Helping Users to Assess Content Online:

Integrating Behavioural Tools through an Adaptive Randomised Controlled Trial

Economics Discussion Paper Series – Issue 16

Published 24 February 2026

Contents

Section

Executive Summary

Ofcom is the United Kingdom's independent communications regulator. We regulate fixed and mobile telecoms, TV and radio broadcasting, post, the radio spectrum used by wireless devices, and online safety. Under the Communications Act 2003, Ofcom holds statutory duties to promote and to carry out research into media literacy. We aim to promote public understanding of online content and support users in navigating digital information environments. Our work is grounded in fundamental rights, particularly the right to freedom of expression under Article 10 of the European Convention of Human Rights. The Online Safety Act (OSA) bolstered our media literacy duties, including by adding a requirement to take steps to improve users' ability to assess the reliability, accuracy and authenticity of content, and ultimately to understand and reduce their exposure to the phenomena of disinformation and misinformation.[1] In practice, this involves enabling UK users of online services to critically assess the content they encounter online.

The experimental study set out in this Discussion Paper is specifically targeted at exploring the efficacy of various interventions aimed at enabling users to assess the reliability, veracity and authenticity of online content they see, thereby supporting users in making more informed judgements about it. The research described in this paper tests a new methodology (Adaptive Randomised Control Trials) for exploring this area.

The study does not offer a broad assessment of platform interventions or their overall effectiveness. Rather, it presents findings from a controlled experiment that tested a range of behavioural tools — including educational quizzes, prompts and content labels — to understand how they might support users in identifying potentially misleading content. The insights generated here are intended to inform best practice and contribute to the wider evidence base on media literacy design, in line with our strategic commitment to supporting public resilience to online harms.

The study was designed to evaluate the effectiveness of various interventions aimed at enhancing users' ability to assess reliability, veracity and authenticity. Their ability to assess was tested based on their ability to identify content that had been independently verified as false. For brevity, we will refer to this content in this report as "false content".

Adaptive Randomised Control Trials (ARCTs) employ statistical algorithms to update the experimental design in real time based on interim results. By adjusting the randomisation throughout the experiment, ARCTs enable the testing of a larger number of interventions and more efficient identification of the best interventions, compared to standard Randomised Controlled Trials (RCTs). In this study, 6,000 participants representative of the United Kingdom (UK) population were shown a simulated social media feed with some true and some false posts which were independently verified by fact checking organisations.[2] Participants were asked to assess the

[1] Online Safety Act, 2023. Chapter 8: Media Literacy. This includes a requirement to help users understand the nature and impact of disinformation and misinformation, reduce their and others' exposure to it and learn how to critically assess sources including manipulated content. While there is no agreed definition of 'misinformation' or 'disinformation', we use those terms because they are used in the statute.

[2] In particular for posts containing false content, we used content that had been confirmed false by independent fact-checkers such as Full Fact and The Ferret. For the true posts, we collated content from reputable sources like the Bank of England, ONS, Science.org, and the BBC, as well as from fact-checkers like Snopes.

veracity of the information contained in each post. While all participants were shown the same set of posts, a different version of the feed was shown depending on the intervention.

Our interventions employed (1) user-focused tools, including an educational quiz where participants learned to detect potentially false content through evaluating posts and receiving instant feedback (a technique often referred to as “inoculation”) and a reminder prompt reinforcing lessons halfway through the feed; and (2) post-focused tools or informational labels including fact-checker labels, AI verification labels and crowdsourced notes. The interventions were tested individually as well as in combination.

Based on our experiment, we found:

- A combination of user-focused tools and labels were most effective at improving participants’ ability to identify false content.
- When tested in isolation, user-focused tools generally outperformed labels.
- Interventions did not significantly reduce trust in verifiably true information.
- Crowdsourced notes were marginally less effective than AI and fact-checker labels.

