

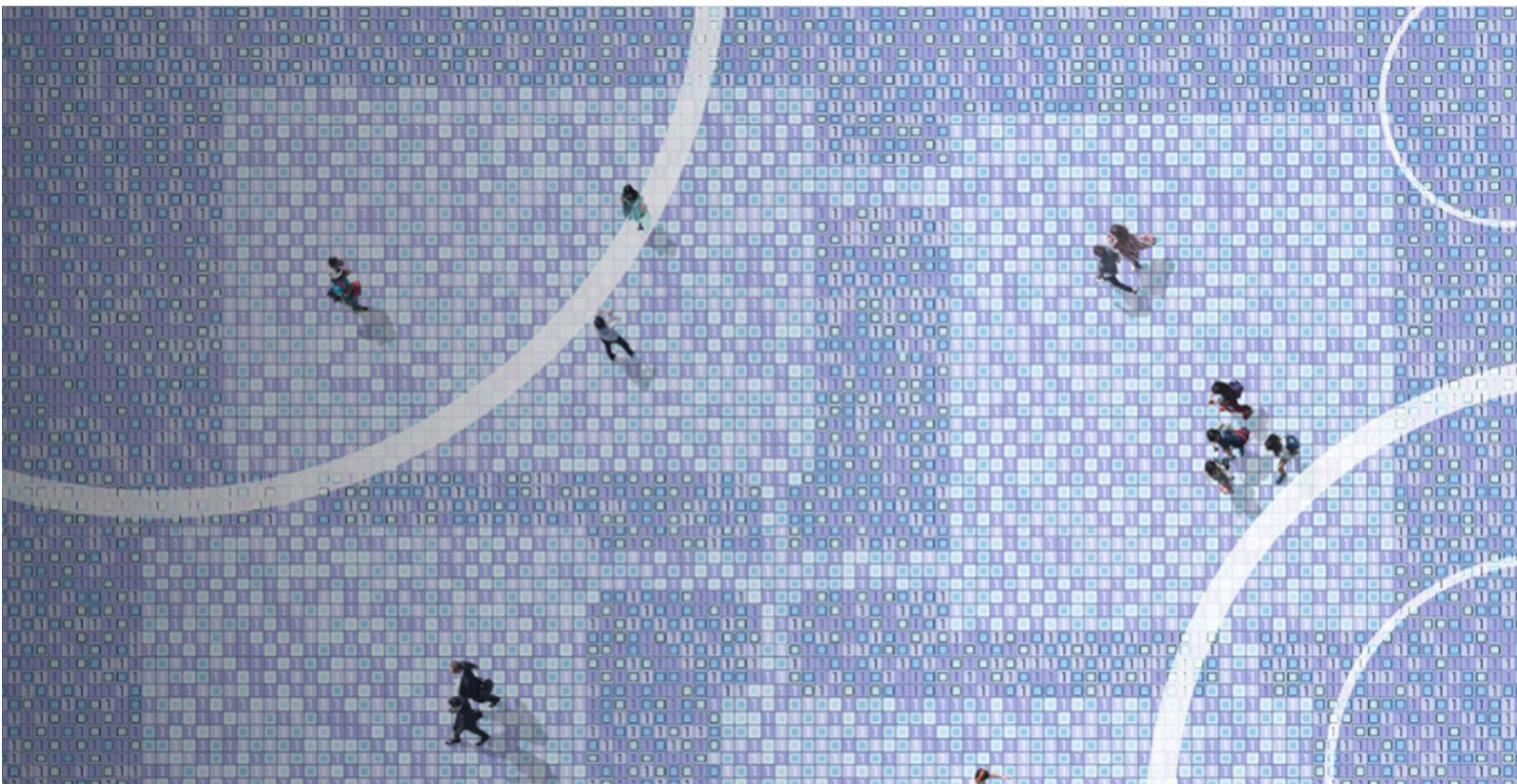
KANTAR PUBLIC

Ofcom online trials: reporting mechanisms of video sharing platforms

Trial report

Kantar Public Behavioural Practice: Michael Ratajczak, Yuchen Yang, Rob McPhedran, and Max Mawby.

For Ofcom



1. Background and Objectives

1.1 Ofcom's remit – Media Literacy

Ofcom has a duty to promote media literacy, including in respect of material available on the internet.

Regulation of UK-established Video Sharing Platforms (VSPs) is a recent but important element of Ofcom's responsibility as the communications regulator in the UK. VSPs - and social media in general - have the capacity to bring an extremely wide range of content direct to any user in a way that encourages immersive engagement. In many cases, this immersive engagement with different types of content will have positive effects e.g. in creating connections or social ties between a diverse array of individuals.

However, in some cases the content may be illegal and users should not be exposed to it. Alternatively, the content could be legal but carries with it the risk of causing psychological, physical, or financial harm to particular groups of individuals and users need to be warned about exposure to such content and have the ability to report such content to platforms to protect others. As of November 2020, Ofcom oversees the regulatory regime which requires UK-established VSP providers to include measures and processes in their services that to protect users from the risk of viewing harmful content.

Ofcom is looking at different methods for researching the effectiveness of the different safety measures used by online platforms to safeguard users from harm. In particular, it is looking to test the use of online Randomised Control Trials ('RCTs') as a method for understanding the impact of the design of online safety measures on users.

1.2 Experiment aims and objectives

This online RCT aimed to contribute to this evidence building process. Specifically, the interventions within the experiment test different ways of increasing both the volume and quality of reporting. Interventions therefore had to encourage three behaviours:

- Prompting users to **start a report**: by increasing salience of the reporting option (in the context of distractions created by the content and the online environment); setting expectations of what reporting involves; and increasing confidence that reporting is worthwhile.
- Encouraging users to **complete a report** once they have started: removing friction to facilitate submission.
- Supporting users to **submit enough information** to allow platforms to act: removing cognitive barriers by restricting options and/or making choices easier.

2. Sample and data collection

2.1 Sample

The target population, in this study, consisted of UK VSP users. As no official statistics were available on the specific demographic breakdown of VSP users in the UK, this experiment took a two-stage approach to providing a sample that was as representative as possible with respect to key demographic characteristics. A total of 2,400 UK participants, **aged between 18 and 69**, were recruited from Kantar's LifePoints panel. All participants indicated that they had used a VSP in the past 12 months in response to a screener question provided at the beginning of the experiment.

Kantar Public conducted this experiment online, using a device-agnostic platform. As such, the experiment could have been completed on a computer, mobile, or tablet, subject to participants' preference. Fieldwork took place in January 2022 over a three-week period.

In the first stage, the experiment started with parallel demographic quotas. These quotas were based on the ONS mid-2019 population estimates¹ and the 2011 Census data.² In the second stage, these quotas were adjusted, based on the relative proportions of respondents in each demographic sub- group who passed the screener. This meant that the quotas for those sub-groups most likely to pass the screener were increased, while those least likely to do so were decreased. The aim of this approach was to achieve a sample of UK VSP users (the adjustment process is described in more detail below).

¹<https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/bulletins/annualmidyea rpopulationestimates/mid2019estimates>

² <https://www.ons.gov.uk/census/2011census/2011censusdata>

Table 1. Demographic quotas set at the start of the study, and the quotas achieved.

Demographics		Start	Finish
Gender	Male	49%	49%
	Female	51%	51%
	Other	-	<1%
Age	18-24	13%	14%
	25-39	31%	34%
	40-54	30%	30%
	55-69	26%	22%
Ethnicity	White	87%	86%
	Mixed/Multiple Ethnic Groups	2%	2%
	Asian/Asian British	7%	7%
	Black/African/Caribbean/Black British	3%	3%
	Other Ethnic Group	1%	<1%
	Refused to disclose	-	<1%
Socio-economic grade	ABC1	55%	56%
	C2DE	45%	44%
Location	London	13%	14%
	East Midlands	7%	7%
	West Midlands	9%	9%
	East of England	9%	9%
	North East	4%	4%
	North West	11%	11%
	Yorkshire and the Humber	8%	8%
	South East	14%	14%
	South West	8%	8%
	Wales	5%	5%
	Scotland	8%	8%
	Northern Ireland	3%	3%

The quotas were monitored at the end of the soft launch (see section 2.7), and at the half-way point of data collection (n=1,200). This is because the sample at the half-way point of

data collection was intended to reflect the presumed target population with regards to selected demographic characteristics. The quotas were recalibrated at the mid-point of the experiment based on the relative proportions of respondents in each demographic subgroup who passed the screener. Note that nationally representative quotas for ethnicity were kept for the second phase of recruitment, because not enough information was available on the ethnicity profiles of users failing the VSP screener to adequately recalibrate the sample. Consequently, the above quotas were set at the half-way point of the data collection.

At the completion of the experiment (n=2,400) the achieved quotas were the same as the recalibrated quotas with the exception of white ethnic group; 86% of the participants self-reported as belonging to the white ethnic group. This discrepancy in the set quotas and achieved quotas occurred because approximately 1% of participants refused to disclose their ethnicity.

2.2 Data collection

Kantar Public ensured compliance with the Data Protection requirements in the UK, including the UK's General Data Protection Regulation (UK GDPR). In addition, participants were able to opt out of the study; the participants were notified, at the beginning of the study, that they might be exposed to what they could consider to be harmful videos; informed consent was obtained for the collection of sensitive data, such as ethnicity, from the respondents. The consent, questions, and videos were reviewed by Kantar Public's Profiles' Privacy team and Kantar Public's Global Head of Compliance.

2.3 Randomisation

Participants were randomly allocated into one of the experiment's four arms, a control which simulated a VSP interface and three further arms that included interface-based interventions that aimed to increase reporting of harmful content.

To allocate respondents to experimental arms a method of blocked randomisation was used (least-filled quotas). This method ensured that blocks fill at a consistent rate whatever the sample size.

Note that this method of randomisation is frequently used in behavioural economics related studies,³ as well as in clinical trials.⁴

2.4 Incentivisation

Panel participants received 'LifePoints'⁵ on completion of experiments, which can be accrued and exchanged for items in an online catalogue. Respondents received 50 'LifePoints' for completion of this experiment.

2.5 Ethics

The purpose of the experimental environment was to replicate the real-world context, to get as close as possible to actual VSP users' behaviour in the experiment. It would have

³ Dannenberg, A., & Martinsson, P. (2021). Responsibility and prosocial behavior-Experimental evidence on charitable donations by individuals and group representatives. *Journal of Behavioral and Experimental Economics*, 90, 101643.

⁴ For example: <https://onlinelibrary.wiley.com/doi/full/10.5694/j.1326-5377.2002.tb04955.x>

⁵ Further information available at: <https://lifepoints.zendesk.com/hc/en-us>

been difficult, if not impossible, to gain externally valid evidence of the propensity to report legal but potentially harmful content in an experiment that did not expose participants to actual content. However, to reduce the risk from participating in the study, participants were not exposed to content where the risk of psychological harm was likely to be more acute, for example, self-harm or suicide.

Kantar Public's Behavioural Practice team therefore selected legal but potentially 'harmful' content (content that some participants could consider to be harmful) for inclusion in this experiment by:

1. Searching various VSPs for videos that have been made downloadable by their originators so they can be downloaded directly from the website.
2. Searching content that is engaging, recent and relevant to current concerns such as Covid-19 mis/disinformation.
3. Sharing these videos with the Kantar project team and Kantar Public's Profiles' Privacy team (who oversee ethical decisions about any potential impact of research on LifePoints panellists) to confirm that these videos could be considered as harmful by some participants, but that these videos are, nonetheless, legal and acceptable for provision to participants from the LifePoints panel.

