Research Framework for Acceptable Ad Experiences

Background

Modifying the Acceptable Ads Standard is determined by the Acceptable Ads Committee (AAC) – who, to date, have relied on commissioned research studies on ad blocking users' perceptions of ad experiences. Depending on the question, these research studies elicit self-reported responses from ad blocking users on mock-webpages that vary different ad experiences. While these studies have good internal validity and control, there are limitations, such as a lack of external validity or an understanding of how users react to ad experiences in practice. Further, a reliance on these types of studies: a) ignores valuable insights that can be derived from a broader range of methodologies that have the potential to offer deeper and/or more timely insights; b) can incur substantial time, resources and cost to conduct – often resulting in limitations to the set of conditions and implementations that can be tested, and c) do not make use of a range of user behavior data that is now available via the extension.

We propose a portfolio of Acceptable Ads Research Methods and Data Sources for review by the AAC. The purpose of this portfolio is twofold: 1) to identify a broad range of research methods that can be used to generate, test or validate user experience insights at various phases of ideation, and 2) to harness additional user experience and behavior data made available through the extension to generate more timely and real world insights on actual user experiences.

There are two considerations in this framework. First, determining which research methodology/ies are sufficient to address specific research questions (***research method justification***). Second, how to utilise and interpret *indirect* user behavior data to justify new or to change existing AA experiences (***utilising and interpreting user experience and behaviour data***).

The current document presents a framework for harnessing both user research and behavioral data to build confidence in and reach more timely decisions about acceptable ad experiences for ad blocking users.

AA Standard: Role of the User and User Data in the AAC

The Acceptable Ads (AA) Standard emphasises user experience as a core consideration for introducing new or modifying existing AA experiences. Indeed, users play a critical and central role in the AAC as well as at various phases in setting the AA Standard.

User representation on the AAC

Users and user advocates are represented as one of the three AAC coalitions, contributing to the decision-making and implementation process for proposed changes to the AA Standard:

- Ideation phase – the user advocate coalition can bring research ideas to the committee
- Discussion phase - the user advocate coalition are involved in discussions on proposed ideas or suggested changes brought to the committee
- Decision-making phase - the user advocate coalition can use their voting power to affect the proposed AA Standard change

User feedback in setting the AA Standard

In addition to representation, the AAC specifies that user feedback play a role in the decision-making and implementation process:

- Neutrally-derived user data is considered one basis on which suggestions relating to the AA Standard can be made (Ideation phase - user data requirement)
- Proposed changes to the AA Standard are posted online to elicit user feedback for one month prior to the AAC decision-making phase (User Feedback phase)
- More recently, the AAC has requested monitoring processes following a change to the AA Standard, making use of available user behavior data for a pre-specified time period to evaluate the effect of the change on: extension *uninstall rates*, *AA opt-outs*, submissions of *issue reports* and responses to *uninstall* and *AA opt-out surveys*.

There are a range of research methodologies as well as an increasing amount of user behavior data now available through the adblocking extensions that can provide insights into user tolerance and acceptability of ad experiences. **The aim of the current proposal is to broaden the use and interpretation of the user data requirement in the Ideation phase** of the AAC decision-making process as well as **ensure the AAC has the data it needs to monitor and evaluate the AA Standard** over time.

User Data Research Requirements

At present, the AAC requires that any user data presented to the committee in support of or to justify a change to the AA Standard should have a comparable methodology and scope to this study. The purpose of this requirement is to ensure that **any user research data brought to the AAC retains a certain level of methodological quality and a generalisable scope**.

Further, the AAC bylaws specify that **the largest criterion for determining the acceptability of an ad experience is the annoyance level of the ad experience to an ad blocking user**. The annoyance level criterion facilitates comparisons with prior AAC research while also serving as a clear metric derived directly from ad blocking users' experiences.

In summary, the current user data approach has sought to achieve two main aims: (1) ***maintain a level of methodological quality and rigor in AA Standard research*** and (2) ***embed direct user perceptions/feedback in the user data requirement.***

However, there are limitations to the current AAC user research data requirements. For instance, they:

- Focus on a narrow set of research methods with no clear path to incorporate user insights derived from a broader range of methodologies (e.g., qualitative interviews, user behavior data);
- Can incur substantial time, resources and cost to conduct – depending on the complexity of the proposal – often resulting in a limited set of conditions or implementations that can be tested in a single study and lengthy research timeframes and decision-making processes;
- Were established when capacity to collect a range of user behavior data via ad blocking extensions was limited; thus, they are currently outdated given recent advances in the collection of data on ad blocking users' behaviors online.

