# Testing of Google's Privacy Sandbox proposals: CMA summary

13 June 2025

The Competition and Markets Authority has excluded from this published version of the final report information which it considers should be excluded having regard to the three considerations set out in section 244 of the Enterprise Act 2002 (specified information: considerations relevant to disclosure).

# Contents

# 1. Summary

1.1 In 2022, Google provided various commitments in relation to its plans to replace third-party cookies ('**TPCs**') and other functionalities in Chrome with the 'Privacy Sandbox' tools. These commitments included evaluating the effectiveness of the Privacy Sandbox tools that are amenable to quantitative testing against criteria including the impact on publishers, advertisers, and other ad tech providers.[1] In the original commitments framework, the testing would have been one source of evidence to inform the extent to which the Privacy Sandbox tools would mitigate competition concerns identified in relation to Google's original plans to remove TPCs in Chrome.

1.2 The context in which we are publishing the results of this testing differs from that originally anticipated, because on 22 April 2025 Google announced that it intends to maintain the current approach to supporting TPCs in Chrome, and will not be rolling out a new, standalone prompt for TPCs.[2] The testing results therefore relate largely to a hypothetical scenario where Google continued with its original plans to deprecate TPCs (and, to a lesser extent, alternative approaches which may have had similar effects such as providing Chrome users with a prompt on whether or not to retain TPCs). This said, the observed utility of the Privacy Sandbox tools available during 2024 H1 testing remains relevant insofar as Google preserves the tools for use alongside TPCs. In this respect, the results inform our view on the relevance of the tools to competition in online advertising in future, as relevant to our Notice of Intention to Release Commitments.[3]

1.3 We worked with Google to develop a flexible testing framework in which both Google and third-party market participants could test the Privacy Sandbox tools' effectiveness within the context of their own businesses. To this end, we published testing guidance in June 2023[4] and additional guidance in October 2023.[5] Google provided certain testing functionality from November 2023, and from January 2024 Google provided additional testing functionality with a mechanism for coordinated testing involving labelled browser groups (Mode A and Mode B).[6] Testing proceeded across Q1 and Q2 2024. Several third parties devoted significant levels of resource to participate in testing and provided invaluable engagement which greatly assisted our oversight of

---

[1] See the Commitments, paragraph 17c.
[2] See Google, Next steps for Privacy Sandbox and tracking protections in Chrome, 22 April 2025 (accessed on 12 June 2025).
[3] See the CMA's Investigation into Google's 'Privacy Sandbox' browser changes case page.
[4] See the CMA's CMA guidance to third parties on testing.
[5] See the CMA's Additional CMA guidance to third parties on testing.
[6] See Google, Chrome-facilitated testing (accessed on 12 June 2025).

Google's proposals, as well as providing valuable additional datapoints. Testing the effectiveness of an evolving technology using a small sample of traffic required businesses to take pragmatic decisions in their experimental designs to make the exercise achievable. We evaluated the limitations in experimental designs across the range of results we received. Our view is that the testing provides valuable evidence to support a 'snapshot' view of the utility of the evolving Privacy Sandbox tools from Q4 2023 to Q2 2024 and the impact of restricting TPCs. However, as set out in our previous publications preparing for the testing period, our intention was for the testing to act as only one source of evidence in our evaluation of Google's proposals.

1.4 The testing demonstrates that absent signals from TPCs or the Privacy Sandbox, publishers and advertisers would lose a significant proportion of the value they gain from digital advertising today. Google's provision of the Privacy Sandbox tools would mitigate some of the impacts of restricting TPCs. However, the test results show that even with the Privacy Sandbox tools as available at that moment in time, outcomes such as publisher revenue would remain materially below those seen on traffic with TPCs.

1.5 In relation to the magnitudes of these effects, the variation in buy-side outcomes (ie amongst businesses involved in the purchase of space to place advertising) makes it challenging to provide a single overall summary metric. However, on the sell-side (ie amongst businesses involved in selling ad space) where we received a greater number of comparable results, we saw that publisher revenue on each impression was around 30% lower without TPCs even accounting for the availability of the Privacy Sandbox at the time of testing.

1.6 The testing evidence also provides some insights into how these impacts may vary across different business models. In particular, there is some evidence that Google's owned and operated advertising businesses would be less affected by these changes than its open display businesses. However, of the two businesses tested (YouTube and Search Ads), YouTube was the more severely affected, with revenue impacts falling within the range we saw for open display ad techs. Within open display, some individual third parties appeared to experience larger impacts than Google, though others appeared similarly or less severely affected. For example, retargeting-focussed businesses appeared to experience greater impacts, though some of these businesses also identified steps to mitigate these impacts.

1.7 In evaluating this experiment, we sought to avoid an overly static analysis of these results by considering how they would generalise outside the experimental context in which they were generated. We identified several ways in which real-world outcomes could have been better than testing

suggested. For example, only a limited section of the ad tech and ad inventory ecosystem engaged with the testing – and many testers suggested that greater uptake of the Privacy Sandbox tools as TPC restrictions come into force would improve outcomes.[7] As another example, we identified that Google had already made some improvements to the tools in the period after the testing took place, and that further improvements had the potential to increase the utility of the Privacy Sandbox tools over time. We also had regard to the potential for alternative identifiers to complement the Privacy Sandbox tools and mitigate the impact of TPC loss, had Google continued to make these alternatives available in Chrome.

[7] See, for example, Kobayashi, Johnson and Gu, Privacy-Enhanced versus Traditional Retargeting: Ad Effectiveness in an Industry-Wide Field Experiment, 30 September 2024, page 32 (accessed on 12 June 2025).

# 2. Testing framework and guidance

2.1 Google and the CMA identified that three Privacy Sandbox APIs were amenable to quantitative testing. These were the Protected Audience API ('**PA API**'), the Attribution Reporting API ('**ARA**') and the Topics API ('**Topics**').

2.2 All APIs were tested using an A/B framework, involving comparing outcomes with and without the APIs across similar traffic.[8] We describe further the framework for this testing below. ARA was additionally tested through an A/A framework, which involved measuring outcome variables for the same traffic using ARA and a status-quo technology and assessing the difference.[9] PA API and Topics were not amenable to A/A testing in the testing framework Google developed.[10]

2.3 Google established infrastructure within Chrome to enable A/B testing. In doing so, it was necessary to trade off including sufficiently large volumes of traffic to provide statistical power, with minimising disruption to browser users and industry. In particular, Google considered that increased sample sizes could result in stability issues and costs for some websites. It was also necessary to trade off establishing an experimental framework quickly, with providing one which reduced the risks that experimental results could be influenced by behaviour outside of the experiment.

2.4 In recognition of these trade-offs, Chrome provided two experimental designs. These collectively assigned users' Chrome browsers to different groups based on labels, modifying in some cases browser functionality, across around 10% of traffic.[11]

2.5 In November 2023, Google provided a persistent set of labels for a set of randomly assigned Chrome browsers.[12] This facilitated experimental design 1

---

[8] A/B tests compare outcomes in one sample with those of another, typically constructing the samples to be equally representative of a wider population. One sample is typically subjected to a particular change while the other sample is not, in this case allowing the testing to assess the differences in the availability of functionality associated with user tracking. See the CMA's guidance to third parties on testing.

[9] In an A/A test, outcomes on the same data sample are measured in two different ways and then compared. The main advantage of A/A testing is that the internal validity of results does not hinge on differing characteristics of samples as in an A/B test. Traffic not reliant on TPCs (eg conversions measured through first-party cookies) was included in both treatment and control, continuing to be measured in the same way. See the CMA's Additional CMA guidance to third parties on testing, paragraph 12.

[10] Running an A/A test for these APIs would have required running auctions in parallel for the same impressions, which would in turn have raised substantial risks around experimental participation differing from real-world auction participation.

[11] The Privacy Sandbox tools were also available, although with some limitations, across unlabelled traffic.

[12] See Google, Chrome-facilitated testing (accessed on 12 June 2025).

(or mode A). In this design, TPCs and the Privacy Sandbox tools were available across all groups, although testing participants were to bid on impressions in each group using the relevant technologies as described below.

*(a)* One set of labelled browsers represented the continued availability of TPCs without the Privacy Sandbox tools. This set of labelled browsers was called 'Control 1', representing 1% of all traffic.

*(b)* Another set of labelled browsers represented the removal of TPCs but with availability of the Privacy Sandbox tools. This set of labelled browsers functioned as the experimental design 1 treatment group (also called 'label only' traffic), representing 7.5% of all traffic.[13]

2.6 In January 2024, Google provided further labels associated with a small fraction of browsers which had TPC signals suppressed. [14] This facilitated experimental design 2 (or mode B). This design provided auctions where no participants (including ad tech businesses not participating in the experiment) could use TPCs.[15] Google provided two groups of such traffic:

*(a)* One set of labelled browsers for which Google suppressed *both* the Privacy Sandbox and TPCs. This set of labelled browsers formed an additional benchmark for the effectiveness of the Privacy Sandbox tools (holding unavailability of TPCs constant), representing 0.25% of all traffic. This traffic was called 'Control 2'.

*(b)* One set of labelled browsers represented the removal of TPCs but with availability of the Privacy Sandbox tools. This set of labelled browsers functioned as the experimental design 2 treatment group representing 0.75% of all traffic.

2.7 These groups were designed to be broadly representative of the population of Chrome users. As discussed in paragraph 3.13 below, some users (eg those using old versions of Chrome) were excluded from the experiment. However,

---

[13] As set out in the CMA's testing guidance, Privacy Sandbox tools were enabled in Control 1 primarily to support measurement of conversion outcomes using the same technology across groups and due to the technical complexity for Google to suppress all Privacy Sandbox tools except the APIs necessary for measurement. See Additional CMA guidance to third parties on testing, paragraphs 11-13.

[14] The two experimental designs overlapped to the extent that the status quo in both designs was represented by Control 1.

[15] As we explain further below, this design reduced the risk that testers overestimated the impacts of removing TPCs. Under Experimental Design 1, testers may have competed in auctions against businesses using TPC signals sent by the browser, and which therefore had access to more information than would exist under Google's proposals. This was lower risk under Experimental Design 2.

where users were excluded, they were excluded both from the treatment and control groups.

2.8 Using these labelled groups, testers could execute A/B tests comparing the status quo (represented by the Control 1 labels) with the signals available following Google's proposals (represented by the treatment group, which excluded TPCs but included the Privacy Sandbox tools). Experimental design 2 also allowed testers to compare the status quo with a scenario in which TPCs were removed *without* Google providing the Privacy Sandbox tools. Testers were able to choose which design to use (or to use a combination of designs), taking into account their business's specific circumstances, such as the prospect of achieving sufficient sample sizes using only experimental design 2. Testers could also execute A/A tests for ARA using these groups.

2.9 Testing was conducted by publishers, sell-side and buy-side advertising businesses. Publishers supply ad inventory to sell-side advertising businesses. Sell-side businesses provide technology to automate the sale of digital inventory, for example by running real-time auctions based on bids for advertiser budgets and in some cases by managing publisher inventories. Buy-side businesses provide platforms allowing advertisers to buy ad inventory from many sources, which involves bidding on ad impressions based on the buyer's objectives and on data about the final user.[16]

2.10 This particular testing framework meant that buy-side testers were dependent on sell-side businesses passing through Privacy Sandbox labels on all traffic. We encouraged sell-side businesses to pass through these signals and buy-side testers to adopt mitigations where such signals were incomplete.[17] More generally, our testing guidance encouraged all testers to make sure the campaigns in which they participated across the controls and treatment group were not systematically different.[18]

2.11 All testers could use a variety of other signals aside from TPCs to assign ads to ad requests, for example first-party publisher data and contextual information. Our testing guidance recommended that testers retain signals they proposed to use following the removal of TPCs in both the control and treatment groups to the extent that these signals would not be impacted by the removal of TPCs and the introduction of the Privacy Sandbox tools.[19]

---

[16] See the CMA's Digital Advertising Market Study, Appendix M: Intermediation in open display advertising, page M4 for further description.
[17] See the CMA's Q1 2024 update report, paragraph 127.
[18] See the CMA's Additional CMA guidance to third parties on testing, paragraph 25.
[19] See the CMA's CMA guidance to third parties on testing, paragraph 10.

2.12 Whilst our guidance set out interest in some particular types of metrics (such as revenues per impression and latency) and also provided some advice on the reporting of these metrics, testers were able to determine the most relevant outcome metrics and presentation based on their own business models. Given the outcome metrics reported would be based on samples of data, we encouraged the reporting of confidence intervals which communicate the precision with which outcome metrics were recorded.

# 3. Interpretation of experimental designs

3.1 Within the framework above, we consider how different aspects of the testing methodologies used might affect the informativeness of the results in relation to the expected outcomes of Google's proposals if fully implemented. As above, we note that Google no longer intends to implement these proposals.

3.2 In this section, we begin with issues potentially limiting the validity of results within the context of the experiment – this is called the 'internal validity' of the experiments. We then go on to discuss issues which govern the 'external validity' - how far the results generated in the experimental context would generalise outside of the specific experimental context, were Google to have rolled out wider TPC restrictions than existed during the testing period.

3.3 Given the focus of experimenters' submissions was on PA API, and given the complexity of some of the methodological issues raised in these reports, we focus our discussion on PA API. However, as set out above, testers were able to test ARA using an A/A testing methodology and a small number of testers provided results using this approach.[20] We consider that A/A testing was not subject to significant methodological issues.

## Internal validity

3.4 We consider that the most important methodological issues relate to imbalances in participation; imbalanced sample attrition; TPC signal leakage; displacement of treatment effects; imperfect outcome visibility; and different approaches to signal loss. We expand on each of these below, and Annex 1 provides a fuller explanation and additional considerations beyond these key issues.

3.5 **Imbalances in participation**: Random assignment of observational units between treatment and control groups helps ensure valid inference. Google undertook various checks which showed that users' Chrome browsers were randomised across testing groups in a broadly sufficient way. However, the experiment included actors beyond end-users of web browsers, particularly ad techs, advertisers and publishers. We understand that publishers and their ad tech providers had to invest in Privacy Sandbox tools to put inventory for sale in the treatment group, and that this deterred some publishers from providing

[20] Others provided more qualitative or descriptive observations, for example by examining reporting data from ARA debugging. A small number of testers also provided results on Topics: given the varying approaches and more limited overall focus, we consider Topics-specific methodological issues in line with our discussion of the results.

some of their available inventory to the market.[21] It is likely that this behaviour would have reduced publisher revenue in the treatment group, compared to what it would have been if the technologies had been more matured and more widely adopted.[22] Given the increase in participation in the period in which third parties tested relative to when Google tested, this issue is likely to be most acute in Google's test results – but likely affected all test results to a degree.[23] Participation effects also went beyond ad techs: some experiments were more likely to be affected by advertiser and publisher self-selection, due to certain testers choosing to work with advertisers on a case study set of results, and third parties' need to actively integrate publishers into some Privacy Sandbox tool testing across their inventories.