This type of content, while still potentially harmful to some participants, was more acceptable for inclusion because of the content's lower impact and greater prevalence (and hence likelihood of being seen "for real" as Ofcom's own research indicates that 70% of VSP users have been exposed to potential online harm on the services they used during the past three months⁶).

Kantar Public follows the Market Research Society's (MRS) code of conduct. The following steps were agreed with the Profiles Privacy Team to mitigate any residual risk from participating in the experiment:

An upfront consent screen at the start of the experiment informed participants that they would be shown some content that could be considered harmful and allowed them to refuse to participate if they did not want to be exposed to this (Figure 1).

Informed consent was also obtained for the collection of sensitive data, such as ethnicity, from the respondents. A debrief screen at the end of the experiment which provided web links to support on any of the potential harms included in the content shown in the experiment. The consent, questions, and videos were reviewed, for approval, by Kantar Public's Profiles' Privacy team and Kantar Public's Global Head of Quality, Information and Security.

⁶ Ofcom, Video-sharing platform usage & experience of harms survey 2021. Accessed on 20/07/21 from https://www.ofcom.org.uk/data/assets/pdf_file/0024/216492/yonder-report-experience-of-potential-harms-vsps.pdf

Figure 1. The consent screen.

We are interested in how people use online video sharing platforms, and particularly how they engage with various types of videos.

The study will take approximately 10 minutes to complete.

All your responses are kept in the strictest confidence and are completely anonymous. Please take the time you need to complete this study at your own speed.

First, we will ask you to interact with a simulated version of a video sharing platform (for example YouTube or Vimeo). Then, we will ask you some questions about what you think, and finally ask you some more information about yourself.

Some of the videos you will see may show violence, extreme views, or harmful content. If you do not wish to proceed, please opt out below.

If you do not wish to proceed, please opt out below.

Are you happy to continue?

Yes

No

2.6 Disclaimer

Kantar Public's Profiles' Privacy team ensured that the research process complied with the relevant regulations, such as the UK GDPR, and best practice (see also section 2.2). Kantar Public also adhered to the Market Research Code of Conduct 2019.

2.7 Attention test⁷

The Profiles panel conducts a range of quality and validation checks when recruiting their panellists.⁸ In addition, and to keep the quality of data high and remove any skimmers who are attempting to get through the experiment as quickly as possible, two attention checks were included in this experiment.

First, any respondent who completed in less than 40% of the median completion time for all respondents was removed. Second, a specific attention check was used. This question asked participants to: "Please select the 'green' colour option below. We are asking this for quality control reasons to check you are paying attention to the questions in the survey."

The response options were:



The total drop-out rate due to failing the attention checks and completing the study too quickly, over the whole sample, was approximately 8%. The drop-out rate due to failing the second attention check was 5%.

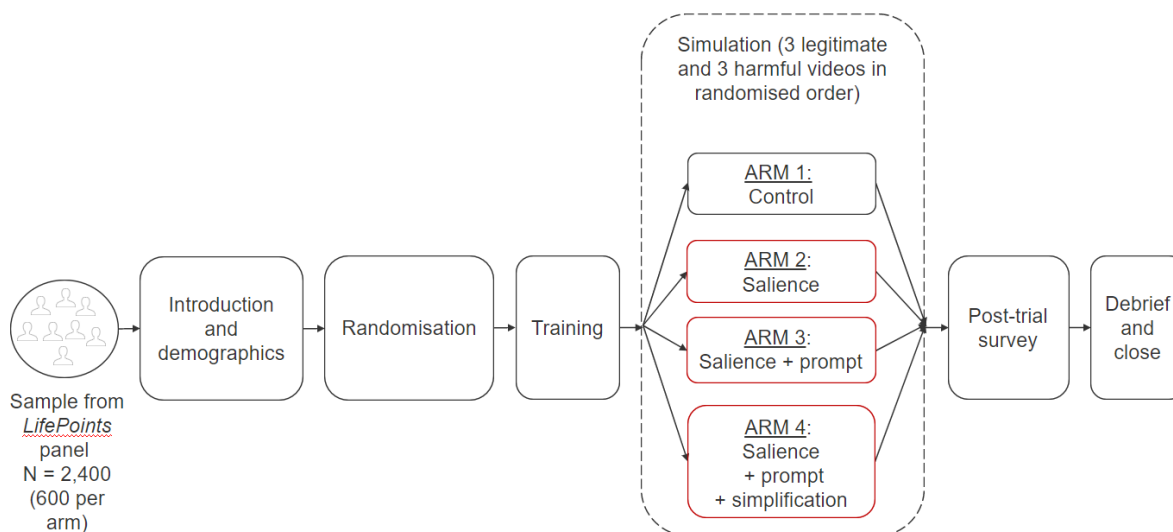
⁷ The attention check that we intended to use throughout the recruitment period was leading to unacceptably high levels of drop out. To deliver on time and within budget we changed the attention check question to the question described in section 2.6. This was a deviation from the attention check originally specified in the trial protocol. However, the new attention check was arguably more in line with the visual nature of the experiment.

⁸ More information available at <https://www.kantar.com/expertise/research-services/panels-and-audiences/lifepoints-research-panel>

To ensure that there were no unforeseen issues with the experimental design and script, an initial soft launch involving 10% of participants was conducted. During the soft launch the following were monitored: the drop-off rate, time to finish the experiment, view time of each of the videos, and the quotas.

3. Trial design and flow

Figure 2. Trial design and flow.



3.1 Introduction and participant consent

Participants were first presented with an introduction screen thanking them for taking part in the study and outlining what it would involve. As per Figure 1 the introduction screen contained a disclaimer about the inclusion of potentially harmful content that read “*Some of the videos you will see may show violence, extreme views, or harmful content. If you do not wish to proceed, please opt out below.*”. **An opt-out button was provided at this point.**

There was also a debrief screen at the end of the experiment which provided links to support on any of the content shown in the experiment.

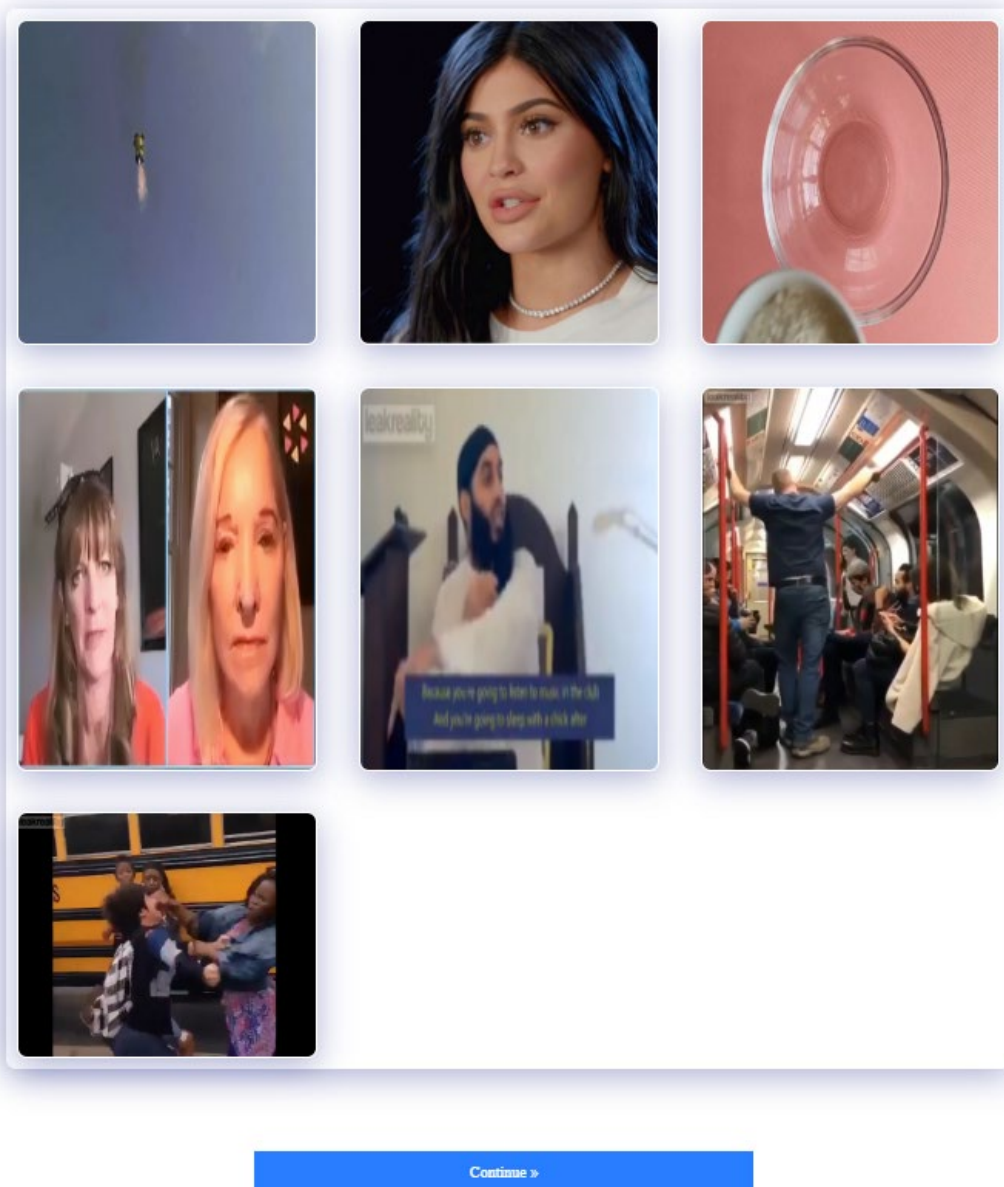
Figure 3. Debrief screen.

Debrief

Thank you for taking part in this study. You may or may not have noticed, while you were going through the study, that some of the videos shown to you could have been classified as 'harmful'.

These videos were sourced and shown to you for the sole purpose of collecting information on the way the UK population reports harmful videos online.

We strongly recommend that you report harmful content to the platform it is on, whenever you see it. For further guidance, and support, refer to the UK Safer Internet Centre (<https://saferinternet.org.uk/report-harmful-content>).



3.2 Demographics and VSP use screener

On entry to the trial, participants were asked demographic questions so that recruitment could be monitored against quotas of interest (as per section 2.1 these were age, gender, socioeconomic background, location, and ethnicity).