To overcome these limitations, **we propose an updated research framework that integrates a broader set of research methodologies and available user behavior data** while retaining methodological rigor and direct user feedback in its approach.

Research Framework

Research Method Justification

The first step in initiating research on the AA Standard should be to **determine which research methodology/ies are sufficient to address a specific research question/s**. The following presents a suggested process that a) streamlines the research process to incorporate a broader range of methodologies dependent on the specific research question, and b) ensures that the user perspective is retained in the process.

There are two types of user data: data on user perceptions and on user experience and behavior. Both of these data sources can provide insights into user acceptability of ad experiences:

- **User experience and behavior data** provides insights into what users actually **do** in response to ad experiences online and therefore provide a valuable source of information about user actions or responses to changes to the AA Standard. User behavior data is most appropriate when a

research idea is clearly defined and the intention is to test implementation or to generate hypotheses based on existing user behaviors.

- **User perceptions data** provides information on how users perceive their online browsing experience, their sentiments or preferences. This type of data has the potential to validate assumptions, identify pain points or opportunities as well as any factors that may be influential in interpreting patterns or trends from user behavior data. It provides a more direct assessment of how users feel about online advertising experiences that may not translate to immediate action but will nevertheless influence their experience online. User perceptions data can be collected at various stages of the research process.

Insights from user perception data incorporate the user perspective in the development of the AA Standard more directly than user behavior data.

Accordingly, **to ensure the perspective of the user is maintained in the decision-making process for the AA Standard, data on user perceptions should accompany any proposed change to the AA Standard**. User perception data can be used to complement user behavior data and can be sourced from a broad range of research methodologies at various phases of the research process.

Thus, research to inform the AA Standard has the following steps.

Step one: Select an appropriate research methodology given the research question and clarity of the idea or proposition (see Figure 1, decision tree).

Step two: Ensure user perception data is collected at a relevant phase of the research process, either prior to or in parallel to user behavior data (see Figure 1, user feedback requirement). This data will serve to generate, identify or refine an idea, or elicit direct user feedback on the idea and its implementation.