3.6 **Imbalanced sample attrition**: Certain factors during the experiments caused groups to diverge further from a fully randomised benchmark. In particular, buy-side testers were affected by sample attrition arising from dependence on sell-side testers to forward labels; invalid traffic imbalanced test group sizes which affected impression-based metrics; and high latency which led to attrition of impressions implying risks of survivorship bias in estimates of spend and conversions. We place more weight on testing results which addressed these risks more thoroughly. However, we recognise some were insurmountable.[24]

3.7 **TPC signal leakage**: To avoid interference with measuring the impact of restricting TPCs, TPC signals should not have affected outcomes in the treatment group. Entirely avoiding leakage of TPC signals between groups was unrealistic in this setting, given for example the multiplicity of actors working across these groups.[25] However, minimising leakage was desirable and achieved to different extents across testers. In particular, some testers used experimental design 1 which involved treatment groups where TPCs were not disabled. Analysis involving these groups risked overstating the impact of removing TPCs, because testers may have been bidding against businesses outside the experiment who continued to use TPCs and therefore

---

[21] Further, in those auctions which took place, testers considered that reduced buy-side participation in the treatment group led to less competition for impressions. This could have been offset by increased willingness to pay for inventory which was auctioned, arising from the reduction in treatment group auctions taking place. It is unclear which effect was the more significant – we return to overall price impacts in the results section.

[22] See for example Kobayashi, Johnson and Gu, Privacy-Enhanced versus Traditional Retargeting: Ad Effectiveness in an Industry-Wide Field Experiment, 30 September 2024, page 32; Google, Testing privacy preserving signals in the absence of third-party cookies on Google's display ads platforms, July 2024, pages 7-8 (accessed on 12 June 2025).

[23] We discuss other challenges under this heading (including participation impact on conversion measurement and mix effects of advertisers, ad techs and publishers) in Annex 1.

[24] We discuss these challenges in more detail in Annex 1.

[25] Browsers were assigned to different groups, however users, advertisers and ad techs operated across groups.

had informational advantages. Further, some testers' results were affected by 'ID bridging', in which TPC signals were effectively reinstated into auctions;[26] and some testers used bidding models more reliant on historic TPC data. Generally, such contamination reduced the observed differences between treatment and control groups.[27]

3.8 **Displacement of treatment effects**: The impact of TPCs could also be mis-measured insofar as experimental traffic interacted with non-experimental traffic, in a way that is unrepresentative of how advertising will work outside the experimental setting. Specifically, advertisers may have shifted budgets to open display traffic retaining TPCs outside the experiment or to the control group – which will not be feasible to the same extent following TPC restrictions. Some testers sought to adjust for this by levelling spend or performance metrics across each experimental group, implying any such budget displacement would affect treatment and control equally and have no distortive impact in comparing the two. However, equalising all displacement may have introduced additional bias by preventing relative differences that *are* of interest – namely budget shifting outside of open display to O&O sites for example. The experimental setting did not provide for strategies to wholly address displacement biases, and assign weight on results whether or not they made adjustments. However, the prevalence of levelled spend approaches implied we could place less weight on third-party sell-side tests (since such strategies artificially reduce publisher revenue differences between groups).[28]

3.9 **Imperfect outcome visibility:** Businesses in digital advertising value various outcomes. For example, a buy-side participant may value both lower spend and greater number of conversions; a publisher would value higher revenue. Each outcome metric therefore only captures some of the treatment effect, and focussing on any individual metric may understate the total impact which flows across metrics. We place more weight on testing results providing various outcome metrics allowing us to see at least descriptively how a range of outcomes varied. We also take into account participants' specific business models, and did not interpret causally any metrics which were held (in part) fixed by testing strategies as discussed above. We place more weight on results with sufficiently large sample sizes to at least give confidence on

---

[26] ID Bridging occurs where the seller declares that a bid request has a TPC, where this did not originate from the user's browser. This can occur for example where sellers are able to recognise users based on first-party identifiers. See Kobayashi, Johnson, and Gu, Privacy-Enhanced versus Traditional Retargeting: Ad Effectiveness in an Industry-Wide Field Experiment, 30 September 2024, page 13 (accessed on 12 June 2025).

[27] We discuss other challenges under this heading (including testing visibility) in Annex 1.

[28] Google did not undertake a levelled spending strategy, and its sell-side metrics capture a time period before other third parties began testing in earnest, implying its sell-side results were not limited by this concern.

directional effects, and interpreted results in the round across testing reports and different outcome metrics.[29]

3.10 **Different approaches to signal loss:** A/B testing seeks to measure the impact of a single consistently defined change of interest, in this case a degradation in signals from TPCs and new availability of signals from Privacy Sandbox tools. However, in this experiment, testers implemented different specific changes in available signals. For example, some testers used only PA API for tracking whereas others also used alternative signals such as Topics. Further, some testers relied on more extensive mitigations to adapt to signal loss, and some interpreted Google's public statements around the types of signals available following its proposals differently. We were careful to interpret results in the context of which APIs they tested. [30]

3.11 The issues above generally reduce the extent to which we treat the experiment as providing truly causal estimates of the impact for PA API. We weight the evidence we received across testers in line with the degree to which they came close to this benchmark, but look at all the evidence in the round.

## External validity

3.12 In this section, we consider whether the experiment was characterised by conditions which would have changed, were Google to have gone on to restrict TPCs more widely than the small sample of traffic used in the testing period. To evaluate this, we consider whether the wider population is representative of the sample tested; the capability of the Privacy Sandbox tools; the capability of participants in using these tools; how far there is increased performance with increased scale; other complementary investments participants can draw on; and equilibrium effects across advertising channels.

3.13 **Representativeness of the sample to the wider population:** Google excluded certain categories of browsing from the experiment, [31] such as browsers without TPCs[32] and enterprise users, implying the experiment did not take place on a sample truly representative of the population. This

---

[29] We discuss other challenges under this heading (including imprecise estimates, challenges estimating confidence intervals) in Annex 1.
[30] We discuss other challenges under this heading (including spillover between testers arising from different treatment approaches) in Annex 1.
[31] These exclusions were generally made to protect users' experience and to prevent attrition of already small sample sizes in the treatment group from these users not having access to the Privacy Sandbox tools.
[32] Either because these were disabled by users or they were browsing in Incognito or guest mode.

affected all testing in open display (including third-party testing). Most exclusions generally seemed unlikely to materially affect interpretation.[33]

3.14 **Capability of the Privacy Sandbox tools:** The experiment took place in a setting where the Privacy Sandbox tools remained under Google's development and in which testers were continuing to adapt their own capabilities to use these tools. Certain testers sought to address this issue by scoping results to traffic where functionality was more advanced at the time of testing. Where testers did not make such adjustments, we consider that improved outcomes may arise in future from expansion of functionality.[34] Even where such adjustments were made, insofar as Google improves the Privacy Sandbox tools regarding currently enabled traffic, results could improve.[35]

3.15 **Capabilities in using the Privacy Sandbox tools:** Given the nascency of the Privacy Sandbox tools, it is likely that market participants would be able to improve their own usage of the tools to increase effectiveness. Google told us that, at the time of testing, it may have been further along in its maturity in usage of and scaling out its Privacy Sandbox implementation than other participants, particularly in relation to its O&O testing where its development work was more advanced.

3.16 **Increased performance with increased scale**: Testers faced problems gaining sufficient Privacy Sandbox data to train bidding models in the context of the experiment. Further, some Privacy Sandbox tools work better with scale: for example, ARA requires various changes to support the API to be made across multiple entities' web technology, for a given conversion on one site to be linked to an ad on another site. Set against this, other APIs (such as PA API) could work less well at scale – because this could exacerbate latency issues observed even in small samples – at least without further investment and mitigations being adopted.

---

[33] We also received impacts from only a portion of the advertiser, ad tech and publisher ecosystems. We consider that Google's testing grant to third parties reduced this risk, given that businesses were required to submit results to the CMA in order to receive the grant. We note that we saw a variety of different business types. Further, Google told us that where experiment participants experience relatively higher negative impacts under testing, such advertisers will be relatively more incentivised to submit their results accordingly. See the CMA's Digital Advertising Market Study.

[34] Further, if Google prioritised the development of Privacy Sandbox tool functionality with the biggest expected uplift to utility, the results conditional only on developed functionality may not be truly representative of the tools' utility across all functionality once developed. This was not possible to fully evaluate.

[35] Particularly in evaluating ARA results, we take into account that some third parties relied on functionalities (such as high values of epsilon and sometimes debugging information provided to help test the APIs) which may not be available had google gone onto implement wider restrictions to TPCs than occurred during the testing period.

3.17 **Complementary investments**: Extending TPC restrictions may have incentivised investments in alternative identifiers where available, to the extent possible given data protection considerations. These were not generally measured in the testing results, and Google's original proposals would have restricted the ability of its advertising rivals to draw on such signals, but its announcements of and subsequent to 22 July 2024 focussed on introducing these restrictions in Incognito mode.[36] The incentives to make these investments depend on the future state of competition and market structures which a small and limited duration experiment does not fully reveal. Google told us that it expects that market participants will develop new and improved technologies complementing Google's APIs. The testing process did not provide a significant amount of evidence assessing the potential for complementary and substitutable technologies of this kind to enhance functionality associated with user tracking beyond the Privacy Sandbox tools on which experiments primarily focussed.[37] In particular, relatively few testers provided evidence on the incremental value of alternative identifiers.

3.18 **Equilibrium effects across advertising channels**: Impacts of wider TPC restrictions might be mitigated by behaviour shifts beyond those a short run experiment can capture. For example, business models could shift to make heavier use of contextual targeting rather than personalised targeting, first-party data for targeting, and other methods of cross-site tracking, including alternative identity signals derived from browsers and/or devices where compatible with data protection legislation.[38] Some changes in the advertising ecosystem have already taken place in this direction, in part driven by uncertainty in the future availability of TPCs.

3.19 Overall, there is material uncertainty in the extent to which testing outcomes would closely reflect the experience of the ad-tech ecosystem, were Google to proceed with TPC restrictions. The scale of TPC restrictions and uptake of alternative signals (including the Privacy Sandbox tools) would be particularly key determinants of whether outcomes would be less severe than measured in the testing period. Further, the applicability of the testing results to individual businesses would vary depending on their relative exposure to the factors assessed in this section.

---

[36] See Google, A new path for Privacy Sandbox on the web, 22 July 2024 (accessed on 12 June 2025).
[37] Testing methodologies provided some variation in mitigations used, which might in principle be informative of the potential for such signals to restore outcomes. In practice however we were unable to draw any reliable inferences from this variation, given that testing reports also varied in many other dimensions, and given the experiment was not set up to test such variation specifically.
[38] See the CMA's Digital Advertising Market Study, Appendix F: the role of data in digital advertising, paragraph 215 onwards.

## Overall views

3.20 Overall, there were challenges in implementing fully robust experiments supporting valid inference to a context in which Google had gone on to implement wider TPC restrictions than occurred during the experiment. These challenges were expected in the context of a novel and voluntary industry-wide experiment into a nascent technology.

3.21 Certain testers were able to mitigate important limitations to experimental validity, but some were not addressable. We are able to assess the direction of some, but not all, of the resulting biases across testing results. We consider that testing results are best interpreted in the round and are indicative of the potential impacts of restricting TPCs at the point in time of the testing period. There are several reasons why the experimental conditions and therefore measured impacts could differ from those which would arise under wider restrictions to TPCs than Google implemented in the testing period.

# 4. Testing results

4.1 Testers followed a range of methodologies, which varied according to their market position and the data they were able to observe, and which affect the interpretation of these results. We first summarise results from buy-side testers, before moving onto sell-side testers.

4.2 We received reports from 25 businesses, including results from Google and 24 third parties. One of these third parties (a DSP ad tech) did not provide testing results but rather qualitative analysis or engineering observations on the Privacy Sandbox tools, which go beyond the scope of this testing report and which we have considered in our API assessment work.

4.3 We discuss Google's results in detail below, relating to: (i) its DV360 and Google Ads buy-side open display businesses; (ii) its GAM, AdSense and AdX sell-side businesses; and (iii) its O&O business for advertising on Google Search and YouTube.

4.4 Following this, we summarise in turn the methodologies and results from the 23 third parties who submitted testing results. These include nine buy-side third-party reports, thirteen sell-side third-party reports (of which four were from publishers, and nine from sell-side ad techs), and one third party in both categories. We are mindful that the testing reports we received do not cover the full spectrum of third parties which would be affected by TPC restrictions: for example, the testing results we received only related to open display advertising.

## Google

4.5 Google divided its testing of the Privacy Sandbox tools into three main experiments: a combined test of the three main Privacy Sandbox tools in open display; a standalone test of ARA in open display; and a standalone test of ARA in O&O. Google published the results of its combined tests on 22 July 2024,[39] and its respective ARA tests on 19 December 2024 and 28 March 2025.[40] We first discuss methodology across these tests, before turning to the results.

---

[39] See Google, Testing privacy preserving signals in the absence of third-party cookies on Google's display ads platforms, 22 July 2024 (accessed on 12 June 2025).
[40] See Google, Google Marketing Platform Non-O&O/Display ARA E2E CMA Test Study, 19 December 2024; and see Google, Google Search and YouTube Ads Measurement CMA Testing Results without Third-Party Cookies, 28 March 2025 (accessed on 12 June 2025).

### *Methodology*

4.6 We first evaluate the methodology of Google's published combined open display test before turning to its ARA tests.

*Open display combined test*

4.7 Google undertook a 'combined test' for its open display businesses which evaluated the effectiveness of the Privacy Sandbox tools in combination– specifically, the PA API, ARA, and Topics. The test used an A/B methodology and took place between January and March 2024.

