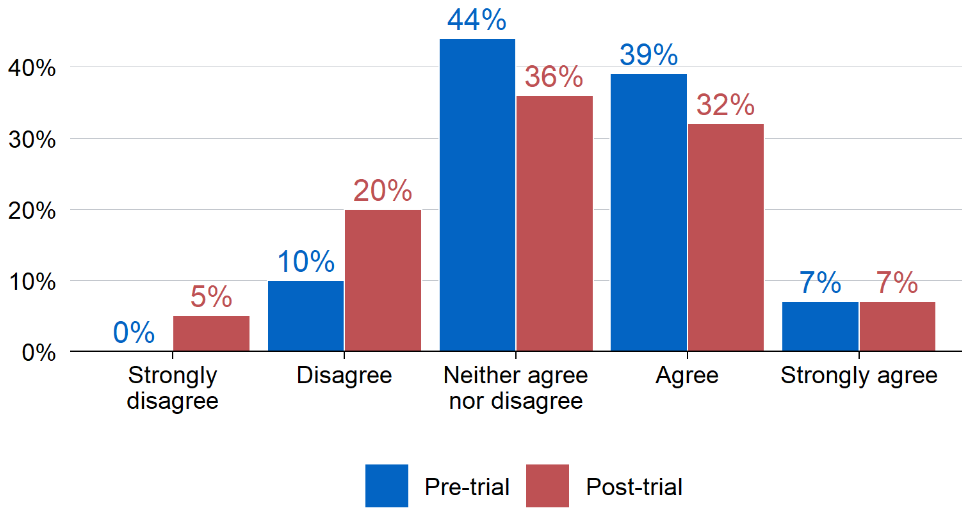
Figure 17 Participant ratings of the impact of Copilot on work outputs



Source: Post-trial survey: 'What impact has Copilot had on your work outputs to date? (e.g., work outputs could include a draft project plan, meeting minute summary, draft brief)'
Note: N = 148.

When considering the quality of work produced, 39 per cent of post-trial survey respondents agreed or strongly agreed that Copilot had improved the quality of their work. Thirty-six per cent were neutral and 25 per cent disagreed. Consistently, 65 per cent of managers reported that Copilot had no average impact on team members' work quality during the trial period.

Figure 18 Participant ratings of the impact of Copilot on work quality



Source: Pre-trial survey: 'To what extent do you agree or disagree with the following statements? I believe Copilot will: Improve the quality of my work'; Post-trial survey: 'To what extent do you agree or disagree with the following statements? I believe Copilot has: Improved the quality of my work'
Note: N = 100; including only respondents who could be matched across pre- and post-trial surveys by their provided email address.

Several factors may have contributed to limited progress in this area, including a lack of familiarity with the product and that the usage of the product did not reach maturity to contribute to work outputs. It is also possible that the lack of Copilot's influence on work outcomes is somewhat harder to capture (versus processes). This is because work processes (such as meeting minutes) tend to undergo fewer iterations, including reviews for accuracy. This hypothesis is borne out in the qualitative survey responses – some participants reported that Copilot outputs were inaccurate, and required more time to review and edit, which limited Copilot's reported benefits for work outcomes.

> *It has neither been revolutionary nor has it made a negative impact. The outputs are either too high level and largely correct, or too low level and incorrect.*
>
> ***Trial participant***, Pulse survey

> *… for some products it helps me produce an output faster, but for many tasks it does not help at all.*
>
> ***Trial participant***, Pulse survey

The examples from focus group discussions and case studies highlight areas where Copilot has supported quality work outcomes.

> *I asked it a policy question – given x, y, z – what's your opinion? It came out with a good response to the policy problem. You do need a good 15–20 min conversation to provide it with more context.*
>
> ***Trial participant***, Focus Group discussion

> *I used it as a research tool or research assistant – it has pulled out things that we might not have otherwise found.*
>
> ***Trial participant***, Focus Group discussion

When participants provided examples of ways Copilot contributed to work outcomes, many of these case studies described creative ways in which Copilot supported processes to strengthen the results of their work. These included using Copilot to overcome technical data challenges, support data analysis, write speeches, and summarise dense information. Some of these are highlighted in this section (see Appendix E for additional case studies). However, there were a limited number of case studies submitted, suggesting that there was only marginal tracking toward the medium-term outcome: 'Participants are using all relevant functions of the tool to benefit their work' (Table 1).

Case study: Supports strategic analysis of information

What: Reduce intensive manual handling of information, allowing more time for analysis.

Prompt: Based on the media summary table in this document, provide me a summary of each item with the following structure: [Website name] Headline in bold. Date. First sentence of article.

Value: Copilot does basic tasks of moving information into the format I want, at a faster speed and with less chance of a handling error.

It allows me to spend more time on the analysis and commentary at the top of the media summary, where I can provide strategic comments on whether the issues in the media summary are relevant to our work and may influence upcoming policy development. It removes process effort from the task and enables me to do more qualitative, subjective analysis work without increasing the overall time to produce the media summary.

Case study: Solving coding errors

What: Using Copilot chat function in Microsoft Teams to find solutions to coding errors.

Prompt: I am trying to make a chart in R with [XYZ feature] using the below code [insert code]. I am getting the following error message [insert error message]. Can you show me how to fix this?

Value: One participant used Copilot to tell them how to fix a specific error message in the coding program R. They said it saved significant time Googling solutions and searching through answers on online forums, which are usually not 100 per cent applicable to a unique problem. Copilot not only provided a solution to the error message but also highlighted why the solution works and identified best practices for setting up code and data.

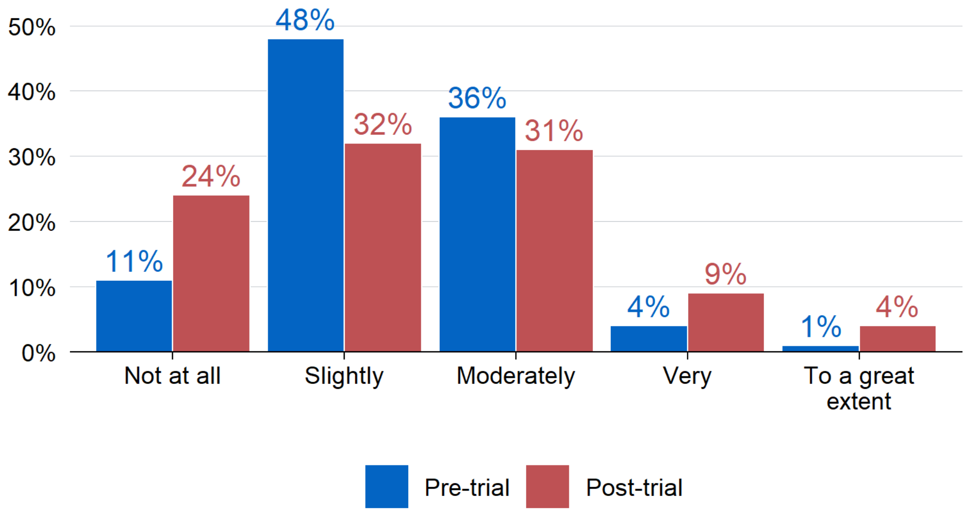
Copilot had small but potentially promising effects on employee's outcomes