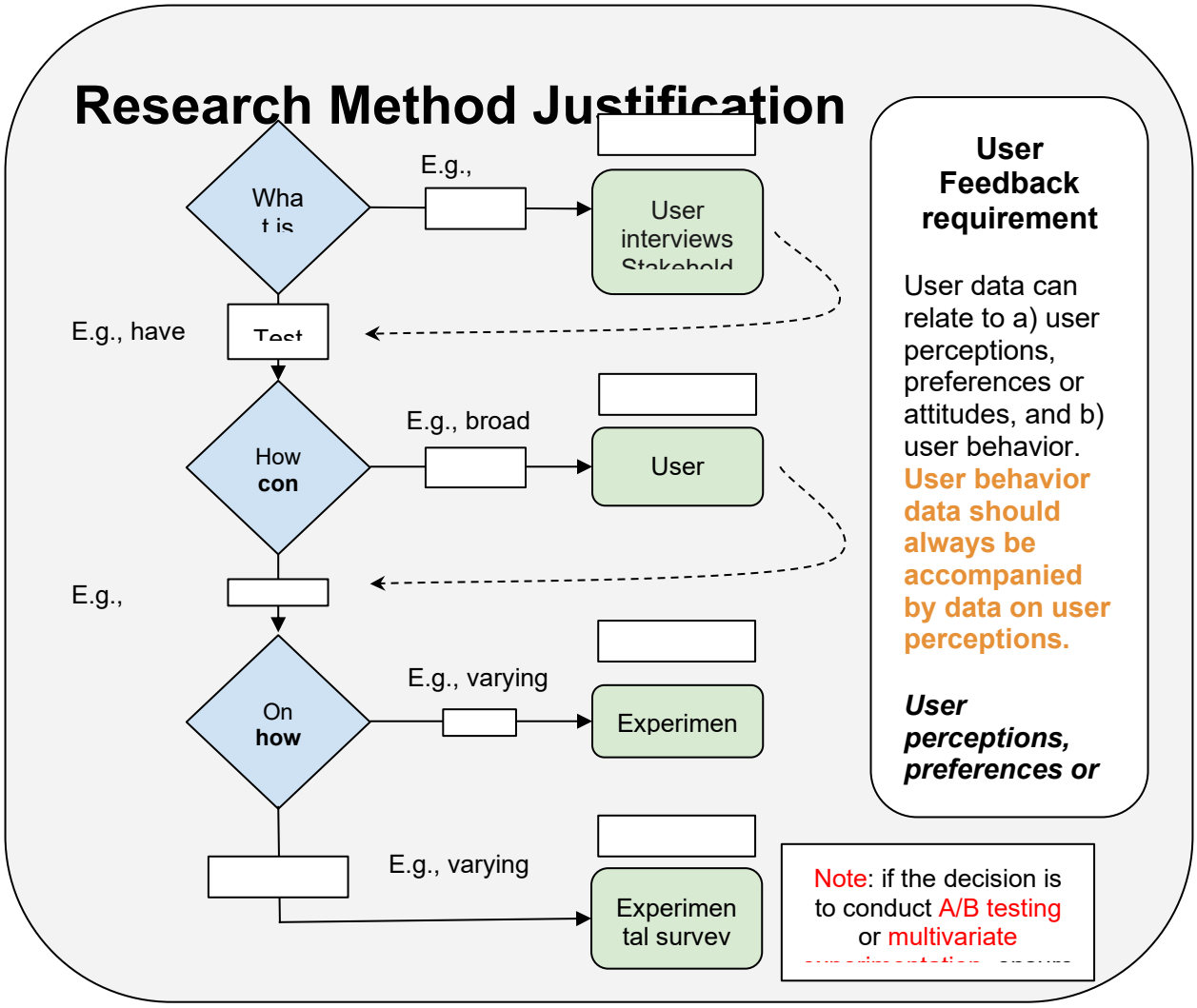

Figure 1. Decision tree identifying research methodologies given specific research objectives and identifying potential sources of user perception data.

To illustrate how a streamlined AA Standard research based on the framework could look like, Box 1 provides multiple possible research approaches to a question about cinemagraph ads. While Box 1 lists multiple approaches to address the same general research question, the selection of approach would be guided by the clarity or specificity of a question and the availability of or confidence that can be drawn from results from prior or related research.

Box 1. Research question: How acceptable do ad blocking users find cinemagraphs?

There is no existing user data on a relatively new ad format, the cinemagraph.

Research approach #1.

1. In-depth interviews with ad blocking users gain insights on disruptiveness and acceptability of these ad formats, along with ideas on where and when they might be acceptable.

Insight: ad blocking users do not seem bothered by these ad formats and even like them.

2. Iterative A/B testing with cinemagraph ads appearing (test #1) nowhere vs. banner -> (test #2) banner vs. banner and right rail.

Insight: A/B test #1 shows no negative effect on user experience when cinemagraph shown in banner; A/B test #2 shows negative effects with cinemagraphs in both the banner and right rail (relative to only in the banner).

Research approach #2.

1. Experimental survey with ad blocking users testing acceptability of a) cinemagraph vs. none and b) cinemagraph placement (none, banner, right rail, both).

Insight: Experimental survey shows users do not perceive a single cinemagraph ad on a page as disruptive to their experience but do find multiple cinemagraph ads disruptive.

2. (optional) User behavior data analytics evaluate the implementation of cinemagraphs on user experience.

Research approach #3.

1. Multivariate experiment with cinemagraph ads appearing in different placements on a page (none, banner, right rail, both) and with different frequencies (5%, 15% or 30% of webpages during a browsing session).

2. In parallel, in-product messaging survey sent to 10% of users within each of the multivariate experiment conditions evaluates user perceptions of their browsing experience on pages with cinemagraph ads.

Insights: Multivariate experiment data shows no negative effects of different cinemagraph placements except when frequency is 30%. IPM survey confirms results with direct user feedback, with the additional insight that users were most bothered by experiences with multiple ads changing on a page (e.g., cinemagraph + ad-refresh on the same page).

Research approach #4.

1. Multivariate experiment with cinemagraph ads appearing in different placements on a page (none, banner, right rail, both) and with different frequencies (5%, 15% or 30% of webpages during a browsing session).