Following the demographic questions participants were screened for VSP use by asking which of 10 common video sharing platforms (YouTube, Facebook, Instagram, Snapchat, TikTok, Twitch, Onlyfans, Vimeo, Bitchute, Fruitlab) they had used within the past 12

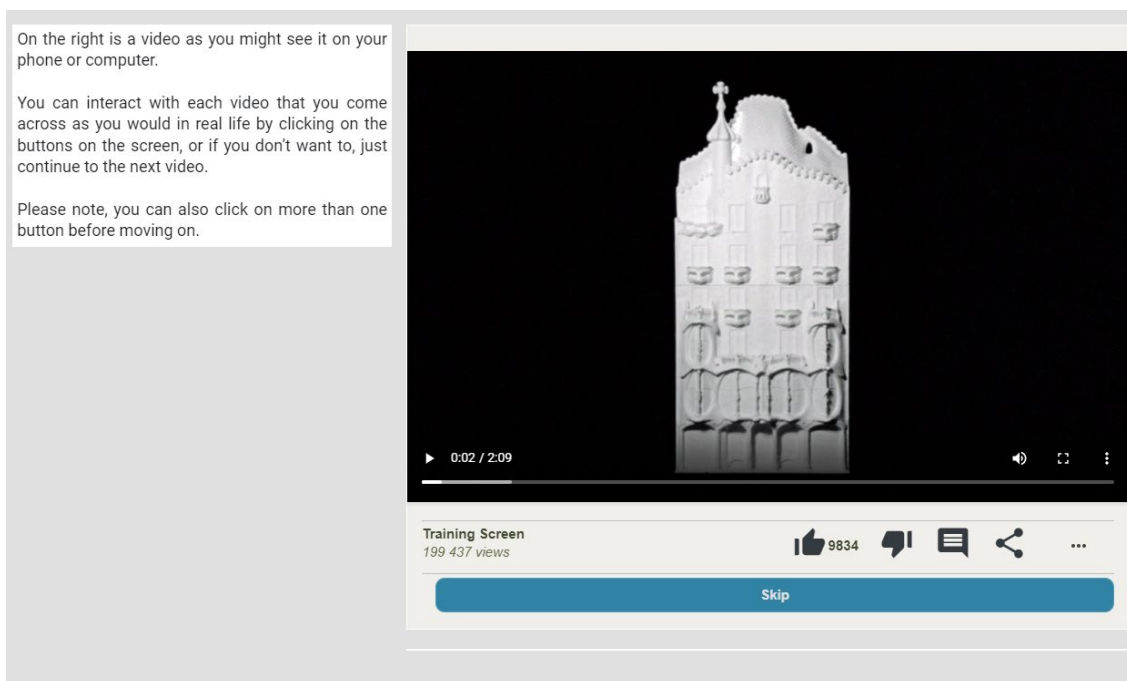
months. Potential participants were screened out if they answered “I haven’t used any video sharing platforms in the past 12 months”.

3.3 Training stage

Once participants had confirmed their demographics, they were randomly allocated into one of the experimental blocks, and the interface that they would be using in the experiment was introduced. At this stage participants had the opportunity to interact with the interface they had been allocated to. First, participants saw a static screenshot of the interface they had been randomly allocated to with instructions for how they could use the buttons available and a short description of how the experiment would proceed.

The intervention arm interfaces were all variations on the generic VSP interface presented in the control arm that incorporated features that are common to many platforms but without any specific branding. After users had seen the labelled screenshot, they were shown a training video that they were able to interact with by choosing to react (like/dislike⁹), comment or share (indicated by adding in comments or pressing the share button in the interface), report (participants were trained on the interface they had been assigned to) or skip past to the next piece of content (see Figure 4).

Figure 4. Control interface training screen.



Participants were able to ‘play’ with this training screen until they were familiar with how the interface worked. The video content shown to the participants at the training stage was selected in the same way as the videos for the main experiment part (see section 3.4). The video content for the training stage was unlikely to be classed as harmful by any participant, as it did not contain potentially harmful content (“This is not a house”:

⁹ Note that each video already had a number of likes and views when participants saw the video. The counts of likes for videos was created using random generation for the Poisson distribution with $n = 10000$. In other words, each video had approximately 10000 likes. Views were generated in the same fashion, but n was 200000. Overall, the number of likes was approximately 5% of the number of views.

<https://vimeo.com/555252697>). After interacting with the training screen, the participants were able to move on to the main experimental section.

3.4 Main experiment

Since reporting content is a rare occurrence, this experimental design aimed to first increase the number of reports of the potentially harmful content (Arms 2 and 3) and then increase the accuracy of those reports by making it easier to report content (Arm 4).

In the main experiment, participants were exposed to six pieces of video content presented in a random order within the simulated VSP interface. Three pieces of content were neutral, and three were legal but potentially harmful. The aim was to encourage participants to report the legal but potentially harmful content.

All videos were chosen, or trimmed, to be engaging in the first 20-45 seconds to hold participant attention. In addition, recent and relevant potentially harmful content was prioritised for the same reason.

Video content:

Neutral One: Vegan Matcha Pancakes: <https://vimeo.com/248973738>

Neutral Two: Blue Origin Booster Landing: <https://vimeo.com/577391557>

Neutral Three: Celebrity Breakups: <https://vimeo.com/247515393>

Potentially Harmful One: Covid-19 Vaccine Misinformation (trimmed): <https://vimeo.com/496630435>

Potentially Harmful Two: Tube Racism Fight: <https://leakreality.com/video/25086/repost-fight-breaks-out-after-british-man-racially-harass-asian-woman>

Potentially Harmful Three: Homophobic (trimmed): <https://leakreality.com/video/26960/uk-muslim-cleric-music-makes-you-gay>

3.5 Post-trial survey

After the experiment, all participants were prompted to report a new piece of harmful video content:

Kids Fighting: <https://leakreality.com/video/9236/never-relax>

First, participants were prompted to report this video using the interface they had been randomly assigned to. The aim was to investigate if participants could accurately categorise harmful content when prompted. Participants still had the option to skip, but if they chose to skip then were asked to indicate why:

- I don't know how to report
- I don't want to report
- Reporting takes too much time
- Other

A further question was then asked to check whether participants actually considered this additional potentially harmful content 'harmful': "To what extent do you disagree or agree that the below video could cause harm?."

The response options were:

"Strongly disagree; Disagree; Slightly disagree; Neither agree nor disagree; Slightly agree; Agree; Strongly agree".

Finally, participants were asked survey questions to understand their internet and social media usage as well as their attitudes toward moderation of VSP content and its effectiveness.

The same questions were used for all trial arms to ensure that results were comparable across all arms. Responses to the questions were used as attitudinal **secondary outcomes** that also constituted as **descriptive metrics** in the study.

4. Interventions

There were four arms of this experiment, each outlined below. These were selected and developed selected in collaboration with Ofcom:

1. *Arm 1 – Control*: The control arm included an interface that is a generic version of a VSP. This meant that the reporting option was available as a secondary action behind an ellipsis below the video. The option to report was therefore only visible after a click on the ellipsis.
2. *Arm 2 – Salience*:¹⁰ If users cannot see the reporting mechanism, then they are less likely to know that it is available. In this arm the report action, which is normally hidden behind an ellipsis in most VSP interfaces, was brought forward to the same level of the visual hierarchy as other primary actions, such as “like” and “share”. and the reporting action was represented by a flag icon. Reporting was therefore visible without any interaction from the user. Visual salience is a key tool for attracting attention as the neural processes that allow the selection of items because we are paying attention to them are often not subject to conscious control.¹¹ In addition, the user journey is simplified and shortened as clicking on the flag icon opens a pop up window taking users straight to the “select a reason” stream of the reporting flow.

Hypothesis 1: When the report button is visible at the top-level of the visual hierarchy participants will be more likely to start a report compared to the control.

3. *Arm 3 – Salience + prompt*: If a user believes there is something wrong with a video, the minimum effort required to express that is by clicking the thumbs down button. Slightly more effort is needed to comment on the video. Reporting a video may feel like it requires more effort than both these options. To test whether dislikes and comments were perhaps being used as an alternative to reporting harmful content, and in addition to the changes in the salience of the report action from Arm 2, users were provided with a salient,¹² low friction¹³ prompt to report harmful content after clicking the thumbs down or comment buttons. This prompt was a message that popped up asking participants if they also wanted to report the post for violating community guidelines.

Hypothesis 2: If a user dislikes or comments on content as a low-effort alternative to reporting, that action can be redirected to actual reporting, leading more participants to report compared to both the control and Arm 2 - Salience.

4. *Arm 4 - Salience + prompt + simplification*: This arm of the experiment was an intervention designed to test all previous behavioural interventions and simplify the reporting process. Specifically, the salience approaches from arms 2 and 3 were combined alongside an intervention to simplify accurate reporting of the type of harmful content. This intervention aimed to decrease choice overload¹⁴ (which can

¹⁰ Desimone, R., & Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annual review of neuroscience*, 18(1), 193-222.

¹¹ Desimone, R., & Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annual review of neuroscience*, 18(1), 193-222.

¹² Desimone, R., & Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annual review of neuroscience*, 18(1), 193-222.

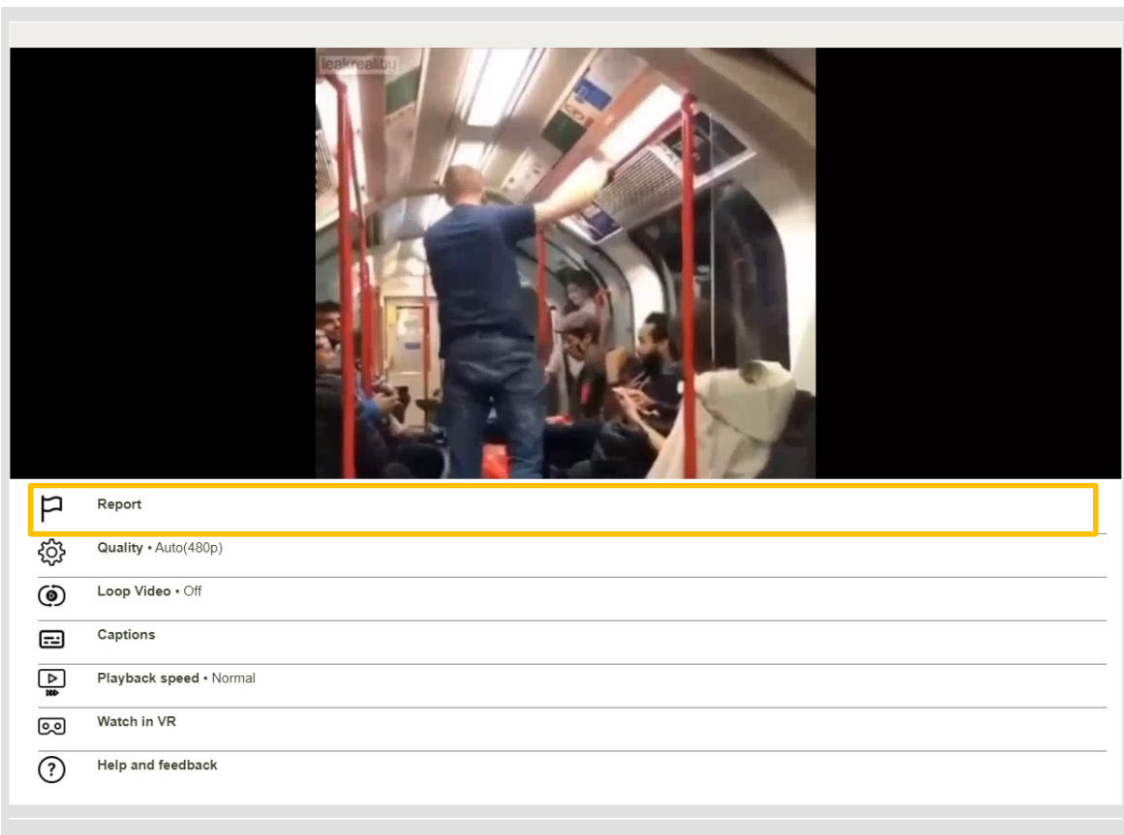
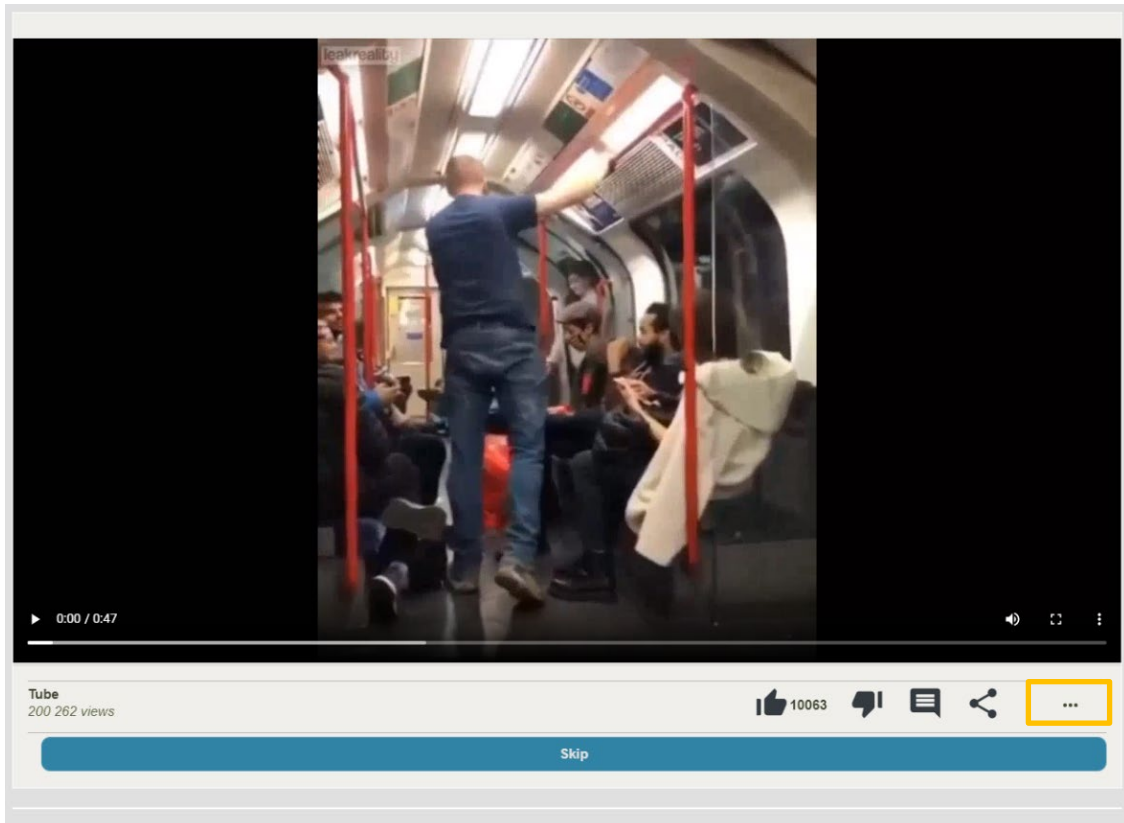
¹³ Bettinger, E. P., Long, B. T., Oreopoulos, P. & Sanbonmatsu, L. (2012). The role of application assistance and information in college decisions: Results from the H&R block FAFSA experiment, *The Quarterly Journal of Economics*, 127(3), 1205-1242.