4.8 Google identified that its open display testing could be affected by similar internal validity challenges in its experimental report to those set out in paragraphs 3.4 to 3.10. In particular, Google's observations covered (i) imbalances in participation;[41] (ii) imbalanced sample attrition; (iii) TPC signal leakage; (iv) displacement of treatment effects; and (v) imperfect outcome visibility. Google has also discussed a large proportion of the full issues identified in Annex 1 with us.

4.9 We consider the caveats Google provided alongside its testing report to be appropriate. In our view, Google's testing was particularly subject to some of these biases, although less materially subject to others. In particular, we assess the degree of bias on each aspect as follows:

*(a)* **Imbalances in participation**: Google's tests were materially subject to ad tech participation biases, primarily because Google conducted its testing in a period before third parties had begun testing the APIs at scale. Google's results may have been less subject to advertiser or publisher self-selection, given Google did not provide direct options to opt-out of the experiment.

*(b)* **Imbalanced sample attrition**: Google's tests varied in the extent to which they were subject to labelling attrition biases (ie third-party sell-side platforms not passing through testing labels). Google's DV360 results were likely to be more affected than those of Google Ads, because of the former's greater participation in third-party sell-side auctions. Due to the participation bias above, Google was less exposed to material volumes of

[41] Google explained key concepts we have discussed under these headings under the names (i) 'Absence of industry 3PCD readiness'; (ii) 'Chrome Mode A/B label and its implications on the experiment'; (iii) and (iv) 'Limitations of A/B testing'; (v) 'Limitations of Conversion-based Metrics' respectively. See Google, Testing privacy preserving signals in the absence of third-party cookies on Google's display ads platforms, 22 July 2024, pages 7-8 (accessed on 12 June).

third-party SSP traffic in these tests, implying this factor may not have introduced a substantial bias in addition to that from imbalanced participation.

*(c)* **TPC signal leakage**: Google's tests were somewhat less materially subject to these biases, because it drew only on traffic in experimental design 2, although such biases could have entered through bidding models and the history of training data or potentially through ID bridging.

*(d)* **Displacement of treatment effects**: Google told us that its ads business processed bid requests in a largely independent fashion, rather than optimising across bid requests and that this should lead to limited implicit interaction between the control and treatment in ad selection.

*(e)* **Imperfect outcome visibility**: Google's tests were less materially subject to these biases, because it had a particularly large sample size which allowed reasonably precise estimates even of rare outcomes such as conversions, and because Google evaluated a range of metrics. However, Google's testing was still subject to many other sources of this bias, such as differing definitions of conversions between advertisers.

*(f)* **Impact of different approaches to signal loss across testers**: Given lower third-party ad tech participation in this period, Google's tests were less materially subject to bias being introduced by different approaches across testers in auctions.

4.10 Google also identified that external validity challenges, similar to those set out in paragraphs 3.13 to 3.18, affected interpretation of its open display testing results. In particular, Google's observations covered (i) equilibrium effects across advertising channels; (ii) utility of the Privacy Sandbox tools; (iii) capabilities in using the Privacy Sandbox tools; and (iv) benefit of complementary investments, such as in alternative privacy-preserving signals.[42] Google has also told us that some APIs (particularly ARA) will see increased performance with increased scale. Overall, Google expected performance of the Privacy Sandbox tools to continue to improve over time as more publishers, advertisers, and ad tech partners continue to adopt and

[42] Google explained key concepts we have discussed under these headings under the names (i) 'Long-term effects of 3PCD'; (ii) 'Upcoming implementation & optimization for 3PCD-readiness'; (iii) and (iv) Absence of industry 3PCD. See Google, Testing privacy preserving signals in the absence of third-party cookies on Google's display ads platforms, 22 July 2024, pages 7-8 (accessed on 12 June 2025).

optimize their use of the Privacy Sandbox APIs and other privacy-centric solutions.[43]

4.11 In relation to external validity challenges, we particularly observe Google's decision to exclude from its headline experimental results traffic which was not yet PA-enabled but which Google planned to address in future. A broadly similar approach was followed by many third parties. This implies Google's testing results measured the efficacy of a given scope of Privacy Sandbox capability (aligned to Google's intentions at the time of the experiment relating to the development of Privacy Sandbox). Google's headline results do not therefore fully evaluate the impact of Privacy Sandbox tools having a materially more limited scope than exists in open display advertising through TPCs today.[44] To assess the impact of this decision, we evaluate additional data cuts Google provided to us without such exclusions.[45]

4.12 Google also varied its approach from that commonly adopted by third parties, by using different measurement capabilities in the treatment group to the status quo (by using ARA in the former, and its existing measuring technologies in the latter).[46] It also means that Google's testing would be best interpreted as the cumulative effect of shifting to all Privacy Sandbox tools in combination. This meant that Google's published open display test could not assess how far each tool individually contributed to outcome metrics.

4.13 For these reasons, we consider that Google's combined open display test for PA API, ARA, and Topics was insufficient by itself to evaluate the Privacy Sandbox tools' impacts on open display. We therefore interpret Google's test in combination with its open display ARA test, and provide our overall views on methodology below.

---

[43] See Google Ad Manager Help, Results from Privacy Sandbox APIs testing, 22 July 2024 (accessed on 12 June 2025).

[44] Google highlighted that it had removed certain campaigns from its analysis. For example, Google explained that campaigns with third-party features which need to be integrated and tested by third-party providers to be compatible with Protected Audience API were removed across all experiment arms from the analysis. See Google, Testing privacy preserving signals in the absence of third-party cookies on Google's display ads platforms, 22 July 2024, page 8 (accessed on 12 June 2025).

[45] These showed lower performance for several data cuts. See Google, Testing privacy preserving signals in the absence of third-party cookies on Google's display ads platforms, 22 July 2024, page 3 (accessed on 12 June 2025).

[46] ARA is more likely to systematically undermeasure conversions than overmeasure them. This bias would only exist in the control group. This implies that Google's experiment could not disentangle any reduction in conversions genuinely attributable to the Privacy Sandbox tools, and any further reduction in the observability of conversions which still occurred. Google could have chosen to report conversions measured using the same technology across groups, whilst using different technologies in underlying bidding models to continue to allow its test to capture any impacts of measurement degradation in prediction and bid valuation.

*ARA tests*

4.14 Google undertook a standalone ARA test focussed on its open display business. The timing of the standalone ARA experiment broadly aligned with the end of the combined experiment. Google followed an A/A test in which the same conversions were measured both using ARA and TPCs, for all browsers and devices where the API integration was completed.

4.15 We identify no material internal validity concerns with this test. In relation to external validity, we observe the following elements which affect how we would interpret the headline results in a context where Google proceeded with wider TPC restrictions:[47]

*(a)* **Increased performance with increased scale**: Google identified that results were highly dependent on the ecosystem enabling ARA registration. Prior to its decision no longer to restrict TPCs, Google said that it is actively collaborating with third-party publishers and ad tech providers to encourage the adoption of the ARA.

*(b)* **Capability of the Privacy Sandbox tools**: Google identified ongoing work to improve ARA, including optimizing the contribution budget in the Aggregate API per customer. Google expected further work could improve results.

*(c)* **Capability in using the Privacy Sandbox tools**: Prior to its decision no longer to restrict TPCs, Google said that it was working with third parties to optimize their conversion setup in the ARA to help improve conversion loss.

4.16 Moving beyond open display, Google also tested the impact of moving to ARA instead of TPCs when measuring conversions for advertisement campaigns on its O&O businesses, focussing on Search and YouTube Ads. Google measured bidding/revenue impacts through an A/B test,[48] and underlying measurement quality through an A/A test.

4.17 We identified no material internal validity concerns with this test. We particularly observe two elements in relation to the capability of the Privacy

---

[47] See Google, Marketing Platform Non-O&O/Display ARA E2E CMA Test Study, 19 December 2024, page 7 (accessed on 12 June 2025).

[48] Bidding/revenue impacts arise because degraded attribution of ad interactions on Google's O&O sites to conversions occurring on third-party websites could affect the prices paid in ad auctions for O&O impressions. Google measured this by randomising browsers into a treatment group using ARA rather than TPCs to measure conversions, and a control group which continued with status quo measurement of cookies (including using TPCs). Google trained separate automated bidding models on the available data in each group.

Sandbox tools which affect how we would interpret the results, in a context in which Google proceeded with wider TPCs restrictions.

*(a)* **Scoping**: Google scoped the test to its AdWords conversion tracking tool (given Google's ad business integration of (ie readiness to use) ARA with its other tools is not ready) and particular campaign strategies.[49] As in the combined experiment, outcomes could be worse if Google's experiment had measured impacts on all traffic on which Google ultimately intends to use ARA (including that where integration of ARA was insufficient at the time of testing).

*(b)* **Improvements to the Privacy Sandbox tools**: Prior to its decision no longer to restrict TPCs, Google identified that ongoing enhancements to ARA and in Google Ads' implementation of the API would be likely to drive further improvements in the performance of ARA, such as improving its Aggregate API configuration.

4.18 Google submitted (in the context of its original plans) that the results from the O&O testing should be interpreted as directionally indicative of the point-in-time impact on O&O on a standalone-basis. Google also submitted that the results from the open display environment should be interpreted as directionally indicative of the point-in-time impact on the open display advertisers on a standalone-basis. Google also stated that ARA results across O&O and open display are not naturally comparable to each other, and identified some differences in support of this. In particular, Google told us that there were differences in:

*(a)* testing goals and experimentation methodology (including different APIs being tested and use of different experimental and statistical designs and different testing traffic);

*(b)* advertiser goals, campaigns and channel inventory; and

*(c)* stages of maturity and readiness (including different dependency on third parties for both implementation and auction participation, and maturity of Google's own implementation).

*Overall assessment of testing approach*

4.19 We considered these responses, coming to the following views:

---

[49] Google considered Search campaigns following a Target Cost Per Acquisition autobidding strategy, and YouTube campaigns following YT Direct Response (ie conversion optimised) Video Action campaigns.

(a) **Different experimental and methodological designs**: The difference in APIs tested across Google's O&O businesses (which are based on ARA alone) and the combined experiment in open display (which are based on using a combination of ARA, Topics and PA API) do not prevent comparison across these experiments. This is because Google's O&O businesses were not intending to draw on Topics or PA API were Google to have implemented its proposals. We are therefore comparing like-for-like, in that both tests consider the likely impact of Google's proposals, accounting for the Privacy Sandbox tools which would be used by Google's businesses were the proposals implemented. Furthermore, experimental designs and traffic used also appear sufficiently comparable: for example, we consider that Google's analysis implies that traffic used in the open display experiment is sufficiently representative of the traffic used in the O&O test.[50] The differences between the testing methodologies appeared sufficiently small (in the context of the differences we see in the results) to give us confidence in some appropriately caveated comparative analysis.[51]

(b) **Different advertiser goals, campaigns and channel inventory***:* Insofar as such differences contribute to differences in measured effectiveness, this tells us how far the Privacy Sandbox tools would support the viability of advertisers retaining their approaches as at the time of testing, and is therefore a relevant dimension of outcomes.

(c) **Different stages of maturity and readiness**: There would be potential for real-world outcomes to improve from those shown in the testing results, were Google to continue investing in the Privacy Sandbox tools. It is possible that Google would invest in its channels such that open display advertising results improve more than O&O results, for instance through prioritising necessary third-party outreach.

4.20 Overall, we consider that Google's testing collectively has sufficient internal validity to evaluate the effectiveness of the Privacy Sandbox tools. Whilst

---

[50] In line with the headline combined experiment results, the O&O results would likely overstate the effectiveness of the Privacy Sandbox tools if functionality remained underdeveloped. This is because Google's scope excluded traffic where planned functionality was not yet ready (and which consequently would have experienced poorer outcomes in the test). The headline results cited for ARA include 'mitigation not ready' data and therefore are more conservative. We therefore compare the open display ARA data excluding mitigation not ready traffic with O&O data to provide a more like-for-like comparison.

[51] Google's O&O testing also included an adjustment to account for advertiser behaviour not captured within the experiment window, which was not included in its open display testing. However, Google stated that this would not fully capture longer-term advertiser responses to the results. Our interpretation of this is that the adjustment does not represent a material point of difference which might affect the reading of the results.

there are some methodological limitations, many are to an extent inevitable and do not prevent us from forming at least tentative conclusions, and others can be mitigated by reading Google's testing in conjunction with other evidence, including third-party testing.

4.21 In relation to how we interpret these results collectively: whilst there are some differences, we consider that comparisons across Google's different results would be broadly informative of the different outcomes customers could expect across open display and O&O channels, based on the stage of implementation and development of the API at the time of testing. In a context in which Google proceeded with wider TPC restrictions, were Google to continue to invest in its open display implementation and the effectiveness of the Privacy Sandbox tools, these differences may narrow over time.

***Results***

4.22 Google's open display combined test showed that both remarketing and non-remarketing solutions observed performance impacts, with these being greater for remarketing.[52]

*(a)* Advertiser spend fell on Google's DV360 platform by 14%, and on Google Ads by 11%. Spend fell particularly steeply on remarketing campaigns.[53] Across all campaign types, Privacy Sandbox tools appeared to mitigate the impacts by approximately 10 percentage points for both DV360 and Google Ads.[54]

*(b)* Conversions per dollar fell on Google's DV360 platform by 32%, and on Google Ads by 3%. The Privacy Sandbox tools appeared to mitigate the impacts by approximately 55 percentage points for DV360, and approximately 1 percentage point for Google Ads.[55]

---

[52] See Google, Testing privacy preserving signals in the absence of third-party cookies on Google's display ads platforms, 22 July 2024, page 2 (accessed on 12 June 2025).
[53] See Google, Testing privacy preserving signals in the absence of third- party cookies on Google's display ads platforms, 22 July 2024, Table 1.1 and 2.1 (accessed on 12 June 2025). Figures showing percentage point mitigation impacts are CMA calculations taking the difference between Google's 'Control 1 vs' and 'Treatment vs' figures.
[54] Comparing the treatment and control 2 groups.
[55] See Google, Testing privacy preserving signals in the absence of third-party cookies on Google's display ads platforms, 22 July 2024, Table 1.1 and 2.1 (accessed on 12 June 2025). Figures showing percentage point mitigation impacts are CMA calculations taking the difference between Google's 'Control 2 vs' and 'Treatment vs' figures. Across conversion-optimising campaigns the impact was mitigated by 4 percentage points for DV360 and 2 percentage points for Google Ads.