Insight: Multivariate experiment data shows no negative effects of different cinemagraph placements except when frequency is 30% and placement is in the right rail.

2. In-depth interviews with ad blocking users to gain insights on why the right rail placement is most disruptive.

Insight: ad blocking users scroll past the banner ad fast enough for the cinemagraph not to bother them, and movement in the right rail ad distracts them from their browsing.

See also here for a worked example of a research process investigating how users react to animated ads.

Utilising and Interpreting User Experience and Behavior Data

Historically, limited and aggregate user data was collected through the extension. These included uninstall and AA opt-out rates, issue report submissions and responses to uninstall and AA opt-out surveys (if completed). While this data is continually monitored and shared with the AAC (e.g., to evaluate periods immediately after a change in the Acceptable Ads Standard, such as the roll-out of in-view ad-refresh), it offers only a limited view on user acceptability of their Acceptable Ads experience.

Recently, engineering teams at eyeo and Blockthrough have implemented a process to securely **collect additional user experience and behavior data related to their use of the extension and their browsing experience**. The data can be collected via the extension, in connection with data from Blockthrough, and/or via in-product messaging surveys. In combination with eyeo's enhanced capacity to conduct A/B and multivariate experimentation, this user experience and behavior data can be used to evaluate the effects of exposure to different ad experiences across test cohorts.

The collection of this user behavior data allows for the evaluation of user experience in real world online environments.

There are two uses for this data in relation to evaluating ad experiences in A/B tests and multivariate experiments: to ensure 1) treatment fidelity and 2) to evaluate user reactions to different ad experiences delivered to the test cohorts.

Data Reliability

The first step in interpreting outcomes from A/B testing or multivariate experimentation is to gather data to ensure a) control and experimental groups received (enough of) the intervention they were intended to (*treatment fidelity*) and b) that the groups did not differ markedly from one another in any other way that could affect results (*comparability*). User experience and behavior data can be used to increase confidence that the results of any tests are attributable to the ad experience being tested and not other factor/s (e.g., users in a cinemagraph ad experience group may have not received any in-view ad-refresh owing to an implementation issue).