¹⁴ Chernev, A., Böckenholt, U., & Goodman, J. (2015). Choice overload: A conceptual review and meta-analysis. *Journal of Consumer Psychology*, 25(2), 333-358.

result in fewer choices being made and a stronger default effect) by providing simpler headers for a shorter list of types of harmful content. The expectation of this intervention was that users would be more likely to report the potentially harmful videos and provide more accurate information when reporting, compared to other intervention arms.

Hypothesis 3: Participants will be more likely to complete a report when the process involves simpler choices compared to arms 1, 2 and 3.

Figure 5. Arm 1 – Control (reporting flow).



Report

Select a reason

- Animal cruelty >
- Pornography or nudity >
- Illegal activities and regular goods >
- Intellectual property infringement >
- Spam >
- Violent and Graphic content >
- Misleading information >
- Harm to children >
- Suicide, self-harm and dangerous acts >
- Hate speech >
- Harassment or bullying >
- Dangerous organisations and individuals >
- Other >

Report


Select a reason

- Suicide >
- Self-harm >
- Dangerous acts >

Additional comments

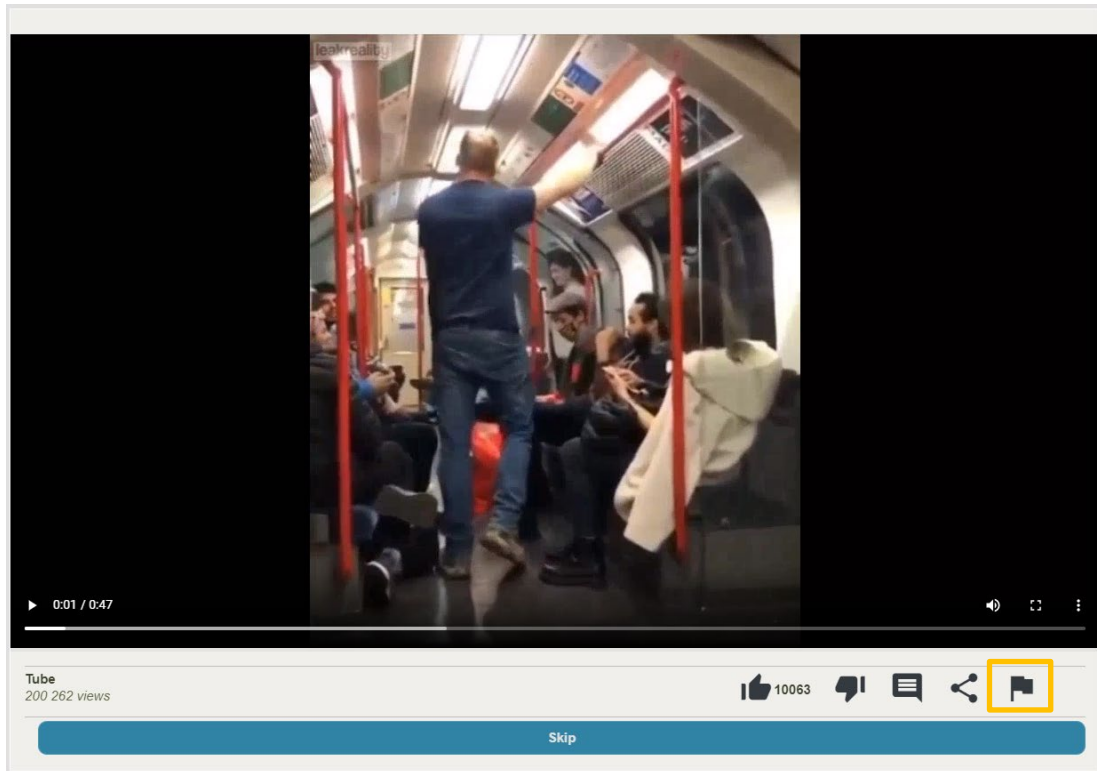
0/500 characters

Report



Thanks for reporting

Figure 6. Arm 2 – Saliency (reporting flow).



Report

Select a reason

- Animal cruelty >
- Pornography or nudity >
- Illegal activities and regular goods >
- Intellectual property infringement >
- Spam >
- Violent and Graphic content >
- Misleading information >
- Harm to children >
- Suicide, self-harm and dangerous acts >
- Hate speech >
- Harassment or bullying >
- Dangerous organisations and individuals >
- Other >

Report

Select a reason

Suicide >


Self-harm >

Dangerous acts >

Additional comments

0/500 characters

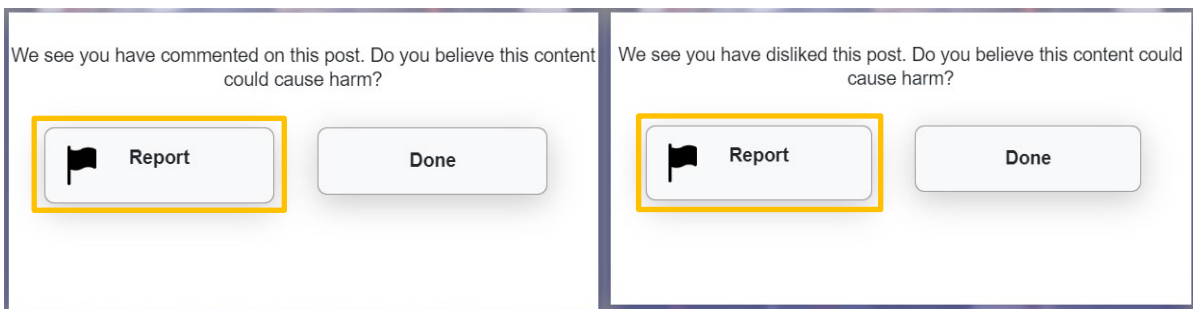
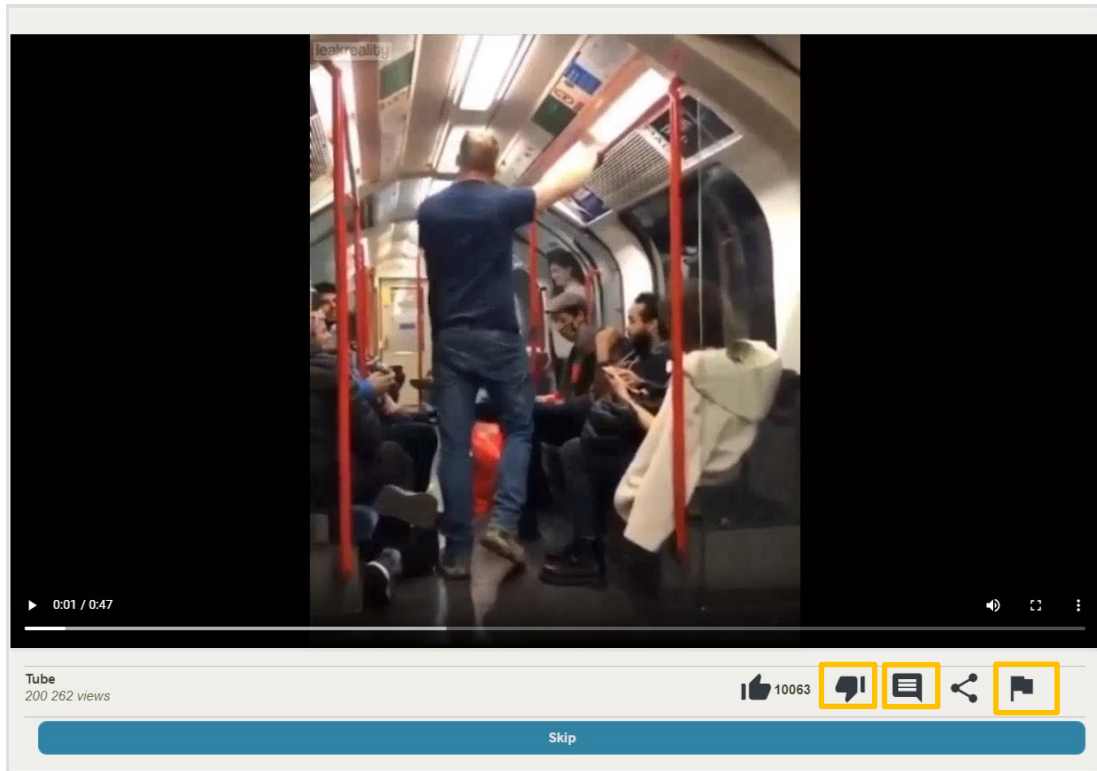
Report



Thanks for reporting

Done

Figure 7. Arm 3 – Salience + prompt (reporting flow).