*(c)* Publisher revenue per impression fell 27% and 15% across traffic scoped to the GAM and AdSense ad serving platforms respectively.[56] The Privacy Sandbox tools appeared to mitigate the impacts by 13 percentage points for GAM, and 3 percentage points for AdSense.[57]

4.23 Comparison of how impressions on Google's sell-side entities varied in the treatment group compared to the control at the time of the experiment highlights the risk that Google's proposals could have benefited its open display buy-side entities under its previous proposals, by allowing Google to increase its share of supply.[58] In Google's test, GAM realised greater impressions of around 9% on traffic without TPCs but with the Privacy Sandbox tools, relative to that with TPCs. Further, its sell-side results when segmented by buying door show that DV360 and Google Ads both observed increased impressions where Authorised Buyers observed reduced impressions.[59] This is consistent with substitution away from Google's rivals towards Google's businesses during the experiment. However, given that third-party testing was not prevalent during the period Google tested, we place little weight on these results in isolation.[60]

4.24 Google's open display-focussed ARA test confirmed that ad tech and ad inventory suppliers would have a reduced ability to measure conversions using ARA compared to using TPCs, absent further investment and under Google's previous proposals. Campaigns more reliant on TPCs at the time of testing had particularly high rates of conversion mismeasurement as a result of greater dependence on ARA.[61] Google's analysis also considered the

---

[56] See Google, Testing privacy preserving signals in the absence of third-party cookies on Google's display ads platforms, 22 July 2024, Table 3.1 (accessed on 12 June 2025). Figures showing percentage point mitigation impacts are CMA calculations taking the difference between Google's 'Control 1 vs' and 'Treatment vs' figures.
[57] See Google, Testing privacy preserving signals in the absence of third-party cookies on Google's display ads platforms, 22 July 2024, Table 3.1 (accessed on 12 June 2025). Figures showing percentage point mitigation impacts are CMA calculations taking the difference between Google's 'Control 2 vs' and 'Treatment vs' figures.
[58] There is a particular risk that these figures are subject to internal validity biases stemming from participation (as discussed above). However, if adoption were to remain low due to barriers to using the tools effectively, and if third parties were not to have other substitutes, these results could still be informative of competitive conditions.
[59] See Google, Testing privacy preserving signals in the absence of third-party cookies on Google's display ads platforms, 22 July 2024, Table 3.1 and 3.2 (accessed on 12 June 2025).
[60] We return to similar third-party analysis below, since these results were produced in a period when both third parties and Google were using the Privacy Sandbox tools.
[61] Google presented its results using the proportion of campaigns with a given APE. A 0% APE would imply equal performance to the status quo, and a 100% APE would imply a mismatch equal to the total number of conversions observed. We focus on figures which include 'mitigation not ready' traffic as this is sliced by TPC reliance bucket. Using figures excluding 'mitigation not ready' traffic leads to similar general conclusions although less stark contrasts. Google defined traffic as 'mitigation not ready' where either 1) Registration or post-processing steps are not fully implemented or 2) the API does not support certain use cases by design. These figures better proxy the performance of ARA *for supported traffic*. The extent to which they represent experiences

impact on different 'slices' (ie groupings) of traffic, depending on how far Google currently depends on TPCs for measurement in that slice. In particular:

*(a)* Click-Through Conversion ('**CTC**') campaign measurements were degraded when ARA rather than TPCs was used to measure conversions.[62] Google considered errors above +/- 20% to be a very high rate. For two of three Google products tested (Campaign Manager 360 and Display & Video 360) over two thirds of CTC campaigns experienced Absolute Percentage Error ('**APE**') of above 20%.[63] For campaigns most dependent on TPCs, the vast majority of campaigns experienced very high error rates (above 60%).[64]

*(b)* View-Through Conversion ('**VTC**') campaign measurements were substantially more degraded using ARA rather than TPCs. For the same two Google products tested, over 95% of campaigns respectively experienced APEs above 20%.[65] The vast majority of campaigns experienced very high error rates (above 60%).

4.25 Google's O&O results allowed us to compare measurement impacts on two of its main O&O business segments, Search and YouTube. Google computed similar statistics as for the open display ARA test. Google tested the impact on click-through conversions for Search, and (in aggregate) on both click-through and view-through conversions for YouTube. Google's analysis showed the following.[66]

---

of the ecosystem had Google gone on to implement wider restrictions to TPCs than occurred during the testing period would depend on whether Google continues to develop the Privacy Sandbox tools - but likely represent an upper bound, given ARA is unlikely to provide utility for the fraction of mitigation not ready traffic comprised of use cases unsupported by design.

[62] Conversions attributed to preceding ad-clicks are referred to in the ad tech industry as Click-Through Conversions, while conversions attributed to preceding ad-engaged views are referred to as View-Through Conversions.

[63] See Google, Google Marketing Platform Non-O&O/Display ARA E2E CMA Test Study, 19 December 2024, page 6-7 (accessed on 12 June 2025). The overall figures fall to 56% for Campaign Manager 360 and 66% for Display & Video 360 excluding 'mitigation not ready' traffic.

[64] See Google, Google Marketing Platform Non-O&O/Display ARA E2E CMA Test Study, 19 December 2024, page 7 (accessed on 12 June 2025). For Search Ads 360, the equivalent percentages were 10% (overall) and 75% (most TPC-dependent).

[65] See Google, Google Marketing Platform Non-O&O/Display ARA E2E CMA Test Study, 19 December 2024, page 6 (accessed on 12 June 2025).

[66] Google provided the CMA with additional analysis to support its comprehension of the test results, for example revenue impacts and data on TPC dependence.

4.26 A relatively lower fraction of campaigns experienced mismeasurement rates of above 20%: specifically, only 15% of search ad campaigns and 25% of YouTube ad campaigns experienced error rates above this level.[67]

*(a)* For similar levels of TPC dependence, Search experienced lower levels of measurement degradation than YouTube. Google noted that some ARA functions were only utilised on Search Ads but not yet on YouTube Ads, which may have contributed to the variance in the performance results.[68]

*(b)* Again, traffic which depended more on TPCs (rather than other technologies to measure conversions) saw greater increases in mismeasurement, though traffic which depended the most on TPCs still appeared to experience lower impacts than broadly equivalent traffic in open display.[69]

*(c)* YouTube was more dependent on TPCs for measurement than Search was. YouTube was therefore in overall terms correspondingly more affected than Search by measurement degradation.

*(d)* These measurement impacts would lead to a small overall loss of revenue for Search [0-10%] and a moderate loss for YouTube [10-20%]. When considering the traffic dependent on TPCs for a majority of conversions, this rose to [10-20%] for Search and [30-40%] for YouTube.

4.27 We compared the impacts shown for Google's O&O business segments to those shown for its open display businesses segments. Noting the limitations inherent in this comparison explained in paragraphs 4.18 to 4.21 above, this showed greater TPC dependence, and greater measurement degradation in Google's open display compared to its O&O business segments. There was more mixed overall evidence on relative revenue impacts between Google's open display businesses and YouTube, though Search appeared to have materially lower revenue impacts.[70]

4.28 Overall, Google's tests showed that market outcomes were worse without TPCs than with TPCs. Remarketing campaigns observed larger declines in performance than non-remarketing campaigns, due to heavier reliance on

---

[67] See Google, Google Search and YouTube Ads Measurement CMA Testing Results without Third-Party Cookies, 28 Mar 2025, page 4 (accessed on 12 June 2025).
[68] See Google, Google Search and YouTube Ads Measurement CMA Testing Results without Third-Party Cookies, 28 Mar 2025, page 4 (accessed on 12 June 2025).
[69] See Google, Google Search and YouTube Ads Measurement CMA Testing Results without Third-Party Cookies, 28 Mar 2025, page 4 (accessed on 12 June 2025).
[70] As set out above, Google's open display combined test showed revenue impacts of 14% and 11% for DV360 and Google Ads respectively. By comparison, its O&O test showed revenue impacts of [0-10%] for Search.

TPCs. Similar factors drove Google's open display business observing larger impacts relative to its O&O businesses.[71] Of the O&O businesses, the more substitutable offering with open display (YouTube) experienced the greater impacts. The Privacy Sandbox tools helped to mitigate the outcomes of TPC restrictions.

## Buy-side third parties

4.29 We received buy-side results from 10 third parties: an ad tech, Adform, Criteo, CyberAgent, Kobayashi, Johnson and Gu in partnership with a buy-side ad tech (**KJG**), Nextroll, Seedtag, SMN, Teads and Yahoo. We first summarise salient features of the methodologies, before moving onto the results. We first summarise salient features of the methodologies, before moving onto the results.

### *Methodologies*

4.30 As set out above, third-party experiments followed different methodological approaches, each with different advantages and disadvantages. These often reflected differing interpretations of Google's proposal and different engineering constraints faced across the range of third parties who submitted experimental results.

4.31 Drawing on the themes identified in paragraphs 3.5 to 3.10 above, our review of the third-party testing identified the following summary points relevant to demand-side tests. Relative to Google's testing, third-party A/B experiments were in general:

*(a)* Less subject to bias from imbalanced ad tech participation across experimental arms, but more subject to bias from advertiser self-selection.[72] Third-party ad tech activity over this period was materially higher than when Google tested.[73] Set against this, ad tech participation imbalances remained, and two demand side experiments (ad techs) identified that advertisers participating in the experiment may not be fully

[71] With the general exception of Google Ads, which was comparatively much less affected by removing TPCs than DV360.
[72] Given additional challenges faced by third parties integrating publishers.
[73] Experiments typically involved multiple sell-side parties, mitigating risks identified by Criteo. See Selman, Criteo's Privacy Sandbox Market Testing, 27 June 2024, section 5.2 (accessed on 12 June 2025).

representative of their wider client base, implying some risks of selection.[74]

*(b)* Often subject to greater biases from sample attrition arising from third-party SSPs not forwarding labels. We consider that this is a consequence of third-party results including a greater proportion of bid requests from third-party SSPs in the treatment group. [75] However, two demand-side experiments (Criteo and KJG) mitigated this bias.[76]

*(c)* Equally subject to biases from TPC signal leakage. However, one demand side tester (ad tech) explained that it had mitigated this bias by simulating constraints in its bidding model, and two (ad techs) were subject to greater risks due to reliance on experimental design 1 traffic or non-experimental traffic.[77]

*(d)* Equally subject to biases from displacement outside of the treatment group. Six third parties (all ad techs) introduced strategies which reduced bias stemming from displacement to non-experimental open display traffic. Two testers (ad techs) broadly followed a levelled spend approach,[78] one tester (ad tech) followed an iso-impression approach, and three testers (ad techs) followed measures broadly reflective of cost per engagement (impression or qualified visit).[79] These same strategies could however have introduced new biases relating to unmodelled displacement outside of open display and affected how we interpreted outcome metrics as set out below.

*(e)* Often subject to challenges in interpreting outcome metrics. Third-party experiments were smaller in scale. This generally did not affect

---

[74] Several testers did not provide information to evaluate whether advertisers had self-selected into the experiments. It was however common for testers (such as two ad techs) to evaluate outcomes for a small number of advertisers, which may reduce the external validity of the results even if these tests did not involve advertiser self-selection. It is not possible to sign the selection bias with confidence, but we consider it more likely than not that advertisers particularly concerned about the performance of their advertisements would be particularly motivated to participate in the experiment, to enable their planning. This would suggest selection bias leads to overstatement of the impacts in testing, relative to those arising 'in the real world'.

[75] Criteo's analysis indicated that 45% of spending was unlabelled in Mode A. See Selman, Criteo's Privacy Sandbox Market Testing, 27 June 2024, section 2.1.2 (accessed on 12 June 2025).

[76] See Selman, Criteo's Privacy Sandbox Market Testing, 27 June 2024, section 2.1.2 (accessed on 12 June 2025); and see Kobayashi, Johnson and Gu, Privacy-Enhanced versus Traditional Retargeting: Ad Effectiveness in an Industry-Wide Field Experiment, 30 September 2024, page 7 (accessed on 12 June 2025).

[77] Multiple third parties relied on experimental design 1 for some tests, but most submitted versions of tests only using experimental design 2.

[78] One ad tech did not follow a pure levelled spend strategy, but rather implemented a daily budget. We consider this could have reduced spend differentials between groups, but not eliminated the potential for them.

[79] See Selman, Criteo's Privacy Sandbox Market Testing , 27 June 2024, section 2.4 (accessed on 12 June 2025).

interpretation around the directionality of results, but implied greater challenges in assessing results with limited precision. Two testers' (ad techs) results could largely be interpreted in direction only. Several testers reported on only a subset of relevant outcome metrics, for example excluding conversions (or reporting only click-through conversions).[80] Further, as set out above, some testers held certain outcomes constant across groups which limited the range of outcome metrics we could interpret.

*(f)* Sometimes subject to greater challenges relating to different approaches to treatment (ie the extent and nature of signal loss). Testers did not always provide complete descriptions of mitigations used due to business sensitivities, but in at least one case a third-party (ad tech) used mitigations in the status quo which were not used in the treatment group. The third-party testing period was also subject to many more differing definitions of 'treatment' (ie because different testers made use of different Privacy Sandbox tools), which could have led to more spillovers through auctions. Set against this, however, the evidence we saw did not suggest variation in treatment definitions would significantly explain much variation in outcomes.[81]

4.32 Given the variation in how far individual experiments were affected by these limitations, we place different weights on different experiments. When evaluating PA API, we consistently place weight on results received from two testers in addition to Google's results across a wide range of metrics: Criteo and KJG.[82] [83] We also place weight on various other third parties for specific metrics, and in providing directional checks to our overall conclusions, as well as to evaluate Topics.

4.33 Several third parties tested Topics using similar A/B frameworks as they used for PA API, and as evaluated above. In some cases, third parties tested whether Topics augmented PA API results. We consider that similar risks apply as above. Other third parties conducted 'offline' checks of whether Topics added to offline prediction models. This is helpful direct evidence,

---

[80] Unlike Google, several testers provided latency information: whilst not an advertising outcome metric per se, we considered this information highly relevant to evaluating qualitatively the effectiveness of PA API.
[81] For example, including or excluding Topics when testing PA API did not appear to lead to changes in results.
[82] These testers presented results which were unaffected by attrition of labels on traffic received by non-Google supply side testers. All other tests had some factors which caused us to reduce weight on their testing results, though these did not always impact individual metrics.
[83] Analysis by Criteo indicated that not adjusting for label attrition bias could inflate publisher revenues by around 50%, implying this is a significant risk; this is consistent with labelling attrition figures reported by other testers. See Selman, Criteo's Privacy Sandbox Market Testing, 27 June 2024, section 5.3 (accessed on 12 June 2025).

although more limited insofar as it does not confirm impacts on headline outcome variables. Other third parties focussed on qualitative observations around the prevalence and nature of Topics received. Despite the wide range of approaches, results were very consistent. We interpret this evidence collectively, rather than placing weight on any individual metric.