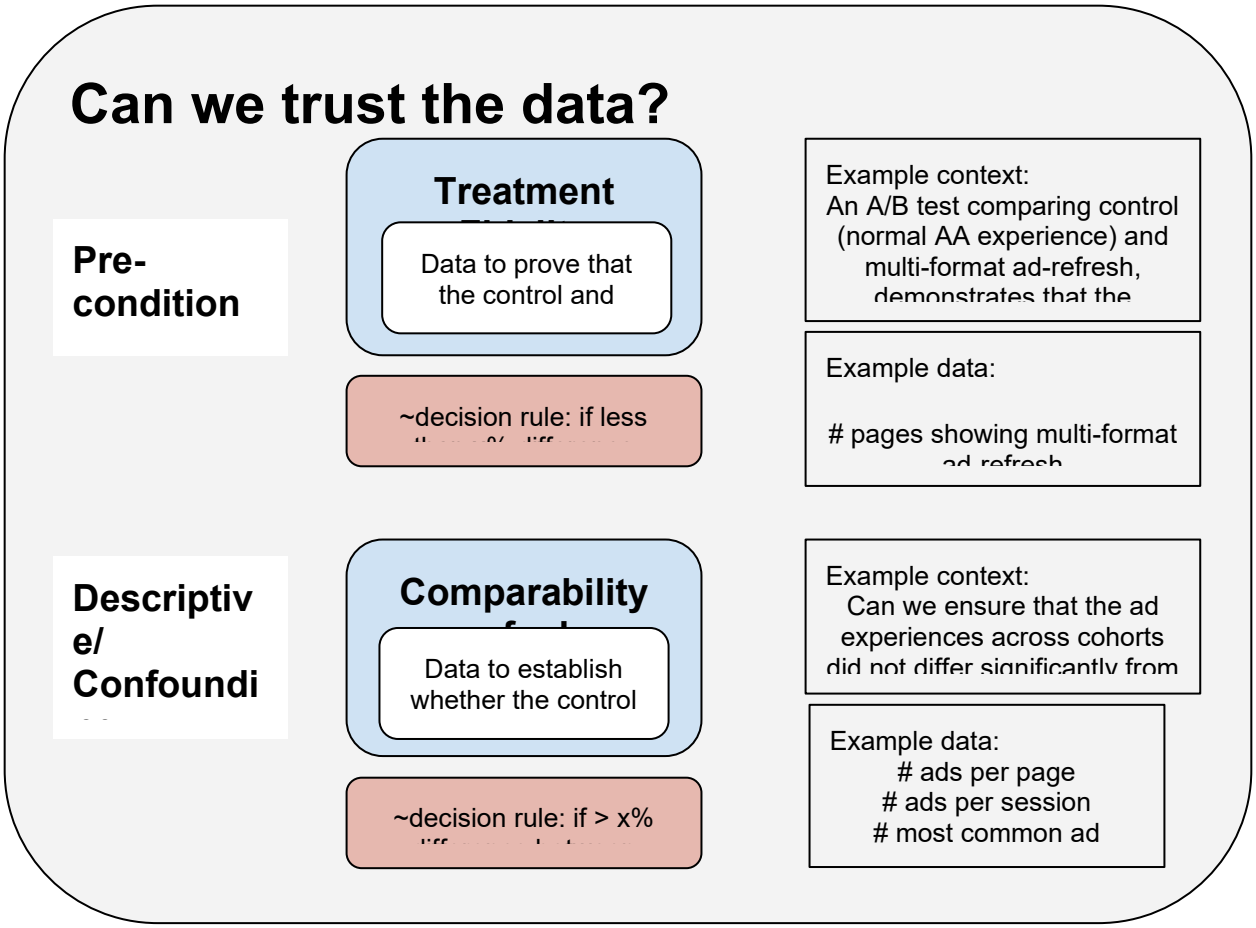

Figure 2. Identifying data analytics and user behavior data to validate A/B testing and multivariate experimentation.

Data Cues and Interpretations

There is a range of potential user behavior data that can be used to make inferences about user experience. **A potential way to refine, structure, and interpret this data is to consider its strength in terms of how strongly it signals negative (or positive) user reactions.** Specifically, we can identify user experience and behavior data to collect prior to the initiation of a test and then delve into the data in pre-specified steps to determine impact – or build confidence in there being minimal negative effects – on user experience.

For instance, we can consider test data like **a decision tree with a process of elimination: if at each node user behavior is not negatively affected by a specific ad experience relative to the control group, then this accumulates evidence in favour of the ad experience**. Justification for this evidence accumulation approach is to avoid setting excessively high criteria for rejecting any tested ad experience (e.g., an increase in uninstall rates) and to ensure potential subtle effects that could impact users or other stakeholders are not overlooked (e.g., increase bounce rates on affected webpages).

The approach could look something like that presented in Figure 3 below. At each node, there is an exit where a decision/interpretation of the results could be made. For example, if there are indications that strong data cues point to a more negative user experience, a decision about user tolerance or acceptability could already be made. If not, examination of moderate cues could ensure less obvious negative effects of an ad experience are not overlooked. **Importantly, the approach does not suggest that an ad experience should only be considered acceptable if all nodes are passed without negative effects on user experience – data still requires interpretation.** Rather, the approach suggests that certain data be given more weight and scrutiny than others.

eyeo is still establishing its capacity to collect a broad range of user experience and behavior data. Nevertheless, Figure 3 lists some example cues that are currently being explored, categorising them according to whether they would provide strong, moderate or weak cues about the user experience.