While these findings provide novel evidence on the effect of different types of media-literacy tools and their complementarities, they should be interpreted in light of the study’s limitations. First, the experiment was conducted in a controlled, simulated environment, which cannot reproduce many key features of real-world online platforms and therefore limits the generalisability of the results. Second, the findings depend heavily on the specific posts shown to participants, meaning results could vary if different posts were used. Finally, the effects of the interventions might differ in live platform contexts or over longer time horizons. Therefore, the findings should be viewed as indicative evidence in a controlled setting rather than direct estimates of real-world impacts.

1. Introduction to the Research

We carried out the research described in this paper as a means of testing a new methodology to explore how users' ability to assess the reliability, veracity and authenticity of online content might be improved. Insights from this research will inform the evidence-based recommendations and guidance Ofcom provides to platforms that might help build public resilience to online information threats. This experimental study contributes to the evidence base in two important ways.

The study provides comprehensive, robust, UK-specific insights into how different interventions can help people assess reliability, veracity and authenticity of content online.

Previous research has typically tested two or three interventions at a time, often using varied outcome measures, experimental settings, and samples. This variability made it difficult to compare findings across studies and to identify the most effective treatment. By evaluating a wide range of interventions within a single, unified experimental framework, our study overcomes these limitations and delivers a clearer, evidence-based understanding of which strategies work best within a controlled environment.

To make the study relevant to the UK we created a mock social media feed with UK-specific content and recruited a representative UK sample. Most previous research is focused on the US; this study helps fill a gap in the literature by providing insights that can help inform UK policy and practice.

The study explores how interventions work in combination.

This study also tested how interventions interact when combined, an area that has received limited attention in prior research. This was made possible using an ARCT methodology. Unlike standard RCTs that allocate participants evenly across trial arms, ARCTs adjust allocations based on observed performance in real-time, sending more participants to more promising interventions. This efficient use of experimental resources allows to identify the most effective among a larger set of treatments without requiring infeasibly large samples. We were therefore able to test combinations of interventions in a resource effective way to identify potential complementarities between them.

These insights are particularly timely given the growing reliance on online platforms for news and information. Ofcom's 2025 News Consumption Report shows that 71% of UK adults access their news online, with 51% relying on social media. Among 16–24-year-olds, these figures rise to 80% and 75%, respectively.[3] Yet, this shift has coincided with growing concerns about misinformation (including false content) online.[4] Ofcom's 2025 Online Nation report found that around four in ten UK adult internet users self-report having encountered misinformation online in the previous four weeks, making misinformation the most prevalent potential harm identified in the survey.[5]

These concerns are further supported by academic research showing that false news spreads more easily and rapidly across digital platforms than true news. [6] [7] In response, behavioural and social science research has focused on empowering users to critically evaluate online content. According to

[3] Ofcom, 2025, News Consumption in the UK.
[4] While there is no agreed definition of the term 'misinformation', it can be defined as false information that is spread by mistake. When there is intent to mislead, it is called "disinformation" (Lewandowsky et al., 2020, Misinformation and Its Correction: Continued Influence and Successful Debiasing).
[5] Ofcom, 2025, Online Nations Report.
[6] Broda E. and Strömbäck J., 2024, Misinformation, Disinformation, and Fake News: Lessons from an Interdisciplinary, Systematic Literature Review.
[7] Vosoughi et al., 2018, The spread of true and false news online.

the OECD's report on misinformation (2022), behaviourally informed interventions - such as accuracy prompts, labels, educational quizzes - have proven to be effective and scalable strategies for curbing the spread of false information.[8] [9] This study builds on that foundation, offering new UK-specific evidence that can inform future efforts aimed at strengthening public resilience to online information threats.

[8] OECD, 2022, Misinformation and disinformation.

[9] Effective interventions include prompting users to consider accuracy, enhancing digital literacy, and using inoculation techniques to build resilience. For a broader literature review, see the Technical Paper.