Report

Select a reason

- Animal cruelty >
- Pornography or nudity >
- Illegal activities and regular goods >
- Intellectual property infringement >
- Spam >
- Violent and Graphic content >
- Misleading information >
- Harm to children >
- Suicide, self-harm and dangerous acts >
- Hate speech >
- Harassment or bullying >
- Dangerous organisations and individuals >
- Other >

Report


Select a reason

- Suicide >
- Self-harm >
- Dangerous acts >

Additional comments

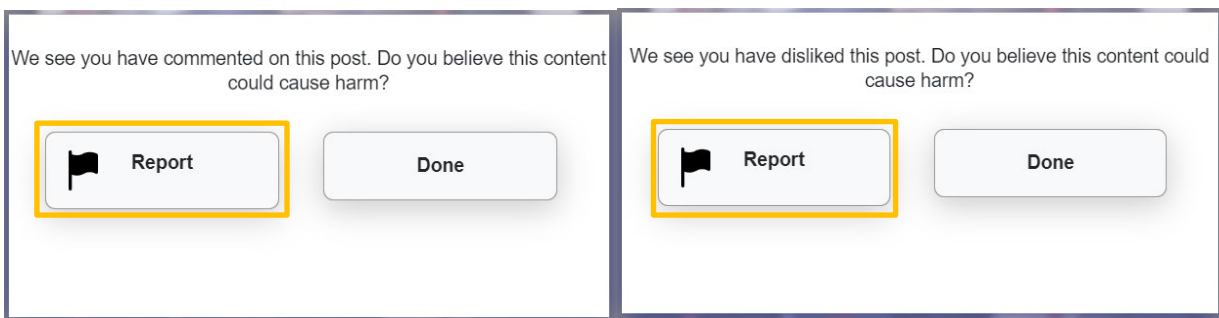
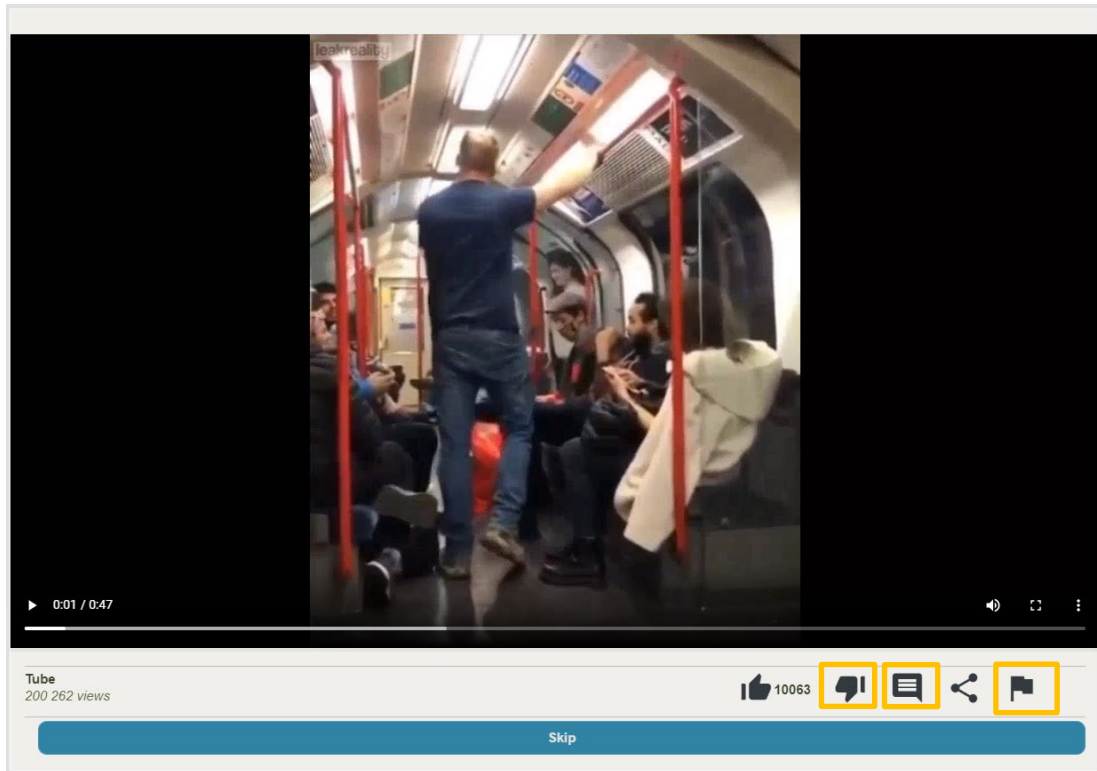
0/500 characters

Report



Thanks for reporting

Figure 8. Arm 4 – Salience + prompt + simplification (reporting flow).



Select a reason

- Violence >
 - Bullying
 - Abuse
 - Harassment
 - Inciting violence
 - Terrorism
 - Graphic content
- False or misleading information >
 - Conspiracy theories
 - Science denial
- Hate speech >
 - Racist
 - Sexist
 - Ableist
- Self harm >
 - Suicide
 - Dangerous acts
 - Intention to self-harm
- Nudity >
 - Pornography
 - Child sex abuse
- Spam >
- Other >


Cancel

Additional comments

0/500 characters

Cancel
Submit

Report



Thanks for reporting

Done

5. Outcomes

5.1 Primary outcome

In this study, the number of harmful videos (out of three) a participant completed a report for were measured. Based on Kantar Public’s experience of running similar online experiments, ‘reporting’ was hypothesised to be a relatively rare behaviour; therefore, the primary outcome was the sum of reported harmful videos. In other words, the primary outcome, in this study, was the number of pieces of harmful content for which a report was submitted (ranging from 0 to 3) (see also section 6).

5.2 Secondary outcomes

The secondary outcomes sought to establish other effects associated with exposure to the interventions. As for the primary outcome, each of these were created by summing the number of posts for which each of the behaviours occurred.

Primary	Number of submitted reports of harmful content when viewing harmful videos
Secondary	<p>Number of submitted reports of neutral content that should not be flagged (over-reporting).</p> <p>Reports of harmful content started but not finished and neutral content started but not finished.</p> <p>Number of skips of harmful content, and of neutral content.</p> <p>Number of submitted reports of harmful content that accurately categorise the harmful content according to the options available.</p> <p>Response to survey question that prompts participants to report harmful content. Specifically, the proportion that accurately categorise the type of harmful content according to the options available.</p>
Descriptive metrics	<p>Number of likes, dislikes, shares, and comments on harmful content and neutral videos.</p> <p>Length of view for both types of video content.</p> <p>Belief that reports will make a difference.</p> <p>Intent to report again in the future.</p> <p>Confidence in VSP reporting mechanisms.</p> <p>Belief that a given content was actually harmful/worth reporting.</p>

6. Statistical methods and analysis

6.1 Statistical methods

Primary analysis

As noted above, the primary outcome for the analysis was the number of pieces of harmful content for which a report was submitted (ranging from 0 to 3). This approach was chosen over alternatives as, based on Kantar Public's previous work, reporting is a rare behaviour in simulated social media environments: the likelihood of a pattern of correlation within individuals' behaviour was therefore expected to be low.

As such, an Analysis of Variance (ANOVA) model was intended to be used to detect significant differences in the mean number of reported pieces of harmful content between the treatment and control arms (using post-hoc tests). The equation for the ANOVA is:

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij},$$

where $\varepsilon_{ij} \sim N(0, \sigma^2)$, for $i = 1, \dots, I; j = 1, \dots, J$ ($I = 4; J =$ number of people in each i^{th} group).

However, since the data can be treated as a count, we also considered modelling changes in the primary outcome variable using a Poisson regression model. This approach was proposed if based on the response, the assumptions of ANOVA were not met. In addition, we noted that an additional complexity was imposed by the data in the presence of many zeros. Given the potential to see a relatively large number of zeros, it made sense to consider a zero-inflated Poisson regression model because using a standard count may lead to bias when there are many zeros in the outcome variable.

A zero-inflated model is based on a zero-inflated distribution that allows for frequent zero-valued observations. The zero-inflated Poisson model mixes two processes: one that generates zeros, and one that generates counts, some of which could also be zero (Poisson process).¹⁵

The interpretation of the estimates produced by the zero-inflation component of such a model may seem counterintuitive. This is because the zero-inflation component of the zero-inflated Poisson model predicts the probability of observing a zero count from the point mass component.

In addition to comparisons between the treatment and control arms, Kantar Public intended to test between the performance of treatment arms: for that reason, the Bonferroni correction to maintain the family-wise error rate was utilised.

Secondary analysis

Three of the secondary outcomes - number of submitted reports of neutral content that should not be flagged (over-reporting); reports of harmful content started but not finished and neutral content started but not finished; and number of skips of harmful content, and of

¹⁵ For a more detailed description, refer to: Lambert, D. (1992). Zero-Inflated Poisson Regression, with an Application to Defects in Manufacturing. *Technometrics*, 34(1), 1–14. <https://doi.org/10.2307/1269547>

neutral content – were analysed using the approach outlined above (ANOVA, or zero-inflated Poisson regression models) to establish other effects (e.g., spillover and halo effects) associated with exposure to the interventions.

In addition to the approach outlined in the paragraphs above, the impact of domain (broad topic area) of the type of harm on reporting behaviour was investigated. To do this, the use of logistic mixed-effects model with the content's domain coded into the model was explored. Mixed logistic models are appropriate in this specific context, because the data are binary, and there were several observations per participant. The intention was to minimise aggregating data to lower the Type I error rate (Type I error is when you spuriously find a significant effect, when there is no significant effect), and to better approximate the underlying distribution of the probability of reporting. In this instance, mixed logistic models were suggested as a secondary approach because there was uncertainty with regard to whether the level of response would allow for this form of analysis.

As part of sensitivity analysis, the number of submitted reports of harmful content that accurately categorised the potentially harmful content according to the options available was used as an outcome. The aim of this sensitivity check was to try to replicate the results of the primary analysis using this outcome, to see whether the effects of the intervention were sensitive to the choice of the data. Last, the proportions of participants that accurately categorised (according to accepted categories for each video) the type of harmful content according to the options available were analysed using a Z-Test (a test of proportions) with Bonferroni corrected p -values.

6.2 Statistical power

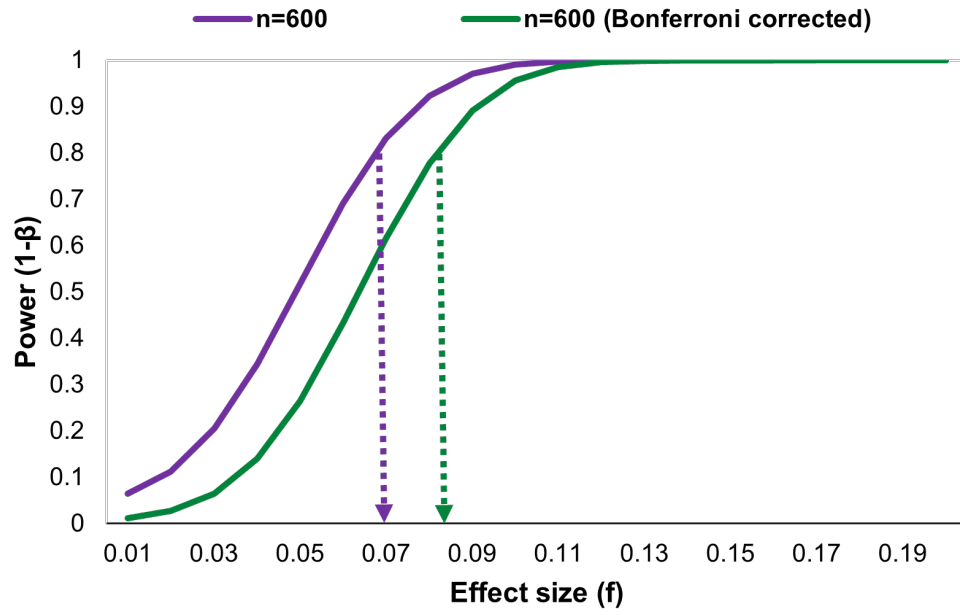
Power calculations were conducted to illustrate the relationship between the sample size, effect size, significance level, statistical power and type of statistical test. In this case, $n=600$ participants per arm were aimed for, leaving a total sample size of $n=2,400$.