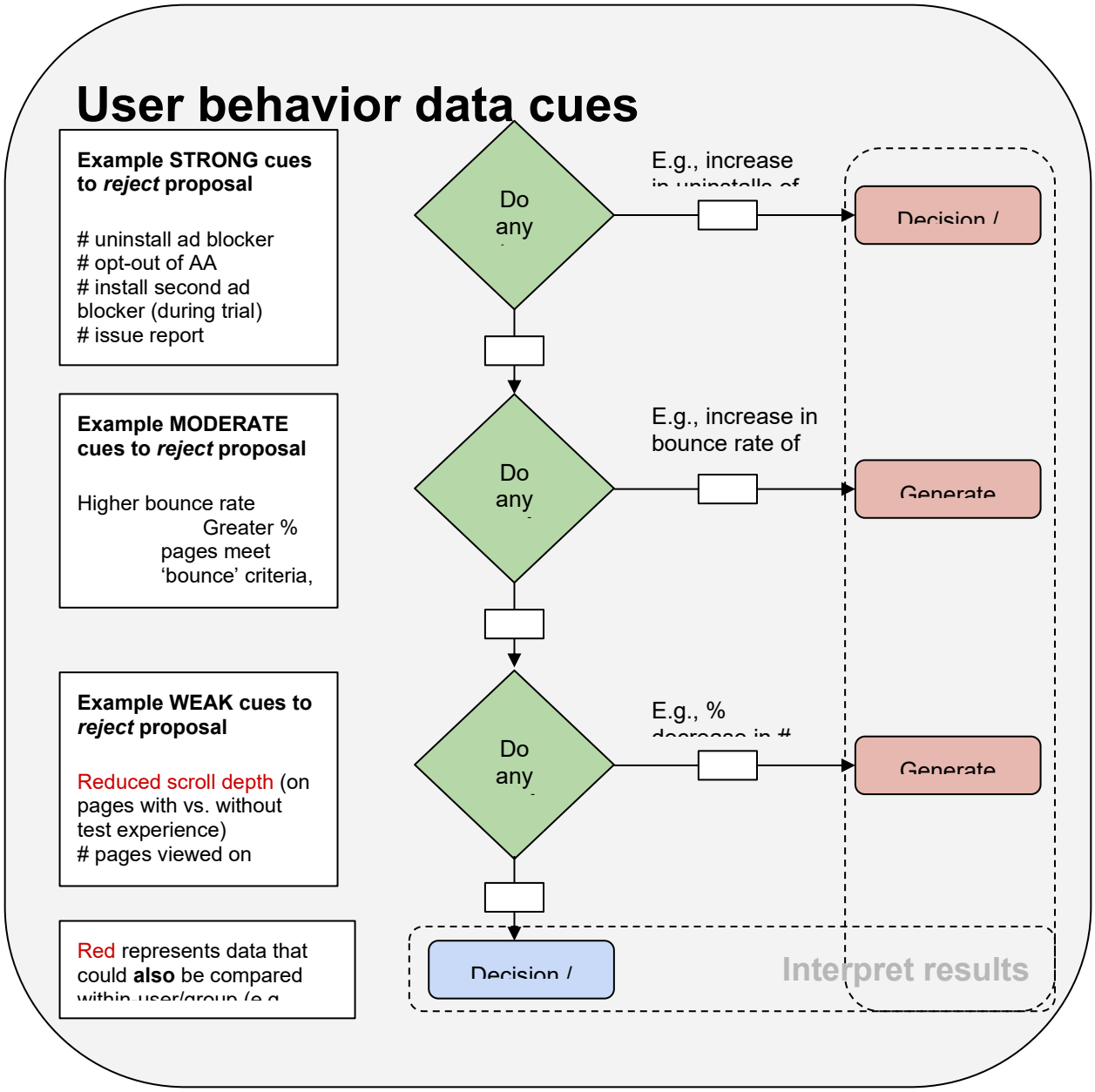

Figure 3. Approach to evaluating user behavior data in support of an AA Standard change.

Reporting and Decision-Making

Reporting the results

Reports should include the research question, research method justification and a summary of the data sources and results.

Decision process and criteria

In the current bylaws, proposed changes to the AA Standard need to be accompanied by user research demonstrating that a new ad experience does not exceed a 35% annoyance level threshold with ad blocking users. This requirement would be too narrow to incorporate the broad range of user experience and behaviour data now available. Incorporating this data would require refinement of the decision criteria and the decision process.

Owing to the diversity of user data available and the various methods that could be used and combined to answer research questions, a single threshold or decision criteria would not accommodate the broad range of research approaches. Nevertheless, the decision process can be informed by three general guidelines:

1. Ensure research quality and adequacy of the data,
2. Weigh and integrate user perception and behaviour data, and
3. Recommendation and subsequent monitoring processes

Research quality checklist

Regarding the first guideline, ensuring data quality, relevance, and coverage can be considered as a first step in the decision process. Given that these criteria are essential, any submission should be able to complete a simple checklist as a prerequisite:

RESEARCH QUALITY CHECKLIST

Data quality and relevance

- Does the sample reflect a relevant or broad enough audience to justify the proposed change (e.g., targeted audience, or representative ad blocking audience from multiple countries)
- Does the data represent a realistic setting that approximates the context in which changes would be applied?
- Is there evidence to demonstrate that the data is reliable (e.g., Figure 2)?