When considering Copilot's potential benefits to employees in Treasury, the evaluation focused predominantly on levels of workload stress employees were experiencing, and participant's reported levels of work satisfaction. Critically, while there were some descriptive changes in trial participants' workload stress and role satisfaction, these changes might have occurred for reasons outside of access to Copilot, and/or normal fluctuations in sentiment toward work associated with deliverables and workload.

There is some encouraging evidence that the trial progressed towards the short-term outcome: 'Participants indicate an increase in work satisfaction using Copilot', but relatively limited evidence of progress toward the medium-term outcome: 'Participants indicate reduction in workload stress' (Table 1).

In relation to workload-related stress, 41 per cent of respondents in the pre-trial survey indicated that their workload was having at least a moderate impact on stress levels. Of note, following the Copilot trial, the distribution of self-reported workload-related stress flattened – descriptively, more respondents reported that their workload was not at all causing them stress (an increase of 13 percentage points from 11 per cent pre-trial to 24 per cent post-trial), but a higher number of post-trial respondents also indicated that their workload was very much or to a great extent causing them stress. These appear to be changes at the margins and cannot be directly attributed to the Copilot trial.

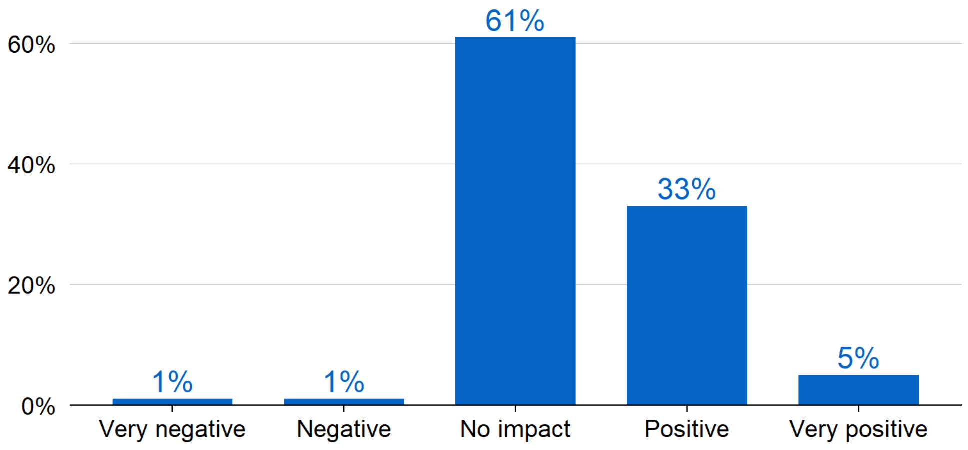
Figure 19 Participant self-reported workload-related stress



Source: Pre-trial survey: ‘To what extent: Do you feel your current workload is causing you stress?’; Post-trial survey: ‘To what extent do you feel your current workload is causing you stress?’
Note: N = 100; including only respondents who could be matched across pre- and post-trial surveys by their provided email address.

Respondents were also asked about Copilot’s specific impact on workload-related stress in the post-trial survey. While 37 per cent did report some level of positive impact of Copilot on their workload stress post-trial, most respondents (61 per cent) indicated that Copilot had no impact on their stress levels.

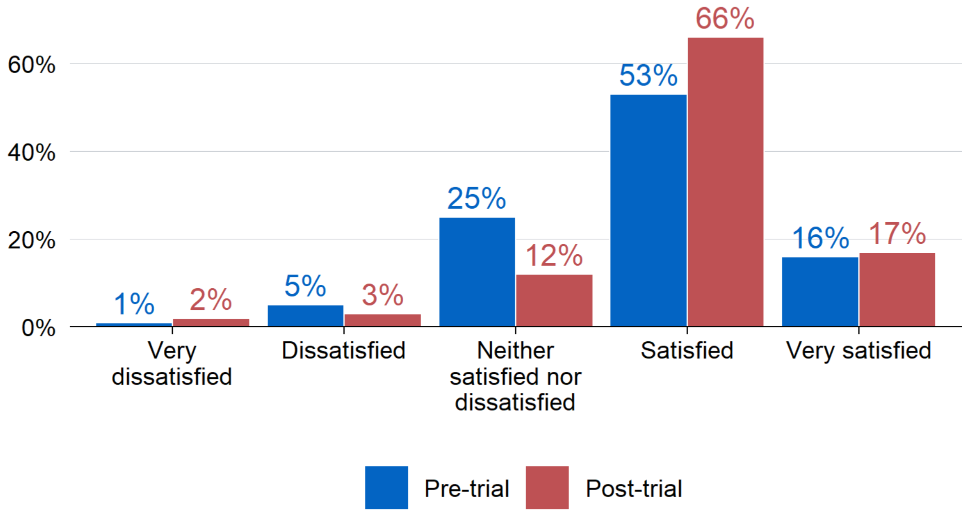
Figure 20 Participant reported impact of Copilot on workload-related stress



Source: Post-trial survey: 'What impact has Copilot had on your workload-related stress?'
Note: N = 133.

There were some descriptive increases in trial survey respondents' self-reported levels of satisfaction with their current role from the pre-trial to post-trial survey periods. Specifically, the number of staff who indicated that they were either satisfied or very satisfied with their role rose from 69 per cent in the pre-trial period to 83 per cent in the post-trial period, a rise of 14 percentage points. These results are broadly consistent with results from the 2024 APS Census, which indicated overall high levels of satisfaction in Treasury roles (Australian Public Service Commission, 2024). These high levels of satisfaction at baseline may have led to a ceiling effect, whereby access to Copilot was only able to shift satisfaction with current role slightly higher.