4.34 Some third parties performed A/A testing of ARA with a similar methodology to Google. Unlike Google, third parties did not set out how results varied across campaigns with different levels of TPC dependence at the time of testing. Further, some results were often provided for a small sample. Others instead used information available to debug the API. This allowed them to draw focussed (but also more limited) conclusions into the impact of ARA adding noise to conversions, when varied under parameters which were held constant in Google's approach. We place weight on all these testing results, although we noted some results were more informative of some questions than others.[84]

4.35 In paragraphs 3.13 to 3.18 above, we also set out various considerations on external validity. Third-party results appear likely to be affected to a similar extent to Google's results by these factors. However, we saw some evidence that third parties had less experience optimising for the use of the Privacy Sandbox tools than Google did. In a context where there are likely diminishing returns to experience, whilst very uncertain, this may suggest third parties' observed impacts in particular could improve 'in the real world' were Google's proposals to be implemented, relative to those measured in testing all else equal.[85]

### *Results*

4.36 Third-party results provide a valuable sense-check on Google's figures, particularly for DV360 given its greater exposure to third-party open display businesses, compared to Google Ads. The experiments also provided valuable methodological diversity and insights into impacts across firms with different business models, although these are hard to separate out in practice. We report datapoints we placed positive weight on, although we had regard in

---

[84] We also received more methodological detail from Google in relation to its ARA testing methodologies, which gave us higher confidence.

[85] One third party (Criteo) included inventory not yet PA API enabled, implying its results measure the efficacy and scope of Privacy Sandbox capability, rather than just the efficacy of a given scope of Privacy Sandbox capability. Privacy Sandbox scope expanding eg to new formats would therefore likely drive improvements in Criteo's results. See Selman, Criteo's Privacy Sandbox Market Testing, 27 June 2024, section 5 (accessed on 12 June 2025).

forming our overall conclusions to different advantages and disadvantages in methodologies as set out above.

4.37 We report all results for important metrics where we did not identify methodological reasons for omitting them. This meant we include some metrics with unexpected signs in our meta-analysis. However, such results were a small minority of datapoints reported to us.[86] [87]

4.38 We first consider how third-party metrics compared to those reported by Google in its published open display results.[88] Third-party data broadly validated Google's results on advertiser spend. For other metrics, third-party data was typically more negative than Google's, however it provided valuable robustness checks given unexpected signs in Google's experiment.[89]

*(a)* **Advertiser spend**: Three testers (two ad techs and an academic) provided comparable datapoints.[90] All of these tests found that using PA API in place of TPC signals reduced ad spend, by magnitudes of 42%, 60% and 67% respectively.[91] These estimates are more negative than Google's overall estimate of a 14% reduction for DV360.[92] The most

---

[86] Three testers (ad techs) provided a range of case studies across advertisers, rather than providing any aggregated metrics. We omit these from the summary of aggregated results below. One ad tech saw generally lower click through rates, though could not draw conclusions on other metrics given their particular testing strategy and low observed numbers of conversions. Another ad tech generally saw higher spend, impressions and higher click through rates across campaigns. Another ad tech saw mixed results in relation to click-through rates and impressions, though its test results which were most generalisable to other market participants generally showed the Privacy Sandbox tools underperformed a TPC comparator.

[87] We also exclude PA API results from an ad tech in this detailed summary, given the ad tech included non-experimental traffic in its results and given its view that the experimental design 2 traffic was too small in volumes to evaluate the impact.

[88] We omit conversions per dollar, where no testers submitted comparable data. One tester (KJG) calculated a similar metric, although unlike Google could only include click-through conversions and chose to keep measurement capability constant across groups. See Kobayashi, Johnson and Gu, Privacy-Enhanced versus Traditional Retargeting: Ad Effectiveness in an Industry-Wide Field Experiment, 30 September 2024, page 24 (accessed on 12 June 2025).

[89] We note that Google tested all APIs in combination, whilst third parties generally tested APIs in isolation. We would expect third-party results to show lower outcome degradation than Google's results, if they were otherwise equally affected: this is because third-party results are not additionally affected by degraded measurement capabilities, whereas Google's are. Whilst Google's results included signals from Topics, which might in principle offset the above, in practice third parties generally ascribed low value to Topics (discussed further below).

[90] Although an ad tech also provided measures of spend, the ad tech's iso-spend testing strategy excluded reliance on this metric. We exclude these results from the summary above.

[91] We use Criteo's figure scoped to PA-enabled supply for comparability. We use an ad tech's daily budget spend ratio estimate which is broadly comparable. All figures statistically significant. See Selman, Criteo's Privacy Sandbox Market Testing , 27 June 2024 (accessed on 12 June 2025); Kobayashi, Johnson and Gu, Privacy-Enhanced versus Traditional Retargeting: Ad Effectiveness in an Industry-Wide Field Experiment, 30 September 2024, page 24 (accessed on 12 June 2025).

[92] See Google, Testing privacy preserving signals in the absence of third-party cookies on Google's display ads platforms, 22 July 2024, page 11(accessed on 12 June 2025).

negative estimates are however broadly comparable to Google's estimates for DV360 remarketing campaigns (a 47-51% reduction), implying that variation in campaign types may explain some of the differences.[93]

*(b)* **Click-through rate**: Four testers (KJG and three ad techs) provided broadly comparable datapoints. All of these tests found that using PA API in place of TPC signals reduced click-through rates, by magnitudes of 12%, 24%, 30% and 67% respectively.[94] These estimates are more negative than Google's overall estimate of a 4% increase for DV360, although more in line with its estimates for DV360 remarketing campaigns of 18-43%.[95]

*(c)* **Clicks per dollar:** Two testers (one ad tech and an academic) provided broadly comparable datapoints. These testers found using PA API in place of TPC signals reduced clicks per dollar by 12% and 14% respectively.[96] These estimates are more negative than Google's overall estimate of around a 5% increase for DV360,[97] although Google cautioned that its ads business does not optimise for this metric and therefore it may have low meaning.

*(d)* **Conversion rate:** One tester (ad tech) provided a broadly comparable datapoint to Google.[98] This tester found using PA API in place of TPC signals reduced conversion rates by 56%.[99] This estimate is more

---

[93] See Google, Testing privacy preserving signals in the absence of third-party cookies on Google's display ads platforms, 22 July 2024, page 11 (accessed on 12 June 2025). The figures reflect Google's different estimates across campaigns and impressions.

[94] All figures statistically significant, although KJG's figure marginally significant. An ad tech did not provide PA-enabled supply scoped figures, so we use their figures scoped to all supply. See Kobayashi, Johnson and Gu, Privacy-Enhanced versus Traditional Retargeting: Ad Effectiveness in an Industry-Wide Field Experiment , 30 September 2024, page 24 (accessed on 12 June 2025).

[95] See Google, Testing privacy preserving signals in the absence of third-party cookies on Google's display ads platforms, 22 July 2024, page 11 (accessed on 12 June 2025). These figures are statistically significant. The figures reflect Google's different estimates across campaigns and impressions.

[96] All figures statistically significant. Kobayashi, Johnson and Gu, Privacy-Enhanced versus Traditional Retargeting: Ad Effectiveness in an Industry-Wide Field Experiment, 30 September 2024, page 24 (accessed on 12 June 2025). An ad tech provided a combined clicks and conversions per dollar metric, which we report as clicks per dollar.

[97] See Google, Testing privacy preserving signals in the absence of third-party cookies on Google's display ads platforms, 22 July 2024, page 11 (accessed on 12 June 2025). This figure is statistically significant.

[98] KJG estimated a click through rate based on conversions per dollar only: this estimate was negative and marginally significant. See Kobayashi, Johnson and Gu, Privacy-Enhanced versus Traditional Retargeting: Ad Effectiveness in an Industry-Wide Field Experiment, 30 September 2024 2024, page 24 (accessed on 12 June 2025).

[99] An ad tech did not provide PA-enabled supply scoped figures, so we use their figures scoped to all supply.

negative than Google's overall estimate of a 36% reduction on DV360.[100] Google's conversion rate estimates for its remarketing slice showed more positive results in the treatment group relative to traffic with TPCs: Google considered these results particularly liable to bias from imbalanced auction pressure or from short term algorithmic responses in the experiment which led Google to buy only higher quality ad inventory (even if less inventory). We placed correspondingly less weight on Google's results for these metrics.

4.39 Third-party experiments evaluated PA API against the status quo using a range of other outcome metrics. These generally showed that PA API materially underperforms the status quo at the time of testing, although the magnitudes of these estimates also varied materially across testers. These metrics were generally consistent with outcomes being constrained by low participation in the treatment group.

*(a)* **Total impressions**: Four testers (three ad techs and an academic) provided broadly comparable datapoints. All four experiments found that using PA API in place of TPC signals reduced impressions, by magnitudes of 35%, 47%, 67% and 69% respectively.[101] Testers which reduced sample attrition saw lower reductions in impressions.

*(b)* **Cost per Impression** ('**CPM**'):[102] Three testers (ad techs) provided broadly comparable datapoints.[103] Two of these tests found that using PA API in place of TPC signals reduced CPM, by magnitudes of 14% and 65% respectively. The third saw increased CPM by 225%.[104] We interpreted the variation around this metric as in large part a product of different testing strategies, and did not assign CPM results high weight.[105]

---

[100] See Google, Testing privacy preserving signals in the absence of third-party cookies on Google's display ads platforms, 22 July 2024, page 11 (accessed on 12 June 2025). This figure is statistically significant.
[101] All figures statistically significant. An ad tech did not provide PA-enabled supply scoped figures, so we use their figures scoped to all supply. See Selman, Criteo's Privacy Sandbox Market Testing, 27 June 2024, section 3 (accessed on 12 June 2025); Kobayashi, Johnson and Gu, Privacy-Enhanced versus Traditional Retargeting: Ad Effectiveness in an Industry-Wide Field Experiment, 30 September 2024, page 24 (accessed on 12 June 2025).
[102] Industry convention typically uses cost per thousand (mille) impressions, rather than per individual impression.
[103] KJG's results are consistent with slightly increased CPM. We could however not evaluate statistical significance on this metric, and have excluded it from the summary above. See Kobayashi, and Johnson and Gu, analysis provided to the CMA on 16 October 2024.
[104] All figures statistically significant. An ad tech did not provide PA-enabled supply scoped figures, so we use their figures scoped to all supply. See Selman, Criteo's Privacy Sandbox Market Testing, 27 June 2024 (accessed on 12 June 2025).
[105] Levelled spend approaches would, all else equal, be expected to result in higher CPMs in the treatment group (algorithms would bid more on the fewer impressions available in the treatment group). Levelled performance approaches might expect to see lower CPMs.

*(c)* **Clicks**: Two testers (KJG and an ad tech) provided broadly comparable datapoints. These experiments found that using PA API in place of TPC signals reduced clicks, by magnitudes of 50% and 79% respectively.[106]

*(d)* **Cost per click:** Two testers (ad techs) provided broadly comparable datapoints. These experiments found that using PA API in place of TPC signals increased cost per click by 48% and 124% respectively.[107]

4.40 Where third-party evidence provided comparators to the differences between various data cuts in Google's publication, these insights were broadly consistent with Google's findings. Retargeting-focussed experiments generally saw greater impacts of using PA API in place of TPCs.[108] Test results scoped across all ad inventory (regardless of whether PA API was enabled) also showed greater impacts.[109] Results showed mixed evidence on whether advertisers and publishers of different sizes saw greater impacts.[110]

4.41 Third-party reports provide further insight into other datapoints of more indirect relevance to advertiser and publisher outcomes:

*(a)* **Google sell-side share of supply**: Two testers (ad techs) provided comparable datapoints.[111] These experiments found that using PA API in place of TPC signals saw AdX and GAM increase their share of spend by

---

[106] All figures statistically significant. See Kobayashi, Johnson and Gu, Privacy-Enhanced versus Traditional Retargeting: Ad Effectiveness in an Industry-Wide Field Experiment, 30 September 2024, page 24 (accessed on 12 June 2025).

[107] All figures statistically significant.

[108] KJG's DSP partner and Criteo both focussed on retargeting campaigns, and showed more negative results for example on spend than those who focussed more broadly. Criteo further provided data slicing by campaign type, consistent with this explanation. See Kobayashi, Johnson and Gu, Privacy-Enhanced versus Traditional Retargeting: Ad Effectiveness in an Industry-Wide Field Experiment, 30 September 2024, page 30 (accessed on 12 June 2025); See Selman, Criteo's Privacy Sandbox Market Testing, 27 June 2024 (accessed on 12 June 2025).

[109] Criteo's data allowed us to compare the impact of scoping results to PA-enabled supply only, relative to those on all supply. Criteo's data implies an 18% improvement in spend when scoping only on PA-enabled supply. Google's data allowed us to make a similar comparison, finding a similar impact for DV360 (scoping on PA-enabled supply led to a 16% increase in spend). See Selman, Criteo's Privacy Sandbox Market Testing, 27 June 2024 (accessed on 12 June 2025); and See Google, Marketing Platform Non-O&O/Display ARA E2E CMA Test Study, 19 December 2024, pages 6-7 (accessed on 12 June 2025).

[110] KJG saw that smaller advertisers are more adversely affected by the absence of third-party cookies. An ad tech and Google did not identify this pattern in their data. See Kobayashi, Johnson and Gu, Privacy-Enhanced versus Traditional Retargeting: Ad Effectiveness in an Industry-Wide Field Experiment, 30 September 2024, page 24 (accessed on 12 June 2025).