2. Trial Design and Setup

Experimental Design

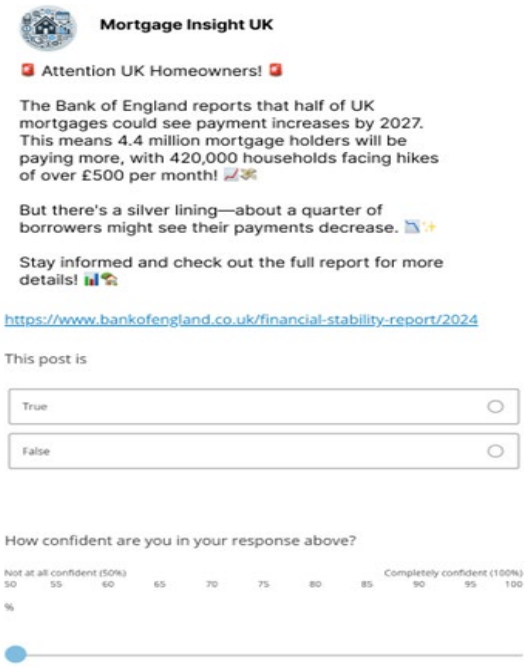
In March 2025, 6,000 adults (18+) representative of the UK population took part in the experiment.[10] Participants were presented with 15 social media posts, shown in random order. Out of 15 posts, 10 posts contained true information ("true posts"), while 5 included false content ("false posts").[11] The posts covered a range of topics including economy/finance, health, migration, politics, and science.[12]

After each post, participants were asked:

- To choose whether the post was true or false (binary choice)
- "How confident are you in your response above?" rated on a scale between 50 (not at all confident) to 100 (completely confident).

An example of a "true" post, as shown to participants, is given in Figure 1.

Figure 1: Example Post



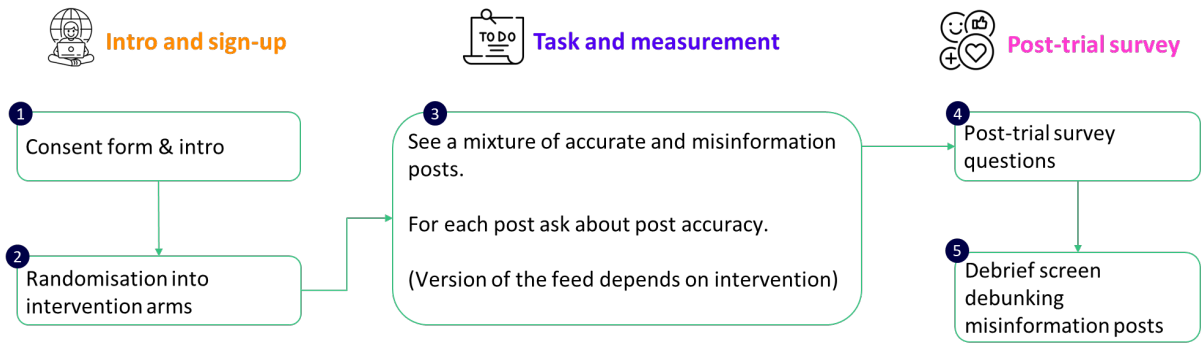
[10] Participants were recruited through Prolific. Prolific recruits participants by connecting researchers with a diverse pool of eligible individuals who are willing to participate in studies. Those who sign up are screened for eligibility and invited to take part in studies based on criteria specified by researchers. They were paid for participating (£1.20) and received a bonus for correctly identifying whether posts were true or false to incentivise accuracy (£0.06 per correct post).

[11] For the false posts we used content that had been confirmed false by independent fact-checkers such as Full Fact and The Ferret. The posts featured randomly generated common usernames and AI-generated profile pictures. The text was kept close to the original but not word-for-word to mitigate privacy risks. For the true posts, we collated content from reputable sources like the Bank of England, ONS, Science.org, and the BBC, as well as from fact-checkers like Snopes.