Figure 21 Participants' self-reported role satisfaction



Source: Pre-trial survey: 'How satisfied are you with your current role?'; Post-trial survey: 'How satisfied are you with your current role?'
Note: N = 100; including only respondents who could be matched across pre- and post-trial surveys by their provided email address.

6. Unintended consequences

This section addresses the fifth and final evaluation question: *Were there any unintended outcomes of using Copilot (positive/negative)?*

Copilot supported staff in ways that were not intended or expected prior to implementation including accessibility and inclusion within teams; staff confidence in their work; and building cross-Treasury networks.

Copilot contributed to accessibility and inclusion within teams

Focus group discussions highlighted that Copilot improved accessibility and inclusion for those new to the public service, working part-time, identifying as neurodivergent and participants who may be experiencing mental health challenges. Some respondents described it as levelling the playing field for those who encounter challenges navigating workplace norms, organisational systems or new working approaches. For example, one respondent found the meeting summary function of Copilot invaluable for identifying the main points in long meetings when they lost focus. Another participant indicated that Copilot supported accessibility if they needed to take unexpected time off work due to mental health challenges. Copilot reduced the time required for their manager to update them on missed work. It also reduced the participant's internalised stigma around taking time off.

> *It's an equaliser, supports people who have time management or dyslexia with spelling to improve access for those staff.*
>
> ***Trial participant***, Focus Group discussion

> *It is …useful for meeting summaries when it is hard to focus or when you lose focus.*
>
> ***Trial participant***, Focus Group discussion

Copilot also provided opportunities for junior staff to free up time for more strategic or complex work. Copilot supported junior staff to undertake basic administrative duties more efficiently. This allowed more time to engage in professional development opportunities and undertake work of more substance such as drafting policy briefs or data analysis. This highlights that Copilot has the potential to create opportunities for junior staff to learn and develop at a faster pace than has otherwise been available to them.

Some trial participants also indicated that Copilot reduced the impact of procrastination and anxiety they experienced at work by supporting them to undertake more difficult tasks. Copilot gave them the ability to brainstorm ideas and crosscheck words or approaches with a professional product.

> *From an Access and Inclusion perspective, CoPilot [sic] has provided me reduced anxiety and a private filter to check my wording when presenting clear instructions to staff. This concept of a private second opinion for any user with anxiety or neuroatypicality makes it critically important to consider long term – irrespective of the current concerns with the version's handicaps.*
>
> ***Trial participant,*** Post-trial survey

> *I did not expect it to assist with ADHD/autism related procrastination. It has been very helpful in this respect.*
>
> ***Trial participant,*** Post-trial survey

Participants also highlighted that Copilot supports part-time staff or those that have a full calendar. Participants accessed Copilot's meeting summaries and recordings to catch up on work efficiently.

> *The automated meeting minutes are also a great gain – especially for areas where you have a lot of meeting clashes, have part-timers who miss meeting or just need more time in your calendar. I'd love it to be the default that something like copilot [sic] generates these for every meeting.*
>
> ***Trial participant,*** Post-trial survey

Case study: Accessing missed information while on parental leave

What: Using Copilot to summarise key meetings and information that was missed during parental leave.

Prompt: Summarise what happened at X meeting on X June 2024. Identify the key points of the meeting, and any notable points of discussion, agreement, and disagreement.

Value:

- Emotionally – helped me manage emotions.
- Socially – unburdened team members of the responsibility of catching me up to the content of an important meeting I missed while on parental leave.
- Functional – saved my team the time taken explaining what happened in the meeting, and possibly summarised the meeting more effectively, in terms of time and memory of key points.

In terms of inclusiveness of an organisation, qualitatively, this improves what Treasury can offer.

These experiences demonstrate the value Copilot brings to supporting accessibility and inclusion in the workplace. It also shows generative AI products have potential to improve professional opportunities for diverse staff.

Copilot increased some participants’ confidence in their work

Some participants indicated Copilot improved their confidence in their work and outputs. These staff reported using Copilot to check their work (for example, treating Copilot as a friendly colleague), or as a virtual administrative assistant. This was largely reported by staff who were early in their career and/or new to Treasury.

> *… Copilot definitely helped me improve my writing and communication skills which has made me feel more confident in my role.*
>
> ***Trial participant,*** Post-trial survey

> *I feel good about how quickly I can get stuff out but there is a small level of guilt because I am doing tasks in a short timeframe, have I missed something? Is the quality good enough?*
>
> ***Trial participant,*** Focus Group discussion

Copilot introduced participants to work and people across different areas

The trial Champions, trial Teams chat, and other trial activities had an unintended benefit of introducing trial participants to others in the department. This networking benefit has unintentionally led to connections and knowledge being shared across Treasury, which may benefit future work.

> *The Copilot trial community chat gave me unexpected insights into the work done by areas I don’t usually interact with.*
>
> ***Trial participant,*** Post-trial survey

7. Summary and recommendations

Summary

The results of this trial demonstrate that generative AI can be deployed within Treasury, but there are some limitations. These limitations stem from the product being relatively new and its features are still being refined, and that the product is subject to security restrictions due to Treasury's protected environment.

High expectations for Copilot's benefits and performance are likely to have led to reductions in usage over time as user's expectations were not met. Regardless, generative AI proved to be useful for some purposes, with users finding it appropriate and beneficial for basic administrative tasks. However, the current iteration of Copilot is not sophisticated enough to support complex work tasks without substantial input on both the prompt and the output to ensure accuracy and suitability. As generative AI products continue to evolve, it is plausible that they could become better at handling more complex tasks in the future.

Perceptions of generative AI were broadly positive. However, many participants reported feeling disappointed by Copilot's functionality, particularly in comparison to other generative AI products. Currently, Copilot is the only generative AI integrated within the Microsoft suite. As a result, options for the rollout of generative AI within Treasury that can operate within the primary applications and tools used by staff are limited.

The current trend of adoption of generative AI indicates it is likely to become the norm to use generative AI for many basic tasks in the future. The critical question in this context is whether Treasury should consider adopting generative AI in its current state or wait until the product is more advanced. If adopted in its current state, staff can build their capability and experience the technology as it continues to be improved and iterated. Alternatively, Treasury could wait for a more advanced product, and staff will receive access to generative AI products as late adopters.

There are immediate costs to adopting this technology associated with the licences required alongside onboarding and training costs. ACE estimates that a staff member on the current APS6 salary would need to redirect approximately 13 minutes of time per week to higher value tasks for the current licence cost to be offset. Although the data collected during this trial did not quantify time savings, the results of this trial suggest that the productivity benefits and time savings associated with Copilot are likely to offset the licence costs.