The assumptions made included:

- application of ANOVA;
- $n=6$ post-hoc tests with Bonferroni correction;
- $\alpha = 0.05/6$; and
- $n=600$ participants per arm.

The expectation was that the trial would be sufficiently powered to detect small effects, such as those observed in previous similar Kantar Public experiments (see Figure 9 for the range of expected effects below). Figure 9 shows the power to detect a statistically significant effect, depending on the effect size, assuming 600 participants per trial arm. Purple line shows the unadjusted power estimates whereas the green line shows adjusted (Bonferroni corrected) power estimates. At the conventional threshold of 80% power, the unadjusted smallest detectable effect is estimated to be Cohen's $f = 0.07$, whereas the adjusted smallest detectable effect is thought to be between Cohen's $f = 0.08$ and Cohen's $f = 0.09$. These estimates can be interpreted as saying that we are able to detect small effect sizes, 80% of the time, if they are truly there.

Figure 9. The range of expected effects.



7. Results

7.1 Randomisation and balance between arms

The randomisation process resulted in relatively balanced split of participants according to demographic variables within each treatment arm. For example, the median age of participants across arms ranged from 40 to 41.

For additional descriptive statistics, please refer to the appendix (section 9).

7.2 Primary outcome

The grey bars in Figure 10 show the observed distribution of counts (square rooted to see small deviations from the expected count) by the total count of the incidence of reports. The red line, and dots, are the expected counts based on our zero-inflated model (see section 7.1). Figure 10 shows two things. First, that the observed number of counts of 0 was disproportionately larger compared to the other counts (1, 2, or 3). Second, that our model predicted the observed counts relatively well (because the red line, and dots, fit the pattern visualised by the histogram).

Figure 10. The distribution of counts (square rooted) by the total count of the incidence of reports.

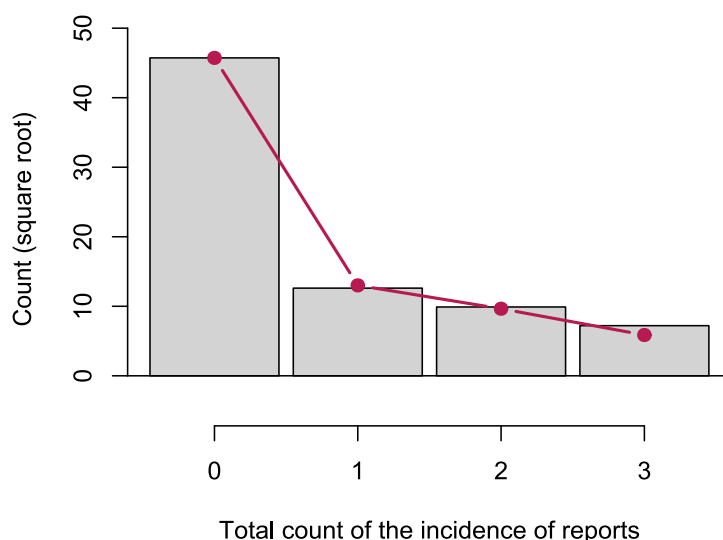
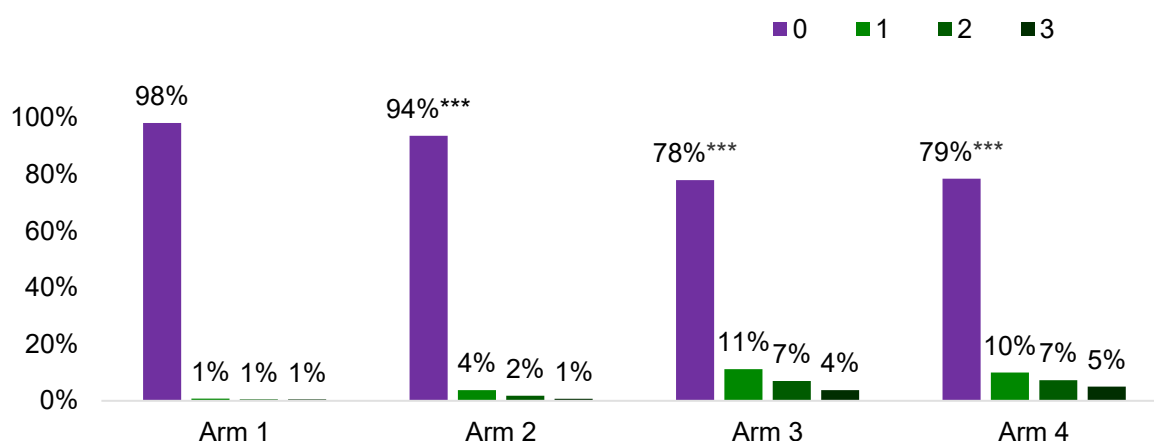


Figure 11 shows the percentage of participants reporting legal but potentially harmful video content by arm. The three stars indicate significant difference ($p < 0.001$) between a particular arm and the control arm (Arm 1). All intervention arms are significantly different from the control arm. Specifically, the percentage of participants reporting legal but potentially harmful content is higher in the intervention arms than in the control arm.

(Note that Figure 11 shows descriptive statistics, rather than any model predictions. Significance of the differences was determined using the zero-inflated Poisson regression model described in section 6.1, and the analysis is described in more detail below.)

Figure 11. The percentage of participants reporting legal but potentially harmful video content by arm.



The primary outcome variable was analysed using a zero-inflated Poisson regression model. As mentioned in section 6.1, the interpretation of the estimates produced by the zero-inflation component of such a model may seem counterintuitive. This is because the zero-inflation component of the zero-inflated Poisson model predicts the log odds of observing a zero count from the point mass component.

All the interventions were found to have a significant effect on the number of completed reports when watching potentially harmful content versus the control arm (Table 2). Specifically, all the interventions decreased the log odds of not reporting potentially harmful content compared to Arm 1.¹⁶

The log odds estimates in Table 2 can be interpreted as: being in Arm 2 decreased the odds of not reporting by 79% ($\exp(-1.5673) = 0.21$; $(0.21-1)*100 = -79$) compared to being in Arm 1; being in Arm 3 decreased the odds of not reporting by 95% compared to being in Arm 1; and being in Arm 4 decreased the odds of not reporting by 94% compared to being in Arm 1. The direction of these effects was not sensitive to the inclusion of age and order as covariates, and the estimates – as well as the standard errors – also did not change substantially.

Table 2. Model based estimates (zero-inflated Poisson regression model).

	Estimate	SE	z-value	p
Intercept	3.6717	0.3411	10.766	< 0.001
Arm 2	-1.5673	0.4129	-3.796	< 0.001
Arm 3	-2.9434	0.3683	-7.991	< 0.001
Arm 4	-2.8928	0.3683	-7.855	< 0.001

Note: Arm 1 is the reference level other arms are compared against.

¹⁶ In this model, and in subsequent models, the standard errors were derived using the Hessian matrix returned by Nelder-Mead optimisation algorithm. Overdispersion in the data, characterised here by the excess of zeros, was accounted for by fitting zero-inflated Poisson models to the data.

Next, the arms were compared against each other. Table 3 shows the differences between the arms, controlling for multiple comparisons.¹⁷ Being in Arm 3 decreased the odds of not reporting by 75% compared to being in Arm 2; further, being in Arm 4 decreased the odds of not reporting by 73% compared to being in Arm 2. There was no significant difference between Arms 3 and 4 in the odds of not reporting.

Table 3. Bonferroni-corrected multiple comparisons' estimates.

	Estimate	SE	z-value	p
Arm 1 vs 2	-1.5673	0.4129	-3.796	< 0.001
Arm 1 vs 3	-2.9434	0.3683	-7.991	< 0.001
Arm 1 vs 4	-2.8928	0.3683	-7.855	< 0.001
Arm 2 vs 3	-1.3761	0.2711	-5.077	< 0.001
Arm 2 vs 4	-1.3255	0.2710	-4.891	< 0.001
Arm 3 vs 4	0.0506	0.1966	0.258	1.000*

*Approximately (rounding error).

Figure 12 shows the model-predicted count of reports by arm, alongside the observed data. Note that the red bars are made of individual data points (red dots), so they are clusters of observed data.

Arms 2, 3, and 4, are seen to have a greater number of observed counts of reports than Arm 1. Consequently, the model predictions for the counts of reports are higher in these arms compared to Arm 1. The 95% confidence intervals surrounding indicate the uncertainty surrounding the point-based estimates of the model.

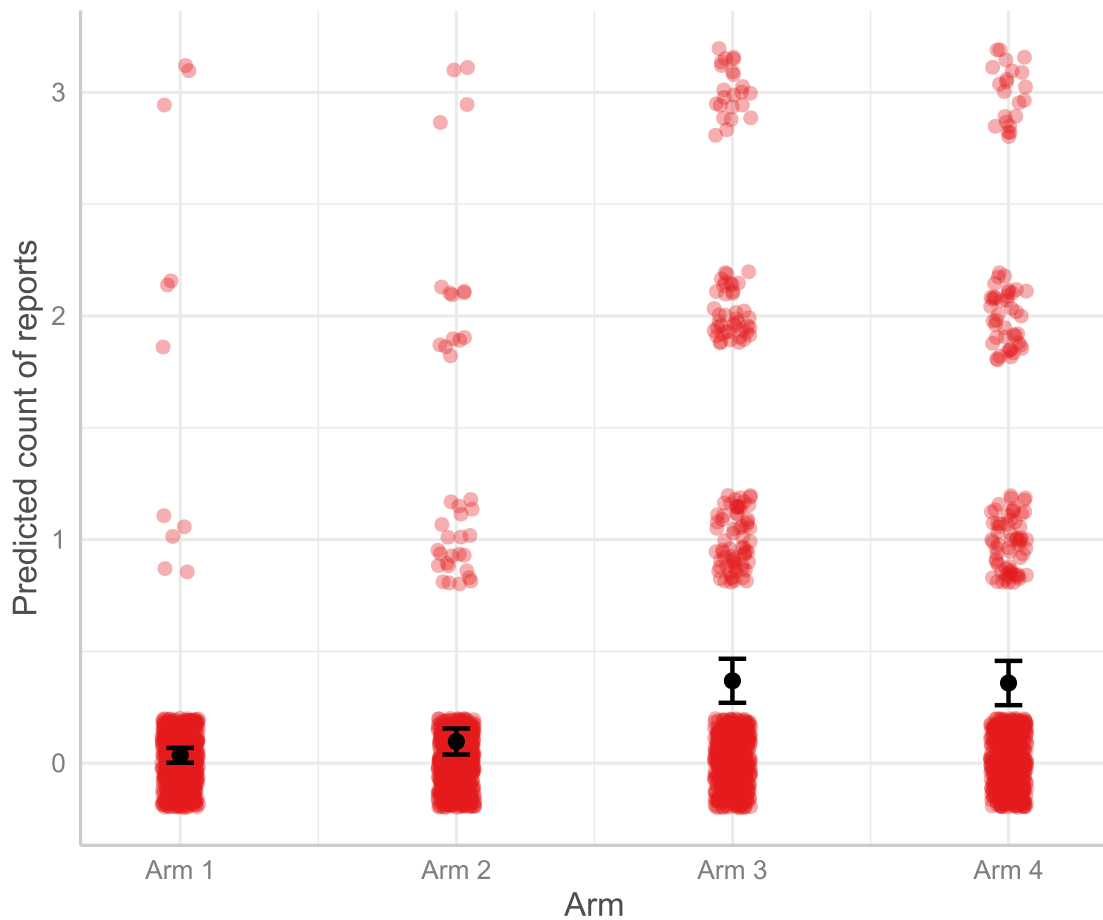
Note that confidence intervals can overlap with each other in Figure 12 (as is the case when looking at Arm 1 and Arm 2), while there still being a significant difference. This is because the difference between the two means is not comparable to the difference in confidence intervals of these means.^{18,19} It is possible to have a precise estimate of the difference between different arms (see Tables 2 and 3) while having a more uncertain estimate of the effect of a particular arm itself (Figure 12).