User perception data

- Does the research incorporate user perception, preference or attitude data (e.g., Figure 1)?

Weighing and integrating research data

Owing the variety of research methods and user data that could accompany a proposal, there is no bright-line test that can be applied across the board. Nevertheless, the following decision rules can guide decisions based on the data:

- Data collected for the purpose of generating hypotheses (e.g., Figure 1, first node) or to refine one (Figure 1, second node) can never be sufficient to justify a proposal. However, evidence from one of these methods could bolster or support a proposed change..
- Consistent with the pre-existing bylaws, Experimental surveys can apply the 35% annoyance level threshold for acceptability.
- User behaviour data needs to be accompanied by some supportive user perception data (see Figure 1, User Feedback Requirement), whether this is collected in parallel or derived from prior or related research.
- User behaviour data collected from A/B or multivariate testing will need to specify a priori a % difference between a control and test condition that would indicate a positive/negative effect.
    - In line with Figure 3, behavioural data cues differ in strength and support for a change would be accumulated through a process of elimination.
        - Strong cues: if any of these data in a test condition exceeds a pre-specified increase of x% relative to the control or other test conditions, **there is strong evidence that it had a negative effect on user experience.** Otherwise, there is no indication that the change would have a strong effect.
        - The same logic applies for moderate and weak cues. If any of these data in a test condition exceeds a pre-specified increase of x% relative to the control or other test conditions, there is moderate or weak evidence that it had a negative effect on user experience, respectively. Otherwise, there is no indication that the change would have a moderate or weak effect, respectively.
        - Ideally, tests should aim to collect a mix of relevant strong, moderate and weak user behaviour data to accumulate evidence that there are no negative effects for strong cues as well as more subtle outcomes that could impact stakeholders (e.g., see Data Cues and Interpretation).
    - **The greater amount of user behaviour data collected, the greater the strength of evidence in support of a proposal. However, in general, evidence from strong cues should be weighted highest and weak cues weighted lowest.**
- If user perception and user behaviour data conflict or show mixed results, greater weight can be assigned to evidence that has been:
    - derived from a greater amount of user data (e.g., experimental survey or user behaviour data with hundreds/thousands of users vs. qualitative data derived from a small sample of users), and

- from sources most similar to a realistic user experience (e.g., A/B or multivariate experiments, experimental surveys that simulate user experience vs. indirect or third-party evidence, such as market research).

Recommendations and next steps

Assuming there is some positive data to support a change to the AA Standard, there could make one of the following recommendations:

Request more data	**Preliminary acceptance**	**Acceptance**
Request additional research to address a gap in evidence or gain additional understanding as to why data sources conflict (e.g., user perception and user behaviour data conflict, and there is no clear justification for why one data source should be weighted more heavily)	There is moderate to strong evidence to support a change, but the AAC would like more data to confirm no longer-term, carryover or other effects. For example, the committee could request: - additional user perception data to be collected in parallel to the implementation; - a monitoring period where additional user behaviour data is collected (e.g., longer-term monitoring of strong, moderate and/or weak cues, Figure 3).	There is sufficient evidence to support and implement the change.

Worked example: Animation

Note, the worked example lists examples of user experience and behavior data that could be used to evaluate ad experiences. The feasibility of collecting these data sources is still being determined.

Research question: How open are ad blocking users to seeing *some* animated ads?

We know from previous AA research that ad blocking users find animated ads to be disruptive to their online experience. However, these studies have focused on user perceptions of animated ads in general or based on viewing animated ads on a single webpage in a mock-web experience. Prior research does not reveal whether ad blocking users are open to or would tolerate seeing *some* animated ads during a browsing session – that is, users may tolerate 5% of AA they encounter being animated provided they occur within the AA Standard experience.

The purpose of this research is to explore ad blocking users' tolerance for animated ads that appear in 0%, 5%, 10% or 20% of AA placements during a browsing session.

Research method justification

What is the purpose of the research? → Test hypothesis: Ad blocking users will tolerate up to 10% animated ads in the AA Standard experience.

How concrete is the idea or mechanism? → Concrete: Test four animated ad quotas (0%, 5%, 10% or 20%).

On how many dimensions do testing conditions vary? → Single: tolerance for different animated ad quotas (0%, 5%, 10% or 20%).