Recommendations

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.** Evidence from this trial suggests that Copilot has specific benefits for supporting process improvement and undertaking basic administrative tasks. As a result, providing specific use cases that outline these benefits will enable staff to determine whether the product will support them in their work. Managers should also review staff role descriptions to determine whether Copilot's use cases would benefit specific team members. Priority for initial licence distribution should be given to a targeted group of staff who are most likely to benefit from these use cases.

Further, communications on the use cases should be specific about what the product will provide to participants to avoid artificially high expectations. This will support participants to self-nominate to receive generative AI products based on the potential for benefits to their work. It will also partially act to prevent disengagement if the product does not immediately meet expectations, supporting users to navigate initial stages of the product rollout towards maturity of implementation within Treasury. Generative AI products could also be provided to those with an accessibility or inclusion use case as a priority. This is because Copilot has been an important contributing factor in supporting staff with different types of access and inclusion barriers to undertake their work.

2. **Any future rollout of new generative AI products should be based on a phased approach.** Future rollouts of any generative AI products should start with a small group of staff for who Copilot is most likely to be valuable. The rollout should then work to reach more staff over time, until the product itself and users' interactions with the product mature. This will require considered investment and a sustained effort to ensure Treasury is enhancing its generative AI capability in line with developments in the technology.

3. **Any future rollout of generative AI products should include an assessment of the appropriate level of investment in education and training.** Providing formal resources and supports to educate and train staff will facilitate use of generative AI in staff work and make the most of the products. Relying predominantly on online Microsoft Teams forums and self-nominated Champions did not provide the technical support participants needed.

   Given the emerging nature of generative AI, training and engagement should be continued over time through more dynamic capability building mechanisms including buddies and communities of practice, in addition to structured training. Any implementation of future generative AI products should include a training plan that accounts for the cost of training as well as the associated time commitment expected from participants (both to participate in training and to experiment with the technology).

4. **In any future rollout of generative AI products, develop guidelines to support the transparent use of generative AI.** Guidelines that set expectations about the use and disclosure of generative AI would address a key concern among staff and reduce the risk of a loss of trust among senior executives, ministers, other stakeholders or the public. These guidelines should continue to emphasise the importance of owning the outputs from generative AI: that is, it is crucial that staff can explain and own any output they use generative AI to create, and so they should critically review all outputs with this lens. Future guidelines should be developed in consultation with relevant parties and consistent with legislative and other APS requirements, and communicated to all staff, managers and executives.

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.** This could be undertaken via regular reviews of work outputs within units and should include both subjective self-reported data including case studies, and objective reports via managers or other senior executives. The benefits of generative AI could also be captured via reporting such as document readability scores, or other appropriate quantitative metrics (however, note that the 'productivity' scores generated by Microsoft on individual users do not provide relevant information in the Treasury context).

Once the implementation of the product has reached maturity, the impact of generative AI products could also be tested in an experimental setting, leveraging the design of similar trials being conducted in other contexts. This would contribute to the nascent evidence base of the benefits of generative AI in the workplace, while providing further evidence on the impact and appropriateness of generative AI within Treasury.

6. **Staff outcomes, including staff wellbeing, job satisfaction, and workload-related stress should be considered as an important secondary outcome of any generative AI product implementation.** While enhancements in staff satisfaction, wellbeing and experiences of stress may not be the primary aim of productivity-enhancing platforms, they are nonetheless an important secondary benefit. This is because staff satisfaction is commonly accepted as a critical predictor of high-quality work outputs and retention, with retention itself also a positive outcome for Treasury. So, while any future implementation of generative AI products should consider work outcomes as primary, these should not be the sole area of consideration for monitoring and evaluation.

   A focus on staff-related outcomes as well as unintended benefits should continue throughout the implementation, but this should be done in a way that balances the need for good monitoring with the burden of data collection.

7. **Conduct periodic assessments of whether emerging generative AI products may be better suited to Treasury's security requirements and existing IT infrastructure.** While security of protected government data and advice is of upmost importance, ideally the core functions of a generative AI product should work alongside security requirements. It is not clear whether products are likely to evolve over time to meet Treasury's strict security needs, or whether Copilot itself will continue to evolve to incorporate external information into its outputs without feeding the algorithm with internal Treasury data.

   Future research into whether open-source generative AI products are fit for purpose (that is, how they comply with Treasury security requirements) may provide additional avenues to integrate other generative AI products into Treasury's work processes.

Appendix A: Copilot Trial Program Logic

This program logic was co-developed with ACE, the Copilot trial Project team, and Working Group.

Table 4 Copilot Program logic

Inputs	Activities	Outputs	Short term outcomes	Medium term outcomes	Long term outcomes
• $$ Funding • 235 Licenses • AI Working group • Community of practice • AI Steering Committee	• Onboarding of participants (n=204) • Participants undertake mandatory training module • Deliver education and CoP sessions for each phased rollout • Managing a learning/feedback platform • Onboard champions • Development of an internal AI policy	• Participants onboarded • Participants attend mandatory training module • Education and CoP sessions delivered • Participants are sharing learning/challenges to inform rollout • Champions recruited and participants engaged • AI policy developed	• Participants have competence to use and confidence to experiment with the Copilot tool • Participants are using Copilot for the identified use cases • Participants indicate an increase in work satisfaction using Copilot • Participants indicate an increase in process improvement using Copilot	• Participants are using all relevant functions of the tool to benefit their work • Copilot is supporting improved knowledge management • Participants indicate reduction in workload stress • Copilot is improving workflows and new approaches to problems	• Copilot is contributing to increases in: • Productivity • Access to timely, quality information (information management) • Staff wellbeing (people) • Good decision-making (risk management) • Innovation

Appendix B: Evaluation Methods

The ACE employed a mixed-methods approach in evaluating the Treasury trial of Copilot. Given the trial timeframes, the ACE used a mixed-methods approach to collect data to determine progress towards the short-term and medium-term outcomes as outlined in the program logic. The ACE considered the long-term outcomes unlikely to be achieved in the trial period, so were considered out of scope for the purposes of this evaluation.

This section outlines the data collection and analytical approach for each element of data, alongside the approach to the triangulation of the information gathered. All data collection materials are shown in Appendix C.

Pre-trial and post-trial surveys of trial participants

A pre-trial survey was undertaken at baseline and a post-trial survey after the trial ended. The surveys were designed to capture how Copilot influenced the daily work of licence holders, with a focus on Copilot usage, productivity, innovation, process improvement and staff satisfaction.