¹⁷ Note that the probability of committing false statistical inferences increases when more than one statistical inference is simultaneously tested: this is known as the multiple comparisons problem. To address this problem, an adjustment for multiple comparisons was made using the Bonferroni correction to control the family-wise error rate.

¹⁸ Schenker, N., & Gentleman, J. F. (2001). On Judging the Significance of Differences by Examining the Overlap Between Confidence Intervals. *The American Statistician*, 55(3), 182-186.

¹⁹ Wolfe, R., & Hanley, J. (2002). If we're so different, why do we keep overlapping? When 1 plus 1 doesn't make 2. *Canadian Medical Association Journal*, 166(1), 65-66.

Figure 12. The zero-inflated Poisson regression model predicted count of reports by arm.



7.3 Secondary outcomes

7.3.1 Over-reporting

No reliable analysis could have been conducted between the arms on the differences in the number of completed reports when watching neutral content. This is because the count of overreporting across these arms was very low (Arm 1 = 2; Arm 2 = 2; Arm 3 = 2; Arm 4 = 12).

Consequently, the number of submitted reports of neutral content appears higher in Arm 4 compared to other arms, but this observation has a high chance of being a Type I error (spurious effect), because the number of over-reports is so low any model predictions will be biased due to the very high levels of uncertainty associated with the estimates. Thus, no inference on over-reporting between arms should be made as it is questionable whether this pattern would replicate in a real-world context.

7.3.2 Started but not finished

No reliable analysis could have been conducted on the differences in the number of started but not finished reports of harmful content because the count of such reports was too low (Arm 1 = 0; Arm 2 = 0; Arm 3 = 0; Arm 4 = 5). In addition, the count of started but not finished reports of neutral content was 0 in all arms.

7.3.3 Skips

No significant differences were found in the number of skips of harmful content between arms (Arm 1 = 389; Arm 2 = 398; Arm 3 = 400; Arm 4 = 410), and of neutral content (Arm 1 = 412; Arm 2 = 413; Arm 3 = 425; Arm 4 = 448).

7.3.4 Number of accurate reports of potentially harmful content

Using the zero-inflated component of the zero-inflated Poisson regression model, the interventions in Arms 3 and 4 were found to have a significant effect on the number of completed accurate reports, when watching potentially harmful content, versus the control arm (Table 4). The coefficients in Table 4 can be interpreted as: Being in Arm 3 decreased the odds of inaccurately reporting by 90% ($\exp(-2.3422) = 0.10$; $(0.10-1)*100 = -90$) compared to being in Arm 1; Being in Arm 4 decreased the odds of inaccurately reporting by 93% compared to being in Arm 1.

Table 4. Model based estimates (zero-inflated Poisson regression model).

	Estimate	SE	z-value	p
Intercept	3.3773	0.5386	6.270	< 0.001
Arm 2	-1.1593	0.6236	-1.859	0.063
Arm 3	-2.3422	0.5749	-4.074	< 0.001
Arm 4	-2.7070	0.5730	-4.724	< 0.001

Note: Arm 1 is the reference level other arms are compared against.

Table 5 shows that Arms 3 and 4 led to significantly less inaccurate reports of potentially harmful content than Arms 1 and 2. There was no significant difference in the odds of inaccurate reporting between Arms 3 and 4.

Table 5. Bonferroni-corrected multiple comparisons' estimates.

	Estimate	SE	z-value	p
Arm 1 vs 2	-1.1593	0.6236	-1.859	0.378
Arm 1 vs 3	-2.3422	0.5749	-4.074	< 0.001
Arm 1 vs 4	-2.7070	0.5730	-4.724	< 0.001
Arm 2 vs 3	-1.1829	0.3730	-3.171	0.009
Arm 2 vs 4	-1.5477	0.3701	-4.182	< 0.001
Arm 3 vs 4	-0.3648	0.2804	-1.301	1.000

*Approximately (rounding error).

7.3.5 Accuracy of reporting (survey question)

To assess the accuracy of reporting we used a logistic regression model. We used this model because every person was asked to submit a report of a potentially harmful video, and only the participants who had submitted the report accurately were marked as having submitted the report. Thus, the outcome variable was binary, the report could either have been accurately submitted (1) or not accurately submitted (0).²⁰

²⁰ Note that mixed logistic regression models require more than one observation per participant and each participant had only one observation relating to this outcome variable. Consequently, they could not have been used with this outcome.

Table 6 shows that all the interventions were found to have a significant effect on the likelihood of submitting an accurate report when watching the follow-up video, versus the control arm. Specifically, all the interventions increased the odds of accurately reporting potentially harmful content compared to Arm 1.

The coefficients in Table 6 can be interpreted as: being in Arm 2 increased the odds of accurately reporting by 78% ($\exp(0.5744) = 1.78$; $(1.78-1)*100 = 78$) compared to being in Arm 1; being in Arm 3 increased the odds of accurately reporting by 118% compared to being in Arm 1; and being in Arm 4 increased the odds of accurately reporting by 599% compared to being in Arm 1.

Table 6. Model based estimates (logistic regression model).

	Estimate	SE	z-value	p
Intercept	-2.0244	0.1272	-15.919	< 0.001
Arm 2	0.5744	0.1643	3.495	< 0.001
Arm 3	0.7780	0.1606	4.846	< 0.001
Arm 4	1.9443	0.1512	12.863	< 0.001

Note: Arm 1 is the reference level other arms are compared against.

Table 7 shows that being in Arm 4 led to a significantly higher likelihood of submitting accurate reports of the follow up video compared to being in Arms 1, 2, and 3. There was no significant difference in the likelihood of submitting an accurate report of the follow-up video between being in Arms 2 and 3.

Table 7. Bonferroni-corrected multiple comparisons' estimates.

	Estimate	SE	z-value	p
Arm 1 vs 2	0.5744	0.1643	3.495	0.003
Arm 1 vs 3	0.7780	0.1606	4.846	< 0.001
Arm 1 vs 4	1.9443	0.1512	12.863	< 0.001
Arm 2 vs 3	0.2037	0.1430	1.425	0.926
Arm 2 vs 4	1.3700	0.1323	10.354	< 0.001
Arm 3 vs 4	1.1663	0.1276	9.139	< 0.001

7.3.6 Survey questions (descriptive statistics)

As part of the online experiment, we also asked participants survey questions regarding their attitudes to reporting of potentially harmful videos. Figure 13 shows the distribution of participant responses to the following question: "Reporting content on video sharing platforms makes a difference. To what extent do you agree/disagree with the following statements?". (Note that all bars are grey because no formal comparison between groups is being made). The median response was 5 (slightly agree), indicating that the majority of respondents agreed, to some extent, that reporting content on video sharing platforms made a difference.

Figure 13. A histogram of participant responses to "making a difference" question (n = 2,234), excluding "don't know" responses. Ratings range from 1 to 7 and the median is 5. The responses are skewed towards 7.

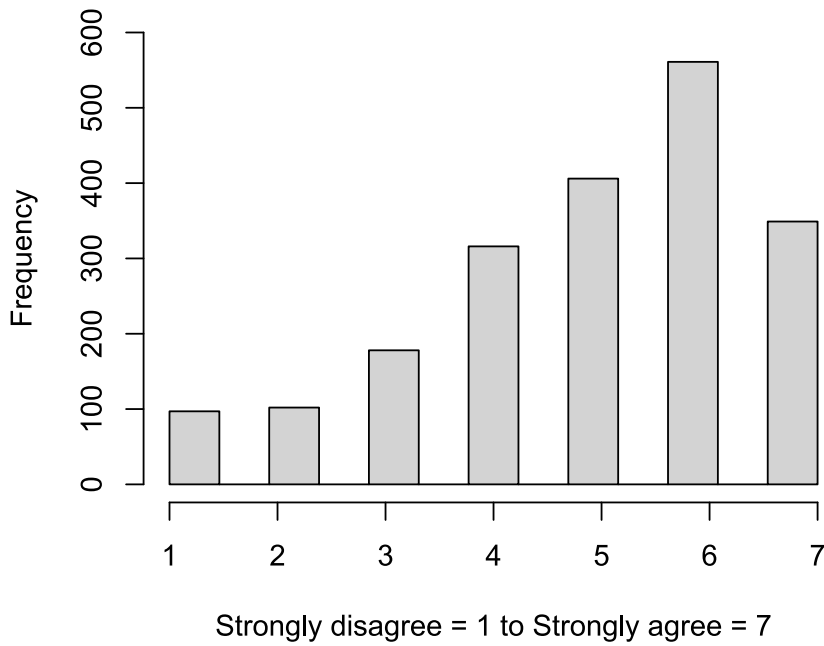


Figure 14 shows the distribution of participant responses to the following question: “Platforms take action on reported content. To what extent do you agree/disagree with the following statements?”. Again, the median response was 5 (slightly agree), indicating that the majority of respondents agreed, to some extent, that VSP took action on the content that is reported.

Figure 14. A histogram of participant responses to “taking action” question (n = 2,156), excluding “don’t know” responses. Ratings range from 1 to 7 and the median is 5. The responses are skewed towards 7.

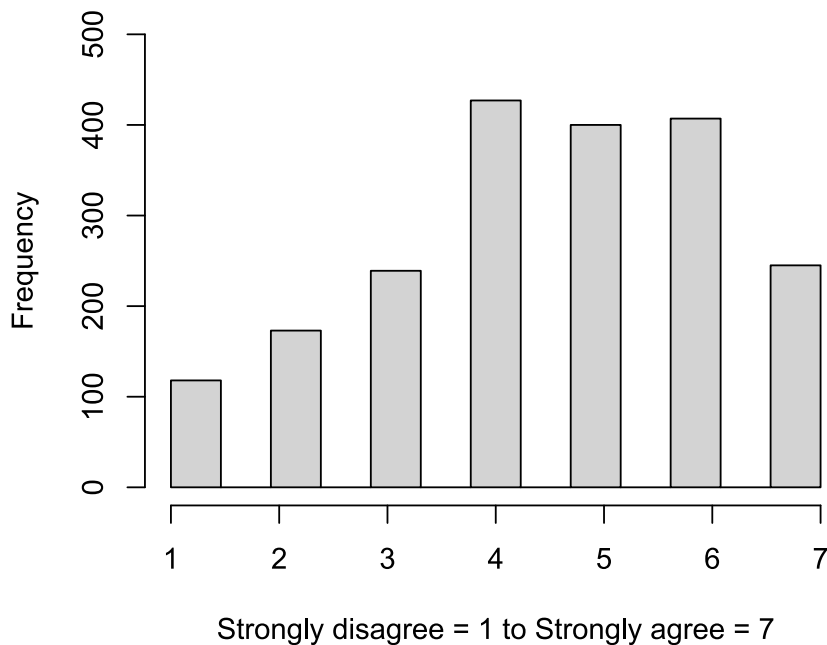


Figure 15 shows the distribution of participant responses to the following question: “User reports are an important part of how video sharing platforms identify harmful content. To what extent do you agree/disagree with the following statements?”.