[12] The topics have been chosen to replicate a representative UK news diet based on Ofcom's previous research (Ofcom, 2021, Understanding online false information in the UK).

After they had responded to all the posts, participants were asked questions about their media consumption habits, social media attitudes and usage, and demographic characteristics including age, gender, ethnicity, education, and political leanings. At the end of the study, participants were shown which posts were false, given links to trustworthy sources, and offered extra information to help them learn how to assess the accuracy of information.[13] See Figure 2 for experimental flow.

Figure 2: Experimental Flow



Primary outcome

The primary outcome is a single **overall accuracy score** that combines participants' answers to whether each post is true or false with their reported confidence level.

The accuracy score for each post ranges from 0 to 100 and is calculated based on two things: whether the participant judged the veracity of the post correctly and how confident they were in their answer. Greater confidence in a correct answer leads to a higher score, while greater confidence in an incorrect answer leads to a lower score. A participant's overall accuracy score is the average of their scores across all 15 posts shown during the experiment.

This scale gives a comprehensive measure of how well participants can assess the veracity of content. It captures not only whether an intervention shifts answers from wrong to right, but also whether it strengthens confidence in correct answers (and weakens confidence in incorrect ones).[14]

Interventions and Trial Arms

The experiment considered two main types of interventions aimed at helping people identify false content online: user-focused tools and post-focused tools.[15]

User-focused tools:

- **Educational Quiz/Inoculation**: Before viewing the simulated social media feed, participants engaged in a short quiz designed to train them to detect false content. They judged whether

[13] This included referring participants to the independent fact-checker websites that were used to collate false content for this experiment.
[14] See the Technical Paper for a formal definition of the primary and secondary outcomes.
[15] We used an existing classification of interventions against online misinformation that have been shown to be effective in the literature (Kozyreva, A., Lorenz-Spreen, P., Herzog, S.M. et al., Toolbox of individual-level interventions against online misinformation.)

the post was true or false and got instant feedback, along with tips on how to spot false content (see Figure 3).

- **Educational Quiz + Reminder Prompt**: In addition to the Educational Quiz, participants also saw a quick reminder of the key lessons from the quiz halfway through the feed (see Figure 4).

Figure 3: Example Educational Quiz

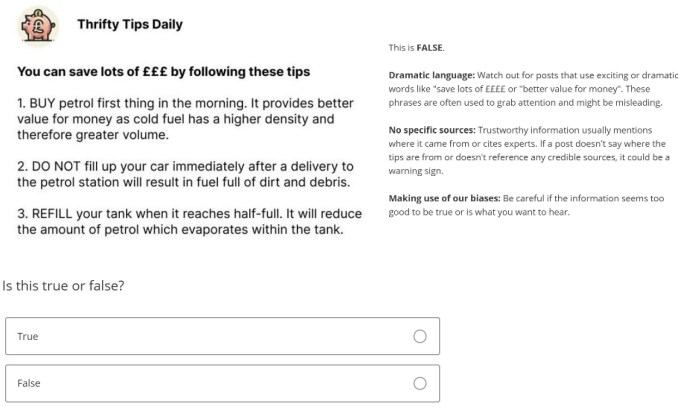
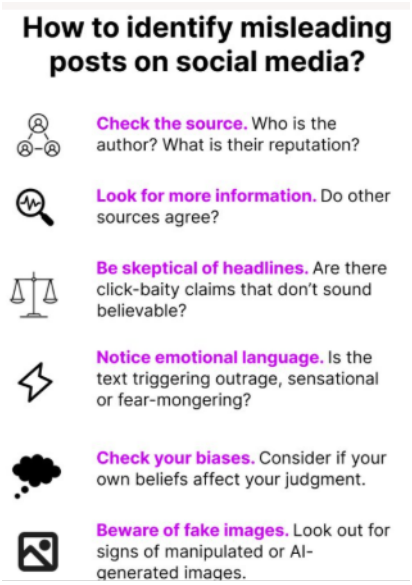