The pre-trial survey was open for 10 business days between the 20 May 2024 (the first day of the Copilot trial) to 31 May 2024. During this period, 153 respondents completed the survey, from a total of 232 trial participants. This is a response rate of 66 per cent.

The post-trial survey opened on 23 August 2024, and closed on 6 September 2024, meaning it was open for a total of 11 business days. During this period, 164 respondents completed the survey, which is a response rate of 71 per cent.

Data collected via the pre-trial and post-trial surveys were downloaded from the online survey platform (Qualtrics). All analysis was completed via R, with basic descriptive statistics generated for all relevant survey questions and all qualitative data analysed thematically.

Pre-trial and post-trial survey data was matched via participant-provided email addresses where a trial participant completed both surveys; there were 100 matches, out of a total of 150 complete and verifiable post-trial survey responses. This is a match rate of 66 per cent. Due to the limited sample of pre and post survey matches, inferential statistics on changes across the trial period were not completed; as a result, all descriptions of changes from the pre-trial to post-trial period are descriptive only.

Pulse survey of trial participants

Trial participants were surveyed every fortnight during the trial to get a pulse check on:

- Copilot usage
- Perceptions of Copilot’s impact on work processes and outputs
- Any issues experienced throughout the trial.

Pulse surveys were analysed descriptively and were reported on a regular basis to the Trial Project Board.

Each pulse survey opened on a Friday, with all trial participants notified of the survey via email. The surveys were open for 3 to4 business days. Only some questions were mandatory, meaning the sample size differs for each question.

Table 5 Pulse survey

Pulse survey	Date closed	Sample size	Response rate
1	19/06/2024	68	29.6%
2	03/07/2024	131	56.5%
3	17/07/2024	79	33.6%
4	01/08/2024	61	26%
5	15/08/2024	68	29.6%

The pulse surveys suffered from a limited sample size. As a consequence, the results from the pulse surveys may not be representative of the experiences of non-respondents and are not discussed in detail in the trial report.

Post-trial survey of managers of trial participants

A survey of the managers of trial participants was used to understand any potential effects of access to Copilot on work quality, efficiency of work outcomes, and (perceived) staff satisfaction. The project team identified 124 managers that trial participants directly reported to, most of who were EL2 staff at Treasury.

Managers were surveyed to provide some external and more objective ratings of the potential impact of Copilot on work outcomes. However, managers were surveyed about all staff they supervise who were participating in the trial, which may limit the ability to detect individual-level impacts for staff. A total of 49 managers of trial participants responded to the survey between 3 September 2024 and 12 September 2024, a total of 8 business days. This equates to a response rate of 40 per cent.

Focus group discussions

ACE held focus group discussions with trial participants and Champions to better understand the benefits and challenges of using Copilot across teams.

The evaluation team facilitated all focus group discussions, with either a member of the evaluation team or the Copilot project team acting as notetaker. Focus group questions are shown in Appendix C, and focus groups were facilitated in a semi-structured manner, with the facilitator asking probing questions if further detail or clarification was warranted.

Table 6 Focus group discussions sample size and composition

Date	Participants	Sample size
02/07/2024	Copilot champions	6
02/07/2024	Copilot participants	8
19/08/2024	Copilot champions	4
19/08/2024	Copilot participants	7

ACE analysed focus group discussion transcripts and notes thematically, with inputs coded, and themes drawn from the codes.

Case studies

The evaluation team provided all trial participants with a PowerPoint slide template and asked to provide examples of ways they had used Copilot throughout the trial. ACE collected a selection of case studies on participant use of Copilot based on feedback through the focus group discussions and Copilot Community of Practice.

Copilot trial participant issues log

The evaluation team distributed a central issues log to all licence holders via email and the Microsoft Teams channel. Trial participants were asked to log any challenges or issues with Copilot to identify unintended outcomes. This process supported the implementation team to mitigate risks throughout the trial.

The evaluation team reviewed the issues log following the trial, and incorporated the reported issues into the overall evaluation findings where relevant.

Appendix C: Data collection tool

Pre-trial survey

Informed consent text (used for every survey)

Thank you for participating in this survey. The aim of this survey is to understand the impact that having access to Copilot has on your work satisfaction, work outputs, and efficiency. We are interested in understanding your perspectives about Copilot and how you use it at Treasury, and there are no right or wrong answers.

- This survey is being conducted by the Australian Centre for Evaluation, and will be used to support Treasury decision-making on investments in future generative AI products, including Copilot.
- This survey should take approximately 10 minutes to complete.
- Your participation in this research is voluntary. You have the right to withdraw from the research at any time up until the point your data is linked for analysis, and therefore de-identified. Please contact [contact] if you would like to withdraw from the research.
- All of the information that you provide will be treated as confidential and will only be used for the purposes of the Copilot trial evaluation. Any data that you provide us will be analysed and reported at an aggregate level. If we include any quotes from you, they will be used in a way that you cannot be identified from what you said.
- We will only ask for your Treasury email address to link your responses to other Copilot evaluation surveys. The email address will only be used for this purpose, and will be stripped from the data and replaced with a unique identifier once the linkage is completed.
- The data collected from you, once deidentified, will be stored in a SharePoint online folder with access restricted to researchers directly involved in the analysis and quality assurance.
- You can contact [contact] if you have any questions about the research.
- If you would like to raise a complaint about any aspect of the research, please contact [contact] at [email].
- If you consent to participate in this research, please click the <Next> button below. If you do not consent to participate in this research, please close this browser window.

Table 7 Pre-trial survey

Question		Response options	Notes
Participant information			
1	What is your Treasury email?	[Email text entry], mandatory question	
2	In an average week, how many hours do you spend … Searching for information required for a task Summarising existing information for various purposes (email updates, talking points, briefs etc.) Taking and preparing meeting minutes Preparing first drafts of a document Undertaking preliminary data analysis Preparing presentation slides Other [Free text]	[For each statement] 0 hours 1-2 hours 3-5 hours 6-9 hours 10-15 hours 16+ hours	
3	In your opinion, how much of your current weekly workload could be performed more effectively with the help of automation? i.e., taking meeting minutes, summarising weekly reports or documents, creating a slide deck or automation of data outputs	Little to none (0-25%) Some of my tasks (25-50%) Most of my tasks (50-75%) Nearly all of my tasks (75-100%)	
4	To what extent: Do you feel your current workload is causing you stress? Do you think Copilot could help you manage workload-related stress?	Not at all Slightly Moderately Very To a great extent	

Table 7 Pre-trial survey (continued)