The median response was 6 (agree), indicating that most of respondents agreed that user reports are an important part of how video sharing platforms identify harmful content.

Figure 15. A histogram of participant responses to the “importance” question (n = 2,265), excluding “don’t know” responses. Ratings range from 1 to 7 and the median is 6. The responses are skewed towards 7.

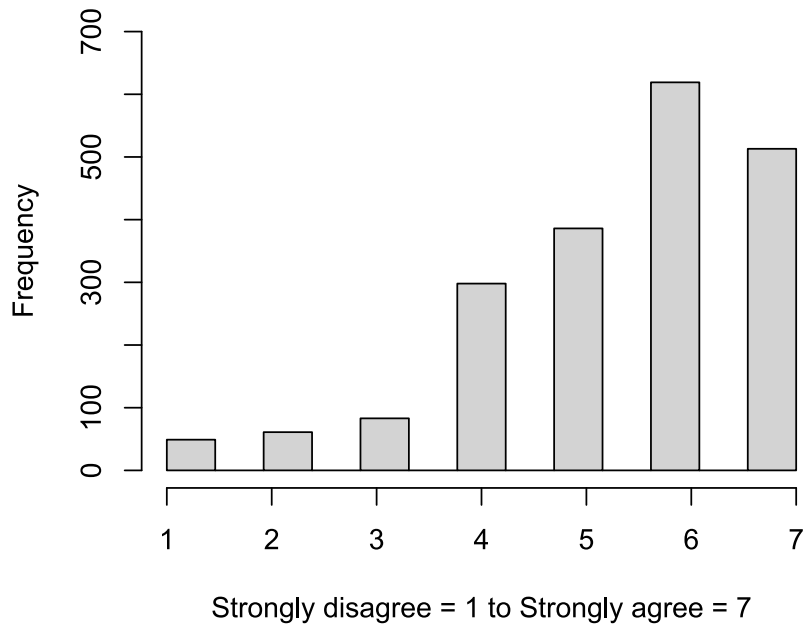
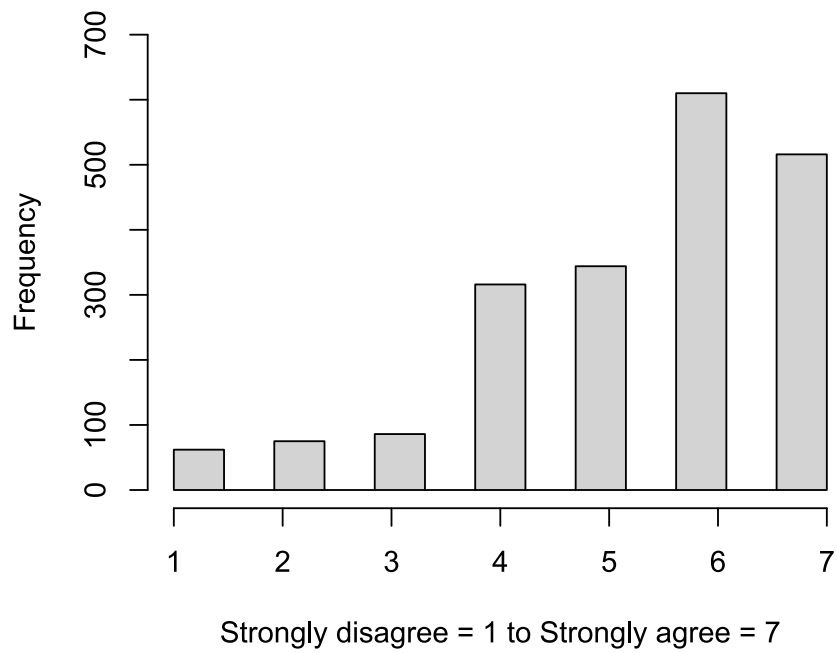


Figure 16 shows the distribution of participant responses to the following question: “If I see potentially harmful content in videos in the future, I will report it. To what extent do you agree/disagree with the following statements?”. The median response was 6 (agree), indicating that most of respondents intended to report potentially harmful videos in the future.

Figure 16. A histogram of participant responses to the “intent to report” question (n = 2,290), excluding “don’t know” responses. Ratings range from 1 to 7 and the median is 6. The responses are skewed towards 7.



8. Comments

These results provide evidence for the effectiveness of using salience in increasing the count of reports made of legal but potentially harmful content by VSP users.

In terms of the primary hypotheses, evidence was found to support hypothesis 1 and research hypothesis 2. Specifically, the reported research results indicate when the report button is visible at a top-level, participants were more likely to complete a report compared to the control (Hypothesis 1). In addition, prompting users to report when users disliked or commented on content (Arm 3) was found to be an effective means of increasing the count of reports compared to both the default VSP interface and the interface containing the reporting button only (Arm 2 – Salience). The evidence did not support the third hypothesis that participants would be more likely to complete a report when the process involves simpler choices (Arm 4) compared to Arm 3. However, both interventions employing double-salience (Arms 3 and 4) were found to be more effective at increasing the count of reports of potentially harmful content than the single-salience intervention.

There was not enough data to reliably examine the incidence of over-reporting and the differences in the number of started both not finished reports (the secondary outcomes). However, we found that both interventions employing double-salience (Arms 3 and 4) were found to be more effective at increasing the count of accurate reports of potentially harmful content than the single-salience intervention and the control. Interestingly, the accuracy of reporting was higher when simpler choices were offered compared to the other interventions used, in the follow-up video where reporting was encouraged. Encouragingly, most of the participants thought that reporting potentially harmful videos on VSP made a difference and that VSPs acted on the reported content. In addition, most of the participants also thought that reporting potentially harmful videos was important and that they will report potentially harmful videos in the future.

9. Appendix

Figure 17 shows the descriptive statistics by other engagements (likes, dislikes, comments, and shares) of the three potentially harmful content videos. Engagement across the three potentially harmful videos is relatively similar, with the biggest difference being in the percentage of responses disliking Homophobic (Music) video (27%) compared to Covid-19 vaccine misinformation video (23%).

Figure 17. Descriptive statistics – Other engagement (potentially harmful content)

When watching legal but potentially harmful videos, the most common form of engagement were dislikes.

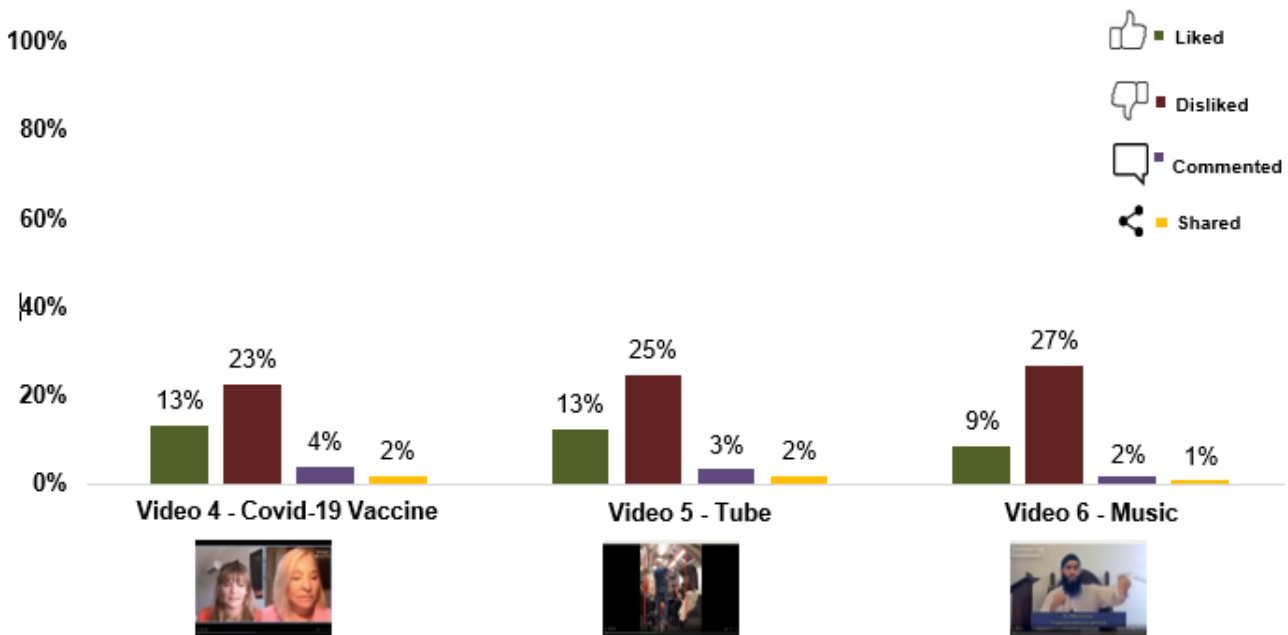


Figure 18 shows the descriptive statistics for other engagements (likes, dislikes, comments, and shares) of the three neutral content videos. Participants most liked watching Vegan Matcha Pancakes Recipe video and Blue Origin First Human Flight Booster Landing video. Participants reported liking Celebrity Breakups of 2017 the least out of these three videos.

Figure 19. Descriptive statistics – Other engagement (neutral content)

When watching neutral content, the most common form of engagement were likes.

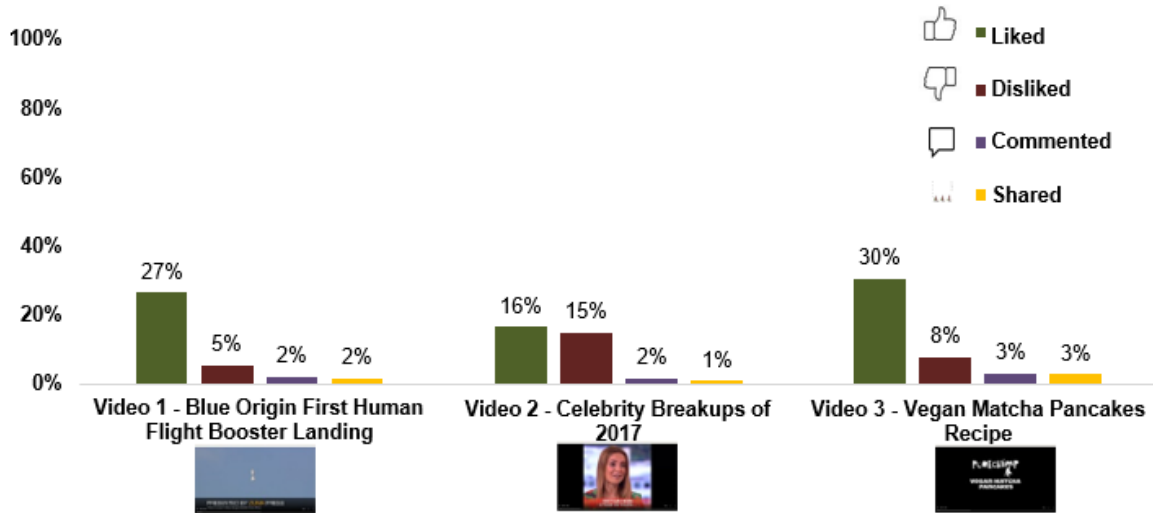


Figure 20 shows the median time spent watching each video. Participants spent most time watching potentially harmful video 5 (tube fight) compared to any other video. Participants spent least amount of time watching potentially harmful video 6 (Homophobic (Music)). The median length of time ranged from 13 seconds to 26 seconds per video.

Figure 20. Descriptive statistics - Median time spent watching videos