Main Research Method: A/B/C/D test

To fulfill the user feedback data requirement, **user perception data will be sourced from**:

1. Prior user research on user perceptions of online advertising (suggests ad blocking users are bothered by animated ads in general but are open to viewing more intense ad experiences with certain trade-offs)
2. In-depth interviews with ad blocking users exploring their openness to or tolerance for viewing *some* animated ads, along with any conditions they foresee as being important
3. In-product messaging survey to collect real-time evaluations of satisfaction with browsing experience

Research methods can be conducted in parallel as shown in Figure 4.

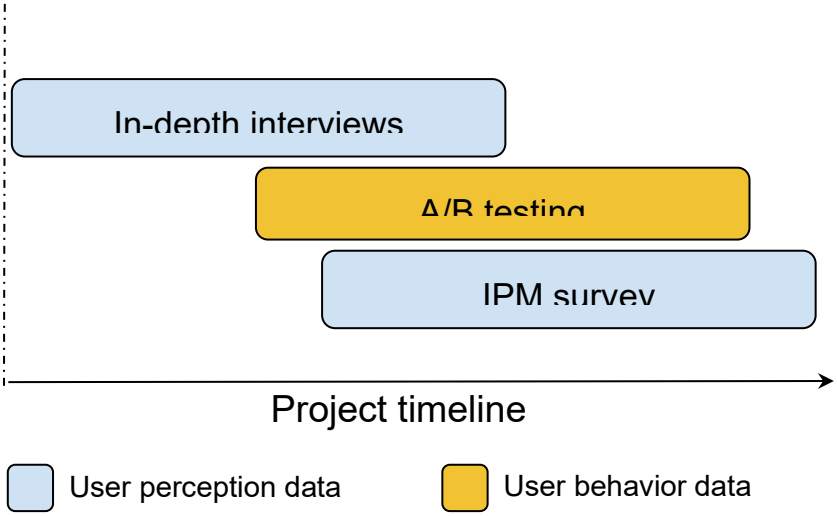

Figure 4. Research plan for animated ads

Data reliability

Treatment Fidelity

To ensure test data is reliable, compare the following user experience data across conditions:

To calculate animation rate per condition (% pages showing animated ads):

> # pages where an animated ad was shown during a browsing session / # pages visited during a browsing session

To calculate animation rate according to viewability (% pages showing animated ad in-view):

> # pages an animated ad was shown AND >x% of the ad was viewable (or equivalent viewability metric) / # pages visited during a browsing session

Comparability of ad experiences

To evaluate how comparable the test conditions are on other relevant metrics, check the similarity of the following user experience data across conditions:

> # ads per page (average)
> # ads per browsing session (average)
> # animated ads per page view (average)
> # violations encountered per browsing session
> …

Test data

The following outcome data will provide cues as to how the user experience was affected by the test conditions:

Strong user behavior cues

If any of these data in a test condition exceeds a pre-specified increase of x% relative to other test conditions, **there is strong evidence that it had a negative effect on user experience**.

- # uninstall ad blocker
- # AA opt-out
- # installed second ad blocker during trial
- # issue report submissions
- # clicked on BT ad label
- # clicked on BT ad label and submitted “xxx” outcome
- # attempt to disengage with ads (e.g., if animated ads have an “x” to close or a pause button)

Moderate user behavior cues

If any of these data in a test condition exceeds a pre-specified difference of x% relative to other test conditions, **there is moderate evidence to suggest there may be a negative effect on user experience**.

Between A/B test groups (compare the data between test groups):

- Bounce rate
    - Greater % pages meet 'bounce' criteria, on average (group level)
- Reduced average session length
- Reduced average page duration

Within A/B test groups (compare data within the test groups on pages with/without test experience):

- Bounce rate
    - Fast(er) bounce rate on pages with test ad experience (vs. not)
- Reduced page duration (on pages with/without test experience)
- Reduced ad/s viewability duration (animated vs. non-animated ads)

Weak user behavior cues

If any of these data in a test condition exceeds a pre-specified difference of x% relative to other test conditions, **there is weak evidence to suggest there may be a negative effect on user experience**.

- Reduced average scroll depth (on pages with test experience)
- # pages viewed on domains (with test experience)
- # pages user is 'engaged' (e.g., >10 seconds, key event, or > 2 page views)
- # interactions with page or domain