Question		Response options	Notes
Current believes about work			
5	To what extent do you agree or disagree with the following statements? I struggle to find the information or documents I need to complete my job I don't have as much focus time as I would like I often feel rushed and don't feel I have put forward my best work I struggle to stay organised	[For each statement] Strongly disagree Disagree Neither agree nor disagree Agree Strongly agree	
6	How satisfied are you with your current role?	Very dissatisfied Dissatisfied Neither Satisfied Very satisfied	
Knowledge of and competence with AI tools and Copilot			
7	How much experience do you have with using generative AI tools: In the workplace? In your personal life?	No experience (Never used AI tools before) Minimal experience (Use AI tools less than once a month) Some experience (Used AI tools between once a month, and once a fortnight) A lot of experience (Used AI tools between once a fortnight and once a week) Expert (Used AI tools for uses other than basic chat or summaries)	
8	How competent do you currently feel using Copilot in your role?	[0-10 scale, with the following anchors] 0 – Not competent at all 1 2 – Slightly competent 3 4 – Somewhat competent 5 6 – Fairly competent 7 8 – Very competent 9 10 – Highly competent	Additionally captured in the Pulse survey and the post-trial survey

Table 7 Pre-trial survey (continued)

Question		Response options	Notes
Anticipated impact of/beliefs about Copilot			
9	Which of the following best describes your sentiment about using Copilot at Treasury:	Very negative Slightly negative Neutral Slightly positive Very positive	
10	To what extent do you agree or disagree with the following statements? I believe Copilot will … Improve the speed at which I complete tasks Improve the quality of my work Reduce the amount of time I spend on low-value or low-priority tasks Be a net positive on my work	[For each statement] Strongly disagree Disagree Neither agree nor disagree Agree Strongly agree	
11	What features of Copilot are you most looking forward to using?	[Free-text]	
12	What concerns do you have about using Copilot?	[Free-text]	
Demographics			
13	Gender: How do you identify?	Man Woman Non-binary Prefer to self-describe another way [Free text] Prefer not to say	
14	What is your current job level?	Contractor APS3 and below APS4 APS5 APS6 EL1 EL2 SESB1 and above	
15	Which Group of Treasury are you currently employed in?	Fiscal Group International Foreign Investment Group Macroeconomic Group Markets Group Revenue Group Small Business Housing Corporate and Law Group	

Table 8 Post-trial survey

Question		Response options	Notes – Link to indicators and outcomes
Participant information			
1	What is your Treasury email?	[Email text entry], mandatory question	
Current behaviours and workload			
2	To what extent has Copilot improved processes to: • prepare first drafts of a document • prepare meeting minutes • search for information needed to do a task • summarise existing information for various purposes	[For each statement] Not at all Slightly Moderately Very much To a great extent	Capturing the following medium-term indicators: • % participants report improved practices for: • Preparing first drafts of a document • Preparing meeting minutes • Searching for information needed to do a task • Summarising existing information for various purposes
3	In your opinion, how much of your current weekly workload has been performed more effectively with the help of Copilot's automation? i.e., taking meeting minutes, summarising weekly reports or documents, creating a slide deck or automation of data outputs	Little to none (0-25%) Some of my tasks (25-50%) Most of my tasks (50-75%) Nearly all of my tasks (75-100%)	Capturing the medium-term indicator '% of participants indicate increased productivity related to Copilot'
4	To what extent: Do you feel your current workload is causing you stress?	Not at all Slightly Moderately Very To a great extent	Capturing data on the medium-term indicator '% participants who report reduced workload-related stress'
5	What impact has Copilot had on your workload-related stress?	Very negative impact Negative impact No impact Positive impact Very positive impact	Capturing data on the medium-term indicator '% participants who report reduced workload-related stress'

Table 8 Post-trial survey (continued)

Question		Response options	Notes – Link to indicators and outcomes
Current beliefs about work			
6	To what extent do you agree or disagree with the following statements? • I struggle to find the information or documents I need to complete my job • I don't have as much focus time as I would like • I often feel rushed and don't feel I have put forward my best work • I struggle to stay organised	[For each statement] Strongly disagree Disagree Neither agree nor disagree Agree Strongly agree	
Knowledge of and competence with AI tools and Copilot			
7	How competent do you currently feel using Copilot in your role?	[0-10 scale, with the following anchors] 0 – Not competent at all 1 2 – Slightly competent 3 4 – Somewhat competent 5 6 – Fairly competent 7 8 – Very competent 9 10 – Highly competent	Capturing the short-term indicator '% of participants indicate they feel competent using the tool'
8	How satisfied are you with your current role?	Very dissatisfied Dissatisfied Neither Satisfied Very satisfied	Capturing the short-term indicator '% of participants indicate an increase in work satisfaction'

Table 8 Post-trial survey (continued)

Question		Response options	Notes – Link to indicators and outcomes
Anticipated impact of/beliefs about Copilot			
9	Which of the following best describes your sentiment about using Copilot at Treasury:	Very negative Slightly negative Neutral Slightly positive Very positive	
10	To what extent do you agree or disagree with the following statements? Using Copilot has … Improved the speed at which I complete tasks Improved the quality of my work Reduced the amount of time I spend on low-value or low-priority tasks Been a net positive on my work	[For each statement] Strongly disagree Disagree Neither agree nor disagree Agree Strongly agree	Capturing the following medium-term indicators: • % participants report spending less time on low priority task (defined by the individual)
Use of Copilot			
11	How frequently did you use Copilot for work-related tasks throughout the Copilot trial period?	Never Less than once a week 2-3 times a week 4-5 times a week More than once a day	Capturing the short-term indicator '% of participants using Copilot'
12	How frequently did you use Copilot for the following types of tasks? Generating structured content (i.e., a first pass project plan, content for briefings, correspondence, or sourcing factual material) Knowledge management (i.e., generating meeting minutes and action items, or summarising existing content) Undertaking process tasks (i.e., drafting emails, summarising calendar appointments, or analysing email traffic) Synthesising and prioritising information (i.e., preparing first-pass data analysis, synthesising stakeholder feedback, or summarising key themes from data)	Never Less than once a week 2-3 times a week 4-5 times a week More than once a day	Capturing the following short-term indicators: • % of participants using Copilot for generating structured content • % of participants using Copilot for knowledge management • % of participants using Copilot for process tasks • % of participants using Copilot for synthesising and prioritising information

Table 8 Post-trial survey (continued)

Question		Response options	Notes – Link to indicators and outcomes
Use of Copilot			
13	What impact has Copilot had on your work processes to date? Why did you answer this way?	Very negative impact Negative impact No impact Positive impact Very positive impact	Capturing the short-term indicator '% of participants indicate an increase in process improvement'
14	What impact has Copilot had on your work outputs to date? Why did you answer this way?	Very negative impact Negative impact No impact Positive impact Very positive impact	Capturing the medium-term indicator '% of participants indicate benefits to the quality of their work outputs'
15	Did you experience any unintended outcomes of using Copilot (positive/negative)? What were these?	[Free-text]	Link to Objective 4 and KEQ 5
16	Would you like to add anything else about your experience using Copilot in your role at Treasury?	[Free-text]	

Table 9 Pulse survey

The evaluation team administered this survey every fortnight during the trial period. The intent was to capture data on Copilot usage and satisfaction early and frequently, to enable evidence-based decision-making on Treasury's procurement or use of AI products.