[111] An ad tech reported a small reduction in PA API requests from GAM in its treatment group relative to Mode A label-only groups. We placed low weight on this statistic given its use of experimental design 1 for this comparison and exclude it from the headline summary above.

multiples of 2.8 and 3.6 respectively.[112] As set out above, Google also identified increases in impressions for GAM although not for AdX and over a different period of time. Interpreting these figures is complex, because third parties reported that testing was primarily skewed towards GAM traffic because very few SSPs were ready for testing PA API at scale. This implies that such increases in Google's share are unlikely to have arisen 'in the real world' under Google's previous proposals, provided these led to wider uptake of PA API. [113]

*(b)* **Latency**: Three testers (ad techs) provided broadly comparable datapoints, although there remain differences in the way in which latency was measured.[114] These experiments found that using PA API in place of TPC signals saw latency increase by 19%; 115% to 152% and around 200% respectively.[115] [116] Google did not provide a latency comparator in its testing report.

4.42 In addition to comparisons with the status quo, both third party and Google's metrics allow calculation of how far the Privacy Sandbox tools recover value lost by restricting TPCs.[117]

*(a)* **Recovery rates (scoped to affected sources):** Two testers (ad tech and an academic) explicitly calculated recovery rates.[118] One of the testers highlighted recovery rates of 17% based on spend, although this figure (as for most other of the tester's figures) includes ad inventory which was not PA enabled. KJG highlighted recovery rates of 46% and 43% based

---

[112] See Selman, Criteo's Privacy Sandbox Market Testing, 27 July 2024, section 3.2.1 (accessed on 12 June 2025); and CMA calculations based on an ad tech's data. Experimenters particularly attributed this result to Google being the only supply side entities that could enable Protected Audience API at scale, with Privacy Sandbox integration being significantly more costly and complex for other SSPs. Testers did not report statistical significance.

[113] See Google, Testing privacy preserving signals in the absence of third-party cookies on Google's display ads platforms, 22 July 2024, page 14 (accessed on 12 June 2025). These figures were statistically significant.

[114] An ad tech provided latency increases split across advertiser case studies. In all cases, the latency increases associated with PA API were high (well over 100%).

[115] See Selman, Criteo's Privacy Sandbox Market Testing, 27 June 2024, section 3.2.2 (accessed on 12 June 2025). The reports did not show statistical significance. All figures used the median impression.

[116] An ad tech has suggested that it believes that transitioning to Bidding and Auction Services could be a solution although has not tested this. See Issue #1207 on the PA API repository on GitHub (accessed on 12 June 2025).

[117] Recovery rates measure the difference between the impact measured between (i) control 1 and the treatment group and (ii) control 1 and control 2, as a proportion of the latter.

[118] Other testers provided data which could allow recovery rates to be calculated, however these were not headline metrics for the relevant testers and were in some cases affected by other experiments conducted on the Control 2 group.

on clicks and click-through conversions respectively.[119] These are within the range of recovery rates reported by Google across different metrics – which when positive, ranged between around 10% (Conversion rates) and around 50% (Spend) for DV360.[120]

*(b)* **Recovery rates (scoped to all sources)**: KJG also calculated recovery rates which took into account conversions advertisers receive from other sources (aside from those directly affected by PA API). In particular, KJG's results imply that open display retargeting advertising contributes 4.4% of advertisers' total conversions in the status quo. KJG estimated, although imprecisely and without statistical significance, that PA API recovers around 20% of these conversions, implying an aggregate impact of around 3.6% from the status quo.[121] Google's report did not contain comparator data.

4.43 Third-party tests results therefore showed that using PA API mitigated some outcomes, but at the time of testing that the API did not close more than half of the degradation in aggregate outcome metrics (such as overall publisher revenue) created by restricting TPCs. KJG's analysis suggested that a key driver of these limited recovery rates was lack of supply side adoption of the Privacy Sandbox tools. On a per-impression basis, KJG found that recovery rates were substantially higher.[122]

4.44 Looking beyond PA API, seven buy-side testers provided evidence on the utility of Topics. Of these, four (ad techs) tested outcomes of using Topics in place of TPCs, and found that this led to lower performance. Three testers (ad techs) focussed on whether Topics augmented PA API or otherwise contributed value absent TPCs: each concluded Topics contributed minimal improvements in outcomes.[123]

4.45 Six buy-side testers provided evidence on the utility of ARA. Of these, five testers (ad techs) conducted A/A tests. These tests showed differential impacts across campaigns particularly depending on the number of conversions identifiable, with high-conversion campaigns less affected.

---

[119] See Kobayashi, Johnson and Gu, Privacy-Enhanced versus Traditional Retargeting: Ad Effectiveness in an Industry-Wide Field Experiment, 30 September 2024 page 18 (accessed on 12 June 2025).
[120] CMA calculations using Google Data. Google's metrics sometimes showed negative recovery rates. See Google, Testing privacy preserving signals in the absence of third-party cookies on Google's display ads platforms, 22 July 2024, page 14 (accessed on 12 June 2025).
[121] See Kobayashi, Johnson and Gu, Privacy-Enhanced versus Traditional Retargeting: Ad Effectiveness in an Industry-Wide Field Experiment, 30 September 2024, page 18 (accessed on 12 June 2025).
[122] See Kobayashi, Johnson and Gu, Privacy-Enhanced versus Traditional Retargeting: Ad Effectiveness in an Industry-Wide Field Experiment, 30 September 2024, pages 1 and 24 (accessed on 12 June 2025).
[123] See Selman, Criteo's Privacy Sandbox Market Testing, 27 June 2024 (accessed on 12 June 2025).

Across these reports, metrics for measurement degradation generally fell between 20 and 50%, broadly consistent with Google's data. Where testers investigated the impact of reducing noise in ARA by varying the API's 'epsilon' parameter, testers found material improvements in the recovery of conversions. Testers generally argued that the highest value of 'epsilon' (implying least noise) would be necessary for acceptable impacts.[124]

4.46 Commentary in testing reports submitted by sell-side third parties generally identified that how closely their metrics reflect outcomes that the market will observe outside the experimental context depends on similar factors to those set out in the discussion of external validity above (paragraphs 3.13 to 3.18).

4.47 Overall, experiment results from buy-side third parties showed that replacing TPCs with the use of Privacy Sandbox tools was associated with materially worse market outcomes compared to those in the status quo. Further, the results showed that some individual third parties (particularly, those with retargeting propositions) experienced even greater degradation in their offerings than Google, though others would be comparably affected. The same experiment results showed that PA API helped mitigate the impact of removing TPCs, though Topics appeared to be of low value and the utility of ARA appeared dependent on the level of noise added.

## Sell-side third parties

4.48 In this section, we also summarise the results from the 12 sell-side third parties who submitted experiments. These represent 10 results associated with sell-side ad tech businesses (ad tech; Criteo; Johnson, Gu and Kobayashi in partnership with a sell-side ad tech (**GJK**); Raptive; Index Exchange; Magnite; Nexxen; OpenX; Pubmatic; and UMT) and 4 publishers (publisher; Axel Springer, Dotdash Meredith, and Wirtualna Polska).[125] We analysed sell-side testers collectively, given similar metrics and testing approaches.

[124] For example, an ad tech found that the default approach during the testing period favours larger advertisers, which receive many conversions in a short period of time, and are thus less impacted by the reporting noise.
[125] One sell-side tester (publisher) provided its report in a format which meant we could not draw data which was aggregable with other testing reports. This tester provided conclusions and qualitative analysis consistent with others.

***Methodology***

4.49 Third-party experiments followed broadly similar methodological approaches to each other. SSP reports for these reasons generally reported on similar variables, focussing on revenue or impression-based metrics.

4.50 Our assessment of relative risks of bias between Google and third-party sell-side results has similarities to the buy-side assessment set out in paragraph 4.31. This is because buy-side testers' approaches were highly influential of the outcomes observed by the sell-side. However, there were some sell-side specific issues identified below. Overall, relative to Google's sell-side results, third-party sell-side results were:

*(a)* Equally subject to biases from sample attrition: Sell-side ad tech third-party testers engaging with us confirmed that they forwarded labels, a key source of sample attrition in demand side results. However, there remains potential for indirect impacts of other SSPs not forwarding labels, given buy-side testers often participated in auctions across several SSPs. However, this indirect risk appears lower than the direct risks to the buy-side tests, likely comparable with the risk to Google's results from attrition. Further, several sell-side submissions (including publishers and GJK) were in a position to observe total revenue impacts. GJK also accounted for the effects of latency reducing the number of impressions when revenue is measured – a form of survivorship bias.[126]

*(b)* Equally subject to challenges interpreting outcome metrics: Sell-side testers generally saw higher aggregate sample sizes than third-party demand side testers, relative to the frequency of the outcome measures they were seeking to observe. This resulted in measurement precision comparable to Google's tests. Third-party sell-side testers reported very similar metrics to Google's sell-side results.

*(c)* More subject to biases related to displacement outside the treatment group: Third-party experiments were more directly subject to bias arising from how demand side partners approached testing, than Google's experiments which occurred when few third parties were testing. In particular, some third-party demand side experiments sought to hold spend constant across groups. Whilst infrequently followed in the reports we reviewed, this strategy could have been used by other demand side

[126] See Gu, Johnson, Kobayashi, analysis presented in 'Privacy Sandbox: Impact on Publisher Revenue & Advertiser Conversions', 25 September 2024, 30.56-33.00 mins (accessed on 12 June 2025); An ad tech also provided data robust to the attrition through latency by providing revenue per request.

participants we see in SSP reports but from which we received no results.[127]

*(d)* On other factors, essentially similar to the buy-side experiments set out in paragraph 4.31 (subject to different biases arising from imbalanced participation across experimental groups; equally subject to biases from TPC signal leakage; and equally subject to different approaches to signal loss).

4.51 Different third-party sell-side experiments were more similar than buy-side results in the extent to which individual metrics were subject to the above risks. This is because third-party sell-side experiments followed much more similar methodologies and reported fewer metrics. Despite some specific issues, we considered the picture presented by third-party SSP results more certain than the picture presented by third-party buy-side results. We therefore place consistently high weight across the third-party sell-side experiments.

4.52 Across all types of sell-side businesses, results rarely differentiated experiments between the different APIs, focussing on evaluations of PA API separately from other APIs. Four testers (ad techs) reported data evaluating Topics API. One tester (publisher) reported data evaluating ARA.

***Results***

4.53 Third-party results provided a valuable sense-check on Google's figures. Further, given their relatively similar methodologies, third-party experiments provided particularly valuable insights into impacts across firms with different business models and working with different demand side partners. Sell-side testers reported broadly similar metrics to Google, focussed on publisher revenue, impressions, and revenue per impression.

*(a)* **Revenue per impression:** Ten testers (eight ad techs, one academic, one publisher) provided comparable datapoints.[128] All saw decreases

---

[127] The risk of bias to third-party results from this source also applies to other buy-side testers, but more indirectly. We therefore consider this a more material risk for third-party sell-side results than buy-side results. Consistent with this, an ad tech provided outcomes scoped to specific DSPs. These results showed highly varying outcomes across buy-side partners which the ad tech considered likely in part to be attributable to their testing strategy.

[128] See Gu, Zhengrong; Johnson, Garrett; Kobayashi, Shunto analysis presented in 'Privacy Sandbox: Impact on Publisher Revenue & Advertiser Conversions', 25 September 2024, 30.56-33.00 mins (accessed on 12 June 2025); Index Exchange, Insights from Our Privacy Sandbox Testing, 2 July 2024 (accessed on 12 June 2025).Statistical significance was not reported for many of the above results, though where the reports did

when using PA API in place of TPCs. On traffic replacing TPCs with the Privacy Sandbox tools, the median tester saw reduced publisher revenues of 30%. Results ranged between around a 20% reduction to around a 50% reduction. Google's results fell at or below the low end of the range, observing a 27% and 15% decrease across GAM and AdSense ad serving platforms respectively.[129]

*(b)* **Impressions**: Four testers (three ad techs, one academic) provided comparable datapoints. All four experiments found that using PA API in place of TPC signals reduced impressions, by magnitudes of 3%, 14%, 34% and 52% respectively.[130] The results were generally more negative than those observed by Google, which saw an 8% increase and 4% decrease across GAM and AdSense ad serving platforms respectively.[131]

*(c)* **Publisher revenue**: Three testers (one publisher, two ad techs) provided comparable datapoints. These experiments respectively found that using PA API in place of TPC signals reduced publisher revenues generated by ad-techs by 43%, 58% and 74% respectively.[132] These results are more negative than those observed by Google, which saw 21% and 18% falls in publisher revenues across GAM and AdSense ad serving platforms respectively.[133]

4.54 Third-party reports provided further insight into other datapoints of more indirect relevance to advertiser and publisher outcomes. We report those where multiple comparators existed, though Google did not include similar data in its testing report.[134]

---

contain the information, the datapoints were precisely estimated. The CMA had regard to the individual breakdowns provided, but in its summary analysis relied on overall weighted averages across these breakdowns calculated from the raw data provided (where third parties did not report overall figures directly).

[129] See Google, Testing privacy preserving signals in the absence of third-party cookies on Google's display ads platforms, 22 July 2024, page 16 (accessed on 12 June 2025).

[130] Statistical significance of these results was not reported, though an ad tech reported relatively small standard errors around the raw values for treatment and control groups.

[131] See Google, Testing privacy preserving signals in the absence of third-party cookies on Google's display ads platforms, 22 July 2024, page 16 (accessed on 12 June 2025).

[132] Statistical significance of these results was not reported.

[133] See Google, Testing privacy preserving signals in the absence of third-party cookies on Google's display ads platforms, 22 July 2024, page 15 (accessed on 12 June 2025).

[134] Some supply side third parties also provided analysis of how Google ad techs' share of impressions or bid requests varied across the groups. These were highly variable: for example, a publisher found that AdX's share of voice increased by 11%, whereas an ad tech found that DV360 reduced its share of impressions substantially. We placed low weight on these figures due to the uncertainty around DSPs' testing strategies with regard to levelled spend, as explained above.

*(a)* **Bid rate**: Three testers (ad techs) provided comparable datapoints. All three experiments found that using PA API in place of TPC signals reduced bid rates by magnitudes of 20%, 29%, and 68%.[135]

*(b)* **Viewability**: Two testers (publisher, trade association) provided comparable datapoints. Both experiments found that using PA API in place of TPC signals reduced viewability, by magnitudes of around 8% and 13% respectively.

*(c)* **Latency**: A number of testers provided comparable datapoints. For example, one ad-tech observed increased latency of 28% for the average impression.[136] An academic (GJK) also saw increased latency, and that the subset of treatment group impressions actually auctioned by PA API saw even higher proportional latency impacts.[137]

4.55 In addition to comparisons with the status quo, metrics provided both by third parties and Google allow calculation of how far value lost by restricting TPCs could be recovered using PA API and alternative identifiers.