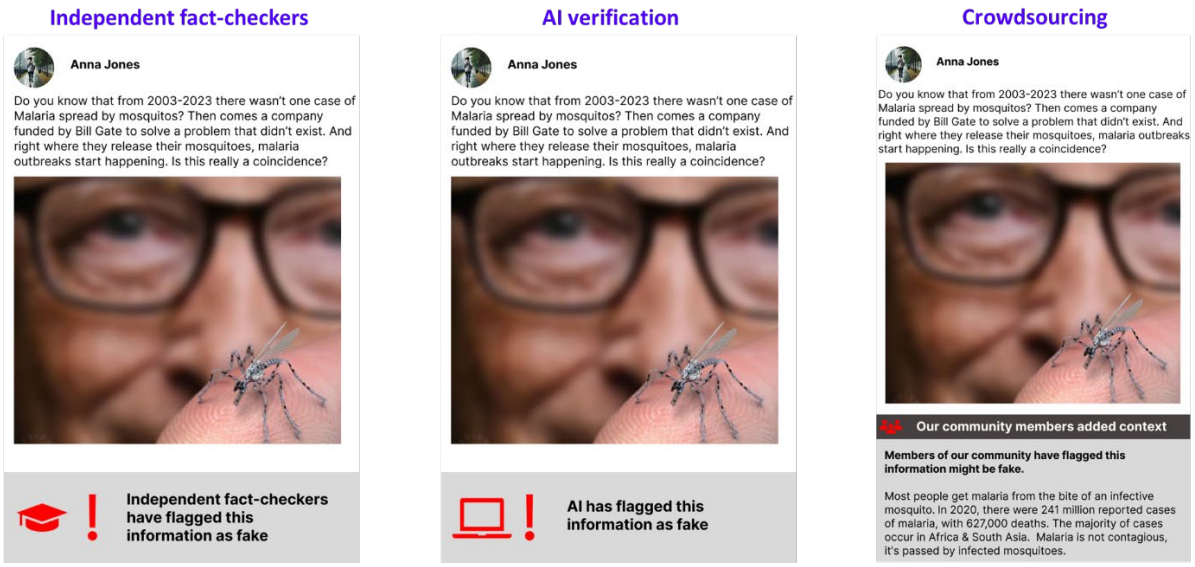
Figure 4: Example Reminder Prompt



Post-focused tools (Labels)

- **Fact-Checker Label**: Posts containing false content were flagged with the message "Independent fact-checkers have flagged this information as fake."
- **AI Verification Label**: Posts containing false content were flagged with the message "AI has flagged this information as fake." This label indicates that an AI system has analysed the content and assessed it to be false or misleading.
- **Crowdsourced Notes**: Posts containing false content were flagged with the message "Our community members added context. Members of our community have flagged this information might be fake", followed by additional explanatory/de-bunking information provided by community members (see Figure 5 for examples of all three types of intervention).

Figure 5: Example Post-focused Interventions



The experiment tested interventions from the two types—both individually and in combinations—alongside a control group that did not receive any intervention (see Table 1 for a list of all trial arms).

Further details of the experimental design and set-up, together with the full set of trial results, are provided in the Technical Paper published alongside this report.

Table 1: Trial Arms

	Educational quiz	Reminder prompt	Label
T1 Control	No	No	No
T2 Educational q.	Yes	No	No
T3 Educational q.+ Reminder prompt	Yes	Yes	No
T4 AI Label	No	No	AI Label
T5 Fact-checker label	No	No	Fact-checker label
T6 Crowdsourced notes	No	No	Crowdsourced notes
T7 Educational q. + AI label	Yes	No	AI Label
T8 Educational q. + Fact-checker label	Yes	No	Fact-checker label
T9 Educational q. + Crowdsourced notes	Yes	No	Crowdsourced notes
T10 Educational q. + AI label+ Reminder Prompt	Yes	Yes	AI Label
T11 Educational q. + Fact-checker label + Reminder Prompt	Yes	Yes	Fact-checker label
T12 Educational q. + Crowdsourced notes + Reminder Prompt	Yes	Yes	Crowdsourced notes

3. Findings

In this section we present the key findings from the experiment. We refer the reader to the Technical Paper for more detailed reporting and full statistical analysis of the results.