Question		Response options	Notes
1	What is your Treasury email?	[Email text entry]	This will be used to ensure data can be linked across data sources.
2	How frequently have you used Copilot for work-related tasks in the last fortnight?	Never Less than once a week 2-3 times a week 4-5 times a week More than once a day	
3	How satisfied are you with your current role?	Very dissatisfied Dissatisfied Neither Satisfied Very satisfied	This will also be captured in the post-trial survey, to collect data on the short-term indicator related to work satisfaction
4	How competent do you currently feel using Copilot in your role?	[0-10 scale, with the following anchors] 0 – Not competent at all 1 2 – Slightly competent 3 4 – Somewhat competent 5 6 – Fairly competent 7 8 – Very competent 9 10 – Highly competent	This will also be captured in the post-trial survey, to collect data on the short-term indicator related to competence in using Copilot

Table 9 Pulse survey (continued)

Question		Response options	Notes
5	What impact has Copilot had on your work processes to date?	Very negative impact Negative impact No impact Positive impact Very positive impact	This will also be captured in the post-trial survey to collect data on process improvements
6	What impact has Copilot had on your work outputs to date?	Very negative impact Negative impact No impact Positive impact Very positive impact	This will also be captured in the post-trial survey to collect data on improved work quality
7	Why did you answer this way?	Free-text	
8	Have you experienced any issues with using Copilot? If yes, which issues have you experienced?	Yes No Issues accessing Copilot Issues with IT Problems finding the correct prompt Irrelevant or unhelpful responses from Copilot Other (including description)	

Manager survey

The evaluation team opted to keep this survey as short as possible to reduce demands on managers, who already have significant workloads. As a result, the evaluation team asked directly about many of the relevant outcomes, rather than assessing them in multiple ways. While this is not ideal, the intent was to reduce demands on managers while still collecting valuable information about any potential impacts of the Copilot trial on staff.

Table 10 Manager survey

Question		Response options	Notes
Trial participant work experiences and outputs			
1	For the staff you manage who are participating in the Copilot trial, based on your experience, what average impact has Copilot had on … • The quality of your team members' work? • The timeliness of your team members' work? • The efficiency of your team members' work? • The team members' work processes (i.e., how they go about creating outputs)? • The team members' knowledge management processes? (i.e., generating meeting minutes and action items, or finding/summarising existing content)	Very negative impact Negative impact No impact Positive impact Very positive impact	Capturing the medium-term indicators: • % of managers indicate participants are using Copilot to produce quality work • % of managers indicate participants are using Copilot to produce work efficiently
2	From your perspective, to what extent are your team members that have participated in the Copilot trial (on average) satisfied with their current roles?	Very dissatisfied Dissatisfied Neither Satisfied Very satisfied	Capturing the medium-term indicator '% of managers indicate an increase in participants' work satisfaction'
3	Do you have anything further to add about you or your team member's experiences of using Copilot in their roles at Treasury?	[Free-text]	

Table 10 Manager survey (continued)

Question		Response options	Notes
Participant information			
4	What is your Treasury email?	[Email text entry]	
5	What Branch/Unit do you work in?	[Drop down listing all options: Group/Division/Branch]	
6	How many staff do you directly manage? How many staff who you manage participated in the Copilot trial?	[Drop-down, with options between 0-15+, for both questions]	

Appendix D: Copilot trial use cases

User goal: Generating structured content

Examples

- Developing a detailed project plan.
- Generating first pass content for briefings, correspondence, job advertisements, agenda items, or other documentation.
- Sourcing relevant factual material for input into documentation.
- Undertaking simple proof-reading and style checks.
- Rewriting drafted content to include more plain English, improve writing style, or alter the structure of the content.
- Developing a template.

Process

Arseny is developing a detailed project plan in Word. He wants to use Copilot for Microsoft 365 to create a baseline plan he can improve on.

Arseny can generate a tailored project plan in Word using Copilot for Microsoft 365. His agency has security assessed Copilot for use with information holdings at the PROTECTED level. Arseny is therefore able to input specific project details that might be sensitive or classified, including the project and agency names, names of systems/software he wants to use and high-level project requirements. Because Copilot can access files that Arseny has access to, it can generate a baseline plan that is relevant to his project.

User goal: Supporting knowledge management

Examples

- Recording and summarising meeting minutes to distribute to attendees and apologies.
- Generating meeting action items based on meeting discussions.
- Finding documentation that has not been updated in a specific period, to ensure its currency.
- Summarising guidance materials, information, or other content.

Process

Ben will be discussing a sensitive issue on a Microsoft Teams video call. He wants to use Copilot for Microsoft 365 to capture notes and summarise the meeting, instead of doing so manually, to help him attend to all details during the call.

Like any recording of a call, Ben should first obtain consent from his team – and let them know he will use Copilot to summarise it. Ben should also consider the security classification of the information they are likely to discuss and whether it is within the classification his agency has approved for use with Copilot.

After the meeting, Ben must review and edit the notes generated by Copilot for clarity and accuracy to the conversation, and ensure they are appropriately classified. Once edited and verified, Ben should share the final document with his team to maintain transparency, allow for their own review or corrections, and maintain the integrity of the recorded information.

User goal: Managing personal or process tasks

Examples

- Summarising calendar appointments or meetings in a day.
- Collating relevant information for a meeting based on email contents.
- Analysing or summarising email content, traffic, or other themes or trends.
- Drafting routine or simple emails.

Process

Ash took an unexpected day off work, and wants to summarise the most important emails that require immediate action before commencing their next day of work.

Ash uses Copilot for Microsoft 365 to summarise all of the time-sensitive unread emails in their inbox, and requests that Copilot list any key actions required of Ash. Based on this, Ash can start attending to the most pressing items.