*(a)* **Recovery rates for revenue per impression**: GJK explicitly calculated recovery rates for revenue per impression,[138] highlighting that PA API brings around 3% of the value that publishers would otherwise receive per impression.[139] This is lower than recovery rates consistent with Google's data of around 11% for GAM and AdSense ad serving platforms.[140] One reason for this could be that GJK adjusted for survivorship bias associated with impressions which do not load due to latency. Without this adjustment, estimates of recovery rates were comparable to Google's.[141]

---

[135] Statistical significance of these results was not reported.

[136] See Index Exchange, Insights from Our Privacy Sandbox Testing, 2 July 2024 (accessed on 12 June 2025). Statistical significance was not reported for this result.

[137] See Gu, Johnson, Kobayashi, analysis presented in 'Privacy Sandbox: Impact on Publisher Revenue & Advertiser Conversions', 25 September 2024, 28.10-30.47mins (accessed on 12 June 2025). Statistical significance was not reported for this result.

[138] Other testers provided data which could allow recovery rates to be calculated, however these were not headline metrics for the relevant testers and were in some cases affected by other experiments conducted on the Control 2 group. Implications for publisher revenue were highly varied and sometimes negative.

[139] See Gu, Johnson, Kobayashi, analysis presented in 'Privacy Sandbox: Impact on Publisher Revenue & Advertiser Conversions', 25 September 2024, 30.56-33.00 mins (accessed 12 June 2025). Statistical significance of this result was not reported.

[140] See Google, Testing privacy preserving signals in the absence of third-party cookies on Google's display ads platforms, 22 July 2024, page 16 (accessed on 12 June 2025).

[141] Other reports provided information that allowed calculation of recovery rates for revenue per impression. However, these reports did not interpret the data in these terms. Implied recovery rates were highly variable and sometimes negative.

*(b)* **Value of alternative mitigations**: One tester (publisher) compared subgroups of the treatment group where it enabled alternative identifiers and those where it did not. Whilst potentially limited by small sample sizes, the third party saw that use of alternative identifiers reduced the revenue impact of using PA API in place of TPCs. The impact varied across different geographies, of which the median was a 9 percentage point improvement in revenue, relative to testing absent alternative identifiers.[142] This indicates that alternative identifiers where available may support outcomes absent TPCs. Google did not provide a comparator in its testing report.

4.56 Four sell-side third parties reported data on Topics. This data generally suggested Topics added some small value, though the evidence was overall more uncertain than the evidence on PA API above.

*(a)* Two testers (ad techs) reported comparable evidence on the impact of using Topics in place of TPCs. One of the testers saw revenue per impression fall by more than half, the other tester saw a fall of around 25%.

*(b)* Two testers (ad techs) reported comparable evidence on the impact of using Topics in place of contextual targeting - and therefore calculated recovery rates for revenue per impression using Topics alone. Both quantified increases of around 4% when Topics was present, though this evidence is somewhat limited because bid requests with Topics could be higher value than those without. [143]

*(c)* Three testers (ad techs) reported comparable evidence on the prevalence of Topics. They found that Topics was present on a minority of bid requests. This is likely to inhibit its utility at the time of testing.[144]

4.57 One sell-side third party (publisher) reported data on ARA. This business found that ARA missed over 45-55% of conversions measured through other technologies.

4.58 Commentary in testing reports submitted by sell-side third parties generally identified that how closely their metrics reflect outcomes that the market will observe outside the experimental context depends on similar factors to those set out in the discussion of external validity above (paragraphs 3.13 to 3.18).

---

[142] Statistical significance was not reported for these results.
[143] Statistical significance was not reported for these results.
[144] See Index Exchange, Insights from Our Privacy Sandbox Testing, 2 July 2024 (accessed 12 June 2025)Statistical significance was not reported for these results.

For example, one party (ad tech) submitted that '*as of now, the PAA landscape is less competitive and hence, DSPs can [bid] lower to get those impressions. As more PAA demand gets added, we expect this to change'*. Several third parties (for example, a publisher) highlighted that particular issues or bugs with the Privacy Sandbox tools over the testing period worsened measured outcomes, and that their effectiveness depends on uptake.

4.59 Overall, experiment results from sell-side third parties showed that removing TPCs was associated with materially worse outcomes compared to those in the status quo. Further, the results showed that some individual third parties would experience a greater degradation in their offerings than Google, though others would be comparably affected. The same experiment results showed that the Privacy Sandbox tools and other identifiers helped mitigate the impact of removing TPCs, albeit to a limited extent at the time of testing.

# Annex 1: Internal validity discussion

1. In the section titled 'interpretation of experimental design' (paragraphs 3.1 to 3.21 above), we summarised the most important considerations for our overall judgements on the internal and external validity of testing results we received, taken in the round. In this annex we provide a fuller overview.

## Imbalances in participation

2. In A/B testing, random assignment of observational units between treatment and control groups is intended to ensure outcomes across each group are statistically independent of the assignment to a particular group. We discuss self-selection issues below that governed how closely or otherwise outcomes were measured over groups meeting this ideal. [145]

3. **Randomisation of experimental traffic**: Inference from A/B testing requires a sufficiently robust randomisation procedure to ensure the treatment and control groups contain a representative sample.[146] Google provided analysis of the extent to which each group was similar on observables to the other and to the characteristics of that group immediately prior to the experimentation period. Google submitted an analysis showing that the testing groups were generally balanced across a range of observable characteristics, such as the proportion of clients in the UK and median time spent browsing per day. Whilst there were some differences, these did not appear to have economic significance. This analysis was based on the period 29 January to 10 March 2024, with reference to a pre-experimentation period 6 December 2023 to 12 December 2024. The data related to Google's own open display experiments, but represents a starting point for assessing third-party experiments relying on the same labels. We consider how the sample may have changed further during the period of third-party testing below.

4. **Auction pressures in control and treatment group:** As control traffic could be won without any investment in the Privacy Sandbox tools but the treatment group required such investment, imbalances arose in participation of ad techs

---

[145] Google sought to address low participation and its impacts by providing testing grants. We considered whether Google's grants to test participants led to a testing environment not representative of market incentives. We considered this risk small, given the size of the grant did not vary with particular actions and therefore would not have affected incentives at the margin.

[146] Google's A/B testing setup also involved removing certain user groups from both treatment and control for pragmatic reasons. Most exclusions seemed innocuous. We noted however that Google excluded all users who have already restricted TPCs from the experiment. However, the estimated impact of moving from TPCs to Privacy Sandbox tools would not apply to these users (as they have already removed TPCs). Therefore, extrapolating impacts to the wider population would risk overstating the impact of restricting TPCs.

across treatment and control group auctions. Reduced buy-side participation could have led to lower auction pressure, reducing publisher revenue per impression. [147] On the other hand, if SSPs had provided less supply of PA API impressions, this could have offset the reduction in demand.[148] The overall per impression (ie price) impacts were ambiguous. However, reduced supply entering the experiment would have resulted in fewer auctions and therefore lower overall revenue – even where per impression figures were similar. The sign of any bias to conversion rates and other metrics from this source is not clear, and stakeholders did not make submissions in this respect.

5. **Self-selection into the experiment**: Businesses who chose to participate in the experiment and share their results with us cannot be guaranteed to be representative of the wider population, and may overall capture those businesses with greater need to adapt. As set out above, ad techs self-selected into the experiment and this may well have led to the mix of ad techs in the treatment group auctions differing from those in the control. There may also have been self-selection of publishers and advertisers. Our guidance advised testers for example not to allow advertisers to choose directly whether or not to participate in the treatment or control arms.[149] Some testers conducted their tests across a wide range of advertisers, whereas others faced practical constraints in doing so and therefore presented results from a small pool of self-selecting advertisers. Third parties also needed to actively integrate publishers into some Privacy Sandbox tool testing across their inventories. Ad tech testers typically did not provide a full analysis of whether the treatment and control arms were balanced according to publisher and advertiser participation, which limited our ability to fully evaluate these issues.

6. **Conversion measurement chains**: Low participation may cause ARA to miss some conversions. This implies conversion measurement in the treatment group of the experiment (and ARA-specific results) may be expected to understate the effectiveness of ARA, assuming there would be higher participation levels were Google to introduce wider conduct restricting TPCs.[150]

---

[147] See, for example, Google, Testing privacy preserving signals in the absence of third-party cookies on Google's display ads platforms, 22 July 2024, page 7 (accessed on 12 June 2025).
[148] Further, insofar as impressions were less expensive, the buy-side may have purchased more impressions for a given budget.
[149] See the CMA's Additional CMA guidance to third parties on testing, paragraphs 23-25.
[150] For a conversion event to be tracked by ARA after a user sees an ad, the advertiser/DSP needs to update their impression and click tags to call the ARA, and the publisher/SSP needs to update the webpage to allow ARA to be called. To the extent that not all parties in the chain are testing the Privacy Sandbox tools, the conversion will not be recorded in the treatment group while it will be recorded in the control group. Changes may also be needed on the website where the conversion occurs.

7. Overall, differing participation between the treatment and control groups appears to be a particularly significant source of bias to estimates. For the sell-side, this is likely to have deflated publisher revenue. For the buy-side, it may have reduced spend per impression but increased the number of impressions bought, leading to more mixed impacts on advertiser utility.

## Imbalanced sample attrition

8. Further risks to the extent to which the control group represents a valid comparator to the treatment group can arise through certain types of traffic systematically entering or leaving the experiment after selection.[151]

9. **Sample attrition due to labelling**: Some buy-side testers identified that a material fraction of labels from certain sell-side firms were not forwarded, leading to attrition of the sample. This attrition was greater in the control rather than treatment groups, inflating publisher revenue in the treatment group. One third party identified that not accounting for sample attrition could have materially biased results.[152] We placed more weight on testing results which addressed this, and on sell-side results which were unaffected.[153]

10. **Invalid traffic**: Several testers identified that the sizes of the experimental groups diverged materially from their expected values due to the presence of invalid traffic, with the 'Treatment 1.1' group particularly affected.[154] Google told us that GAM's approach was to filter this traffic out of its experiment, an approach shared by some third parties.[155]. Some testers reported that this divergence primarily stemmed from invalid traffic entering the experiment. Testers were generally able to address this via normalisation (subject to the concern below).

---

[151] We also considered one factor which remained constant over time under this heading. Differences in experimental group size: Google's experiment provided a treatment group composed of 0.75% of stable Chrome traffic, whereas the control 1 group was composed of 1% of traffic and control 2 group 0.25%. When making comparisons between groups of different sizes on an aggregate (rather than per impression) basis, testers generally reported that they regularised sizes of groups in accordance with their theoretical values (and none reported they made comparisons without regularising these values).

[152] Criteo estimated that publisher revenue could appear inflated by around 50% in the treatment group relative to Control 1. See Selman, Criteo's Privacy Sandbox Market Testing, 27 June 2024 (accessed on 12 June 2025).

[153] This was feasible by accessing the labels via other mechanisms to observe (close to) the full sample. See, for example, Kobayashi, Johnson and Gu, Privacy-Enhanced versus Traditional Retargeting: Ad Effectiveness in an Industry-Wide Field Experiment, 30 September 2024, page 8 (accessed on 12 June 2025).

[154] Further, insofar as this traffic stemmed from 'bots' which did not bid on ads, then spend and conversion metrics would be unaffected. See, for example, Kobayashi, Johnson and Gu, Privacy-Enhanced versus Traditional Retargeting: Ad Effectiveness in an Industry-Wide Field Experiment, 30 September 2024, page 11 (accessed on 12 June 2025).

[155] See, for example, Google's Q1 2024 progress report, page 28.

11. **Survivorship bias due to latency**: As set out in our quarterly reports, during the experimental period latency in Privacy Sandbox tools was material. This implied that certain impressions in the treatment group would not be counted (because certain page views would not lead to ad views). Results normalised by the observed number of impressions would therefore understate treatment group impacts, because they would not account for the negative advertising outcomes associated with such impressions 'dropping out' of the experiment.[156] We placed more weight on testing results whose normalisations addressed this concern.

12. **Bugs:** Certain other issues such as bugs with the Privacy Sandbox functionality affected the composition or functionalities of treatment and control groups asymmetrically. These were typically small in scale and/or duration.[157]

13. Overall, several of these factors (eg sample attrition) appear to have been particularly important in this experiment: we placed most weight on testing reports which were able to solve these challenges given limited resources. Other issues are unlikely to have led to material bias in the tests' internal validity.

## Experimental contamination

14. In A/B testing, the potential outcomes for any unit should not vary with the treatments assigned to other units in the experiment. One way this can happen is if the treatment is not limited to the treatment group. This implies here that TPC signals in the control group should not affect the treatment group.[158] There were however various channels through which TPC leakage and other mechanisms could have led to experimental contamination.

15. **Choice of 'diversion unit'**. Avoiding all interactions between the multiple different groups was extremely challenging. We consider Google's decision to construct the treatment group by randomising browsers (rather than other units, such as advertisers, publishers or users where signed-in) between groups was pragmatic and appropriate. However, this choice did not remove

---

[156] See Johnson, Privacy Sandbox: Impact on Publisher Revenue & Advertiser Conversions (accessed on 12 June 2025).
[157] See also the table below paragraph 102 in our April 2024 update report, which identified actions Google had to take to experiment traffic to mitigate a bug in bounce-tracking mitigation. See the CMA's Q1 2024 update report, page 72.
[158] It would further imply that the Privacy Sandbox tools should not affect the control group. Such impacts were possible – for example by diverting engineering efforts or by providing valid signals which could be used to refine eg measurement of control group conversions. However, any such impacts would likely be small in scale.

the potential for these interactions. For example, conversions could leak between groups if users of multiple devices had these devices assigned to different groups, such as a mobile user viewing ads targeted by treatment group technology but then having the conversion measured only using a control group desktop browser. Other potential leakages exist in relation to advertisers, ad tech and publisher actions as set out below.