User-focused tools and labels worked best in combination

Treatment groups that received combinations of user-focused tools and labels showed the highest accuracy in assessing the veracity of content. In particular, the best performing treatment Educational Quiz + Prompt + Fact-checker label (Treatment 11) produced a significant improvement in the measured accuracy score compared to the control group. Other combinations displayed similar effects.[16]

When given in isolation, both types of interventions still produced improvements relative to the control group, with user-focused interventions substantially outperforming labels. This is particularly striking given that labels explicitly flagged false posts, whereas user-focused tools did not provide any direct indication of the veracity of specific posts to participants.

These findings point to complementarities between these two types of interventions, suggesting that they may act through partly distinct behavioural channels. The evidence on the comparatively stronger performance of user-focused tools also may highlight the importance of empowering users to make their own informed assessment of content on social media.

Interventions did not adversely affect trust in genuine information in this experiment

We found no evidence that any of the interventions—whether in isolation or in combination—reduced trust in the genuine information. In fact, all treatments displayed insignificant differences in accuracy score for true posts compared to the control group.

This finding helps assuage the concern that these interventions might backfire by causing users to doubt the veracity of all content seen on social media.

AI and Fact-checker labels outperformed Crowdsourced Notes

Comparing the effectiveness of user-level interventions, we find that treatments involving AI and Fact-checkers' labels were slightly more effective at improving participants' ability to identify false content. This difference is particularly salient when labels were used in isolation, as Crowdsourced Notes alone (Treatment 6) failed to produce statistically significant effects relative to the control group.

This finding may reflect greater perceived impartiality of content verification mechanisms that are backed by established institutional or technological infrastructure. It may also reflect the different presentation of Crowdsourced notes in terms of wording and information provided, compared to Fact-checker and AI labels.

[16] For posts containing false content, Treatment 11 produced a 18.55% increase in accuracy score compared to the control condition (associated 95% confidence interval: [16.49%, 20.61%]).

4. Conclusion and limitations

This study provides novel experimental evidence on the effectiveness of a broad set of media literacy tools designed to help users identify false content on social media. Drawing on a UK nationally representative sample, our findings show that combinations of user-focused tools and post-focused labels were most effective, while user-focused tools on their own delivered stronger improvements than labels alone. Crucially, the interventions did not reduce trust in genuine content.

This research contributes to strengthening Ofcom's research capability in using Adaptive Randomised Controlled Trials. This approach enabled us to test a large number of treatments in a cost-effective way, including combinations of interventions to examine their complementarity.

Several limitations should be noted when interpreting these results:

First, the effectiveness of any intervention may vary depending on the context in which it is studied. This experiment simulated a social media environment, which may not fully capture real-world user behaviour. On actual platforms, users' motivations and incentives may be different and algorithmic curation and personalisation may play a significant role. Additionally, although this study used a nationally representative sample from the UK population, the findings may not generalise to contexts with different demographic and cultural characteristics.

Second, the results of this study crucially depend on the specific content shown to the participants. Since the reliability of content is known to be highly context-dependent, the impact of interventions could change if different types of false posts (including content generated by AI) were to be used.

Third, the study only measured short-term impacts. This potentially limits our ability to draw conclusions about long-term effects, especially since the effect of educational tools might attenuate over time.

Therefore, while our findings provide valuable insights, further research is needed to test interventions in more realistic settings and over longer horizons such as longitudinal studies. Such evidence is crucial to establish which interventions best allow users to assess reliability, veracity and authenticity of content and how they might be most effectively deployed.