Over the course of the day, Ash reviews all emails to ensure that no urgent information or actions were missed, and information provided by Copilot was accurate and consistent.

User goal: Synthesising and prioritising information

Examples

- Preparing first-pass data analysis and visualisations of structured data.
- Generating code snippets or functions.
- Summarising key themes within structured data (such as from the Australian Bureau of Statistics' (ABS) data releases).
- Synthesising stakeholder feedback from consultations or other data sources.

Process

Kirra is analysing a complex dataset in Excel. It contains columns with labels that might not be self-explanatory. She is considering using Copilot for Microsoft 365 to explore the data and generate insights.

Kirra can generate insights from the dataset using Copilot for Microsoft 365. She should carefully check the results to ensure generated insights are accurate, correctly interpreted and not ‘made up’ or ‘hallucinations’. Kirra should cross-reference generated insights with her knowledge of the dataset and subject matter, and seek to validate them with expert sources or subject experts.

If the dataset or analysis exercise relates to individuals or groups of people, Kirra should conduct research and seek advice to ensure they both are factually accurate and reflect the perspectives of those in question. These groups may include Aboriginal and Torres Strait Islander peoples, people with disability, LGBTIQA+ communities and multicultural communities.

Appendix E: Further Copilot case studies

This section includes a summary of all case studies submitted by participants for the purposes of the evaluation.

Case study: Solving coding errors

What: Using Copilot chat function in Microsoft Teams to find solutions to coding errors.

Prompt: I am trying to make a chart in R with [XYZ feature] using the below code [insert code]. I am getting the following error message [insert error message]. Can you show me how to fix this?

Value: One participant used Copilot to tell them how to fix a specific error message in the coding program R. They said it saved significant time Googling solutions and searching through answers on online forums, which are usually not 100 per cent applicable to a unique problem. Copilot not only provided a solution to the error message but also highlighted why the solution works and identified best practices for setting up code and data.

Case study: Data analysis

What: Quickly absorb dense paragraphs of modelling results, with many different sets of numbers to keep track of and compare.

Prompt: Highlight paragraph with modelling results in a Microsoft Word document, click the floating copilot icon, and select 'visualise as a table'.

Value: This saved me time scrolling back and forth, re-reading numbers and comparing them. It also significantly reduced the cognitive load required for this task, making it easier to absorb the information and leaving me less tired and able to continue onto other tasks.

This gave me more time and headspace to engage in the data analysis, giving me more time to identify any errors or important messages in the results.

Case study: Supports strategic analysis

What: Reduce intensive manual handling of information, allowing more time on analysis.

Prompt: Based on the media summary table in this document, provide me a summary of each item with the following structure: [Website name] Headline in bold. Date. First sentence of article.

Value: Copilot does basic tasks of moving information into the format I want, at a faster speed and with less chance of a handling error.

It allows me to spend more time on the analysis and commentary at the top of the media summary, where I can provide strategic comments on whether the issues in the media summary are relevant to our work and may influence upcoming policy development. It removes process effort from the task and enables me to do more qualitative, subjective analysis work without increasing the overall time to produce the media summary.

Case study: Speechwriting

What: Getting across complex long policy information from multiple sources; with Copilot able to summarise these quickly and accurately.

Prompt: Please summarise the information in / [doc]. Make it shorter. Make it shorter!

Value: I still read the document to check the results against it. But it is like having another person's opinion on what matters and what information is most important, useful or valuable within that document.

I'm yet to achieve any time savings. The benefit comes from having confirmation – 'another set of eyes' on the information that I'm looking at. This means I feel more confident when making decisions about how to structure a speech and what information to prioritise within it.

Case study: Finding a document

What: Finding out which document a screenshot is from.

Prompt: Which document is this quote from: '[TEXT]'?

Value: I was sent a screenshot of graph and needed to know which document contained the graph so I could find out how it was produced.

Unfortunately, I couldn't directly copy in the screenshot and search it, but I was able to type in some text from just above the graph in the image and ask copilot where the text came from.

This saved me a lot of time as I'm not sure I would be able to find the source of the image had it not been for copilot.

Case study: Drafting an email

What: Drafting a friendly email to external stakeholders, requesting some input and updates about relevant information.

Prompt: Can you please draft a friendly email to external stakeholders, requesting inputs in the last meeting minutes and updates on your jurisdiction's current planning system situation.

Value: The result saved me time and helped me structure the email, including effective communication in terms of what should be included in the first lines and what should go to the end of the message.

Receiving a very useful output was exciting because I only had to fix some grammar issues (replace z for s). It made me feel good that the result was very close to perfection because I don't think my prompt was as precise as it could be.

Case study: Supports improvement of work outputs

What: Use Copilot to take minutes in a meeting, and quickly summaries information and reports, write better and more concise summaries, write better first drafts of briefs.

Prompt: For removing personal information from consultation submissions: *check for personal information and inflammatory comments. For finding out who was responsible for a task: from my chats and meetings, who was responsible for XXX.*

Value: Copilot had a very functional value in providing helpful information on communications between staff that have been buried or confused in emails/teams chats.

It was very functional in summarizing large amounts of information that you did not have time to search for yourself. It was also very good at rewriting sentences and paragraphs in the teams app. These were valuable in improving the quality of my work outputs.

References

Australian Public Service Commission. (2024). *2024 APS Employee Census: Highlights Report Treasury.*

Krumpal, I. (2013). Determinants of social desirability bias in sensitive surveys: a literature review. *Quality & Quantity, 47*(4), 2025–2047.

Microsoft. (2024). *Copilot Dashboard update – Features and data interpretation guide*. Retrieved October 18, 2024, from https://techcommunity.microsoft.com/t5/viva-insights-blog/copilot-dashboard-update-features-and-data-interpretation-guide/ba-p/4165494

Nous Group. (2024). *Evaluation of the whole-of-government trial of Microsoft 365 Copilot: Summary of evaluation findings.* Commonwealth of Australia (Digital Transformation Agency).

Nous Group. (2024). *Full report: Australian Government trial of Microsoft 365 Copilot*. Retrieved January 7, 2025, from https://www.digital.gov.au/initiatives/copilot-trial/microsoft-365-copilot-evaluation-report-full/employee-related-outcomes

Prime Minister of Australia. (2023). *Media Release: Australian Government collaboration with Microsoft on artificial intelligence.* Retrieved October 18, 2024, from https://www.pm.gov.au/media/australian-government-collaboration-microsoft-artificial-intelligence