16. **TPC signals present in experimental design 1 bid requests**: As set out above, some browsers labelled as part of the treatment group retained TPC information, where these were included in experimental design 1. Ad techs who did not participate in the experiment would likely have used the TPC information in their bid requests, affecting auction prices for ad techs who were part of the experiment. As per our guidance, we placed greatest weight on experimental design 2 tests.[159]

17. **TPC signals in experimental design 2 bid requests**: TPC signals were sometimes present in treatment group traffic (and control 2) even where suppressed by Chrome. This occurred where sellers reinstated TPC-derived information in the bid request through 'ID bridging'. Where testers evaluated the impact of ID bridging, this increased our ability to evaluate their results.

18. **TPC signals in bidding models**: Given the limited data, time and resources available for model training, models relying on TPC information were typically deployed on treatment group traffic (in either experimental design). Therefore, a specific ad request with no TPC data itself could still 'piggy-back' on inferences derived from historic TPCs. Some testers sought to address model contamination, although their ability to fully address this issue was limited.[160]

19. **Visible testing**: In some experimental contexts, participant behaviour may change where they know how they are allocated between treatment and control. Experimental designs sometimes employ 'masking' to avoid such risks. In this context, that would mean making participants unaware which browsers are associated with the control or treatment groups. Our discussions with Google identified this was not realistic.[161] Further, Google also submitted

---

[159] See the CMA's additional CMA guidance to third parties on testing, paragraphs 5-7.

[160] For example, an ad tech stated that it is impractical to train models exclusively on the experimental data due to low traffic volumes, but it attempted to emulate certain constraints associated with the absence of cross-domain identifiers. The ad tech cautioned this would not fully replicate the conditions in the event of TPC deprecation.

[161] Google identified that avoiding any visibility of the treatment would not be possible. Ad requests with TPCs could not be part of the treatment group; ad requests without TPCs would have a strong chance of being in the treatment group (although could also arise where users opted out of TPCs). Google also identified that it had to communicate with users in cases such as breakage mitigation (where users in the treatment group experienced a

an analysis which showed that its experiment did not generally affect how users browse with Chrome across a range of observable characteristics, such as the numbers of pages loaded. Testing reports did not typically contain equivalent data to let us evaluate the representativeness of advertisers, ad tech or publisher actions based on observables – although reports also did not indicate behaviour that would clearly place pressure on this assumption.

20. Generally, such contamination would reduce the observed differences between treatment and control groups. The potential for such influence reduced the extent to which we treat the experiment as providing truly causal estimates.

## Displacement effects

21. In A/B testing, the potential outcomes for any unit should not vary with the treatments assigned to other units outside of the experiment. This can be invalidated where treatment 'spills out' of the experiment. Risks arose in this context because the experiment covered only a small fraction of traffic, and because advertisers could choose to move activity out of the experiment. We explain how these risks arose and challenges with mitigating them below.

22. Advertisers allocate budgets to maximise their return on overall ad spend. In doing so, advertisers allocate spend to channels based on performance against their objectives (such as products sales) and their prices. Restrictions to TPC availability will affect functionalities associated with user tracking and therefore performance, particularly for more TPC-dependent business models. In equilibrium, advertisers would reallocate budget away from more affected channels, implying both spend and budget will fall.[162] In the context of the experiment, this might translate to advertisers reallocating budget from the treatment to the control group. Budget shifting of this nature would cause spend in both treatment *and control* groups to be affected by TPC restrictions only in the treatment group, invalidating inference. Even more likely, given the small scale of the experimental sample relative to all open display traffic,

---

site without TPCs, Google had to re-enable TPCs in Chrome). Although a portion of traffic today does not carry TPCs (such as users in Incognito mode or those who have opted out of TPCs), there is no reason to assume these users are representative of the wider population.

[162] Depending on the magnitude of relative price and performance reductions – and how advertisers' weight these. Insofar as bidding models adjust during the experiment, lower performance in affected open display channels would reduce bid valuations and therefore prices of affected channels. This would provide a 'rebound' effect mitigating against budget shifting. The rebound effect is unlikely sufficiently large to compensate publishers or achieve advertiser objectives.

budget could be displaced from experimental traffic to unaffected open display traffic.[163]

23. Addressing these issues presented challenges. Modifications to current bidding models would be needed to avoid these interactions, which might be impractical and reduce the relevance of the experiment.[164] Buy-side testing also needed to be compatible with varying types of advertiser constraints around campaign goals.[165] Our guidance recommended businesses take steps to avoid outcomes in the treatment impacting outcomes in the control groups and vice versa, including in relation to budget shifting. However, our guidance recognised that budget shifts may naturally arise, and that interpretation should consider the mechanisms through which this would occur.[166]

24. In practice, testers followed various approaches to these trade-offs, broadly falling into three groups: (i) some (eg an ad tech) enforced level spend at campaign level of pre-allocated budgets across treatment and control groups,[167] assessing the differing performance of browsers assigned to each group; (ii) others (eg an ad tech) targeted levelled performance of the campaign across browsers assigned to each group by varying campaign level spend across treatment and control groups;[168] and (iii) others (eg Google) allowed budgets to flow across treatment and control groups relying on their bidding models' ability to equilibrate spend and performance trade-offs between these groups.[169]

25. More active approaches to addressing these issues, such as levelling spend or performance in the treatment group with that in the control group, may have improved experimental validity. The idea is to preserve the relative difference

---

[163] Google told us that treatment affects which spillover to the control would be small in the context of spillovers also flowing outside the experiment – given the relatively small share of the control amongst wider treatment. Further that any such budget shifting should not lead to any particular directional bias to the experiment.
[164] See the CMA's additional CMA guidance to third parties on testing, paragraph 25(c).
[165] Google also told us that in the context of its own ad platforms, advertisers choose different types of spend limit, and that sometimes spend limits can be increased whilst campaigns are in progress. Advertisers are able to specify goals such as target cost per action in campaigns.
[166] See the CMA's additional CMA guidance to third parties on testing, paragraph 25(c).
[167] Testers following such levelling approaches sought to match control and treatment characteristics, although in doing so would not have managed budget shifting outside the experiment (and may have led control group spend or performance to deviate from a true status quo representation).
[168] See Selman, Criteo's Privacy Sandbox Market Testing (accessed on 12 June 2025).
[169] Testers used various measures of performance and spend to effect such strategies in practice. For example, one ad tech targeted levelling impressions across treatment and control groups, and another ad tech used cost per qualified visit as a measure of performance.

in outcomes of interest observed in the treatment and control groups. For example:

*(a)* In the case of levelled spend, any budget flow to non-experimental traffic would be equal across control and treatment groups, implying equal distortive effect on performance, and preserved relative comparisons between treatment and control groups.

*(b)* In the case of levelled performance, where this was a proxy for return on ad spend, cheaper impressions in the treatment group would have led to net budget flows into the control group. Insofar as the levelled performance measure (working across the two groups) fell below the performance measure outside of the experiment, budget would still have diverted outside of the experiment. However, levelling the performance would have implied broadly equal distortion to spend and preserved relative comparisons between treatment and control groups.

26. However, these approaches also had downsides. By levelling all displacement, testers would have also controlled displacement *outside* of the open display ecosystem which is a relevant outcome which should in principle be captured by the experiment. For example, levelled spend displacement effects outside the experiment would likely lead to 'too much' displacement from the control group (given such traffic would have experienced low performance degradation relative to O&O for example) and fewer destinations following wider TPC restrictions, and 'too little' displacement in the treatment group (given the more significant performance degradation relative to O&O). Further, there were risks in distorting overall outcomes – for example, if testers sought to level performance across very imprecisely measured outcomes such as conversions, testers could inadvertently have increased displacement.

27. Whether levelling spend/outcome metrics across experimental groups reduced overall bias, depends on the extent to which the strategy could be implemented effectively without introducing new biases, as well as the relative attractiveness of open display and non-experimental open display traffic during this period for advertisers and the actual propensity for advertisers to switch in the short-run experimental context. One tester (ad tech) submitted a stylised model which indicated that in the experimental context, the global performance impact of new budgets flowing in other channels was negligible, implying that controlling displacement between and outside of experimental groups would reduce bias (provided executed correctly).

28. In our view, it was reasonable for testers to undertake such levelling strategies, or to conclude that the risks of introducing new biases were too

great. Eliminating all bias was not possible without a fundamentally different and more invasive industry experiment. We consider that these issues reduce the extent to which we can consider testing results to be causal estimates, but do not invalidate considered interpretation from them. We therefore place weight on results even where testing strategies did not make adjustments for displacement.[170]

29. Where testers levelled particular outcomes, we focussed our interpretation on non-levelled metric. For example, for levelled performance results we focussed on spend metrics. We had regard to wider metrics to the extent they were not also held constant by the strategy.

## Different treatment levels

30. In A/B testing, there should be no different forms or versions of each treatment level which lead to different potential outcomes, and relatedly an unambiguous definition of treatment. This implies here that all bid requests in the treatment group should be subject to the same loss of signal that bid requests will experience following the implementation of wider TPC restrictions, or at least that all bidding should be undertaken in conditions approximating the same understanding of the level of signal available. Testers followed different approaches to testing the Privacy Sandbox tools, leading to different extents of signal degradation across bids in auctions than would arise following the implementation of wider TPC restrictions.[171]

31. **Differences in APIs tested**: Some testers used only PA API for tracking whereas others also used alternative signals such as Topics. Whilst differences in API usage across participants could arise with wider TPC restrictions and therefore be reflective of the true treatment of bidding behaviour, it is likely that differential testing strategies introduced some additional noise around this variation which could impact results. This noise is limited insofar as differences in APIs used contribute to low levels of outcome variation.

32. **Differences in mitigations**: Some testers relied on more extensive mitigations to signal loss (such as use of commercial ID solutions). These differences in approach in part reflected testing strategy and uncertainty of

---

[170] We place however correspondingly lower weight where testers sought to follow a particular strategy but there was some evidence this was thwarted by other conditions. In such cases, the interpretation of the results becomes less clear.

[171] Differential use of mitigations likely reflected to some extent different views around the end-state effectiveness of these identifiers and capability to use them at the time of testing, and in part due to differing approaches to risk relating to Google's potential to restrict the availability of alternative identifiers in future.

what signals will be available in the long run. Therefore, the differences in approach influencing the level of signal degradation in bids during the experiment may, for at least some testers, less closely reflect differences we would expect for browsers without TPCs. This implies that certain bids would have experienced variation in treatment levels which must be taken into account in interpreting the data in the round.[172]

33. Whilst testers broadly used the same approach within their experiments, outcomes were determined through auctions between testers using different approaches. This implies a degree of 'spillover' between experiments, and variation in the treatment received by individual bids. Assessing the direction of such biases is not feasible. Based on third-party views about the relative utility of signals that differed across testers' implementation of 'the treatment', and in particular the relatively low utility of Topics, we expect the practical significance to be low.

## Measuring outcome metrics

34. A/B testing is a framework capable of measuring multiple outcomes of interest. However, measurements of these outcomes need to account for the experimental context. We took into account several contextual points in evaluating results.

35. **Multiple outcomes of interest**: Testers typically reported a range of outcomes to cover different objectives, which allowed us to assess impacts across different dimensions which matter to businesses in this sector. Testers sometimes used different terminology for similar metrics. As set out above, testers' strategies (such as holding some outcome metrics constant across groups and including different mitigations across groups) also prevented comparisons on metrics where titled the same. We sought to make appropriately nuanced comparisons across the range of data we received, accounting for expected differences.

36. **Different definitions of conversions**: There are further challenges for conversion metrics, where different ad techs use different measurement approaches, with measurement sometimes differing even within the same ad tech (due to advertisers valuing different actions).[173] Google provided test statistics which addressed challenges with interpreting average conversion

[172] Testers also developed their bidding models to different extents to adapt to signal loss, with similar impacts.
[173] Some advertisers may define conversions as visits to particular pages, others adding a product to a shopping basket, and still others as purchasing a particular product. Some reported status quo conversions using ARA in line with our testing guidance, whereas others such as Google used technologies such as TPCs. Further, choices can vary over time; prospects of conversions vary significantly across products.

rates: Google used Mantel Haenszel (MH) ratios which help compare like-with-like across control and treatment groups. Third parties typically caveated their results where conversion rates were submitted. We placed more weight on the former approach.

37. **Imprecise estimates**: Experimental groups in the Chrome-facilitated testing involved only a small fraction of traffic, as set out in paragraph 2.3 to 2.6. This risked testers being unable to measure rare outcomes (such as certain conversion types) with much precision. This risk particularly applied to third parties, in some cases creating uncertainty around the measured directionality of the results. Testers often removed extreme values in their results which might unrepresentatively skew a small sample.

38. **Estimating confidence intervals**: There were conceptual challenges computing confidence intervals, for example due to testing units not being fully independent and several outcomes being determined simultaneously. Testers took a variety of approaches to estimating standard errors, particularly in the treatment groups, and typically provided a range of caveats to these estimates.[174] Across all testing reports, we treated the precise estimates of standard errors with some caution.

39. These issues did not generally affect the validity of testing that we received, but informed the firmness of any comparisons we made between testing results and militated against placing precise weight on any individual point estimate.

[174] For example, Google stated that its methodology could lead standard errors to be slightly overestimated without TPCs, but that this effect should not materially impact on the statistical validity of its experimental results.

# Annex 2: Glossary

**Control Group**: Impressions served removing data related to new Privacy Sandbox APIs). There were two control groups: Control 1 which represented the status quo (ie included the use of third-party cookies) and Control 2 which excluded the use of third-party cookies.

**Integration of PS tools:** The extent to which an ad tech or ad inventory supplier has made the necessary adjustments in its systems to adopt the Privacy Sandbox tools

**Market outcomes:** The results of the interaction between supply and demand in a market.

**Status Quo**: Availability of all signals which support functionality associated with user tracking today, and in particular third-party cookies.

**Privacy Sandbox Proposals:** A set of proposed changes on Chrome that aim to address privacy concerns by removing the cross-site tracking of Chrome users through TPCs and other methods of tracking; and create a set of alternative tools to provide the functionalities that are currently dependent on cross-site tracking.

**Privacy Sandbox Tools:** A range of alternative tools which provide some support to the functionalities associated with user tracking.

**Third-Party Cookie (TPC):** TPCs are set by a domain other than the one the user is visiting. This typically occurs when the website incorporates elements from other sites, such as images, social media plugins or advertising. When the browser or other software fetches these elements from the other sites, those other sites can set cookies as well.[175]

**Treatment Group:** Impressions served without using data related to third-party cookies.

[175] See Information Commissioner's Office, Cookies and similar technologies (accessed on 12 June 2025).