KANTAR PUBLIC

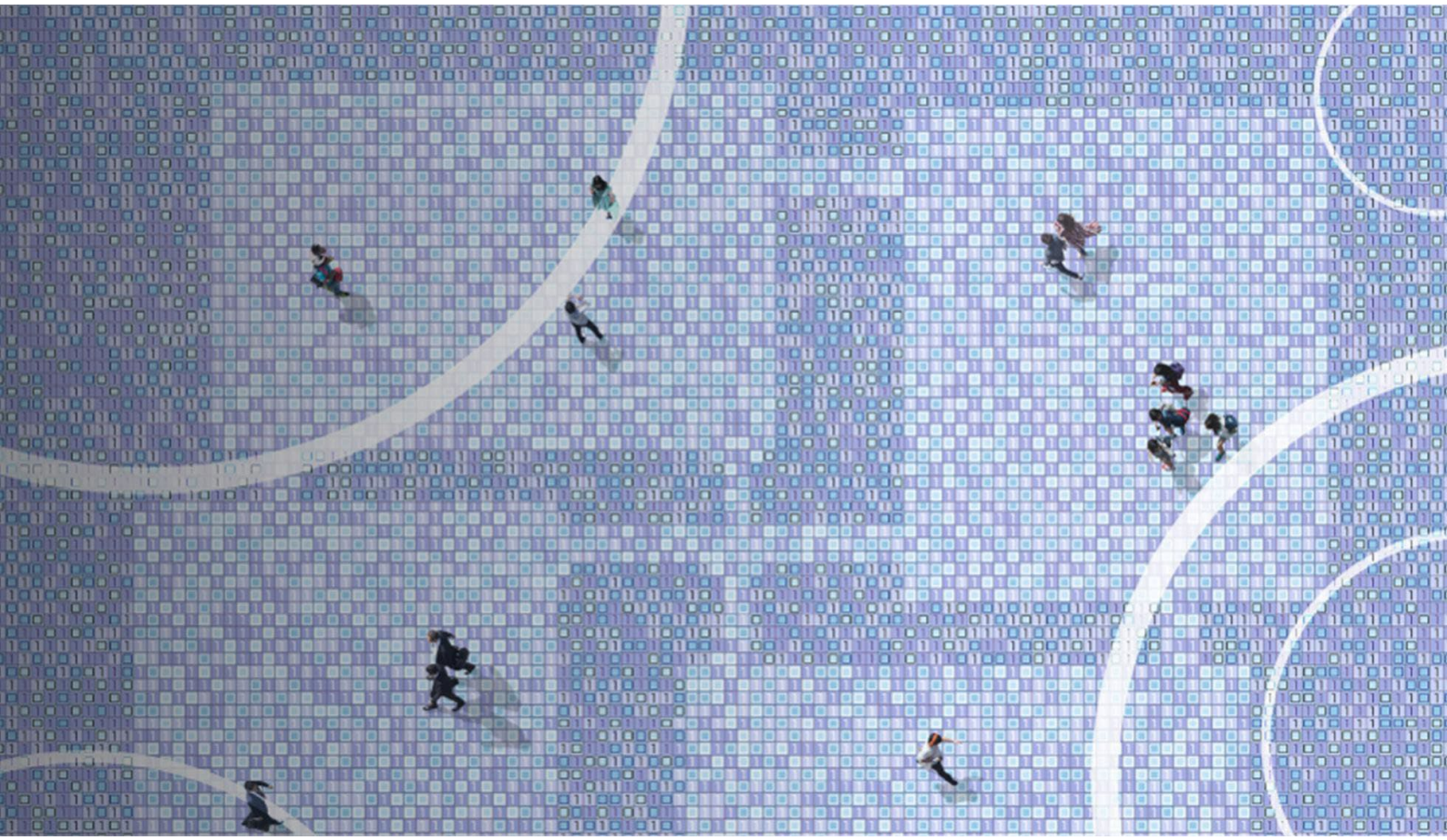
Ofcom online trials: alert messaging of video sharing platforms

Trial report

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

Revision: Michael Ratajczak, Rupert Riddle, and Natalie Gold.

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. For example, 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

In the case of this research, the focus was on testing the impact of interventions that could help users reduce their consumption of potentially harmful video content. To achieve this aim, interventions needed to encourage users to **skip the legal but potentially harmful content** presented in this study by using three types of alert messages (pop-ups that gave users the opportunity to skip each potentially harmful video). The three types of alert messages that were tested in this study included: a generic warning message about the potentially harmful content; a warning message that included high-level descriptive social proof; and a specific warning, describing what the potentially harmful content was.

1.3. Research questions

In this trial, we aimed to answer the following two research questions:

RQ.1. Do alert messages affect the consumption of potentially harmful video content?

RQ.2. Are there differences in the magnitude of the effect on the consumption of potentially harmful video content, between the different types of alert messages?

2. Sample and data collection

2.1. Sample

The target population, in this study, consisted of UK VSP users. This experiment aimed to provide a sample that was as representative as possible, with respect to key demographic characteristics, of UK VSP users. Consequently, demographic quotas were set based on the adjusted quotas used in previous trial of reporting mechanisms. Specifically, the demographic quotas were set with respect to the relative proportions of respondents in each demographic sub-group who passed the VSP use screener in the Reporting Mechanisms Trial (refer to the Reporting Mechanisms Trial report). Critically, participants who participated in the Reporting Mechanisms Trial were excluded from participating in this trial as they might have already developed some understanding of this trial which was very similar to Reporting Mechanisms Trial and therefore were not able to serve as a representative group of general UK VSP users. A total of 2,401 UK participants, **aged between 18 and 69**, were recruited from Kantar's Lifepoints panel. All participants were asked whether they had used a VSP in the past 12 months in response to a screener question provided at the beginning of the experiment. Those who had not used a VSP in the past 12 months were screened out from this experiment.

Kantar Public conducted this experiment online, using a device-agnostic platform; as such, participants were able to complete the experiment on a computer, mobile, or tablet, subject to participants' preference. Fieldwork took place in March 2022 over a two-week period.

Table 1 shows the quotas set before the recruitment began, and the quotas that were met when recruitment ended.

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

Demographics		Start	Finish
Gender	Male	49%	49%
	Female	51%	51%
Age	18-24	14%	14%
	25-39	34%	34%
	40-54	30%	30%
	55-69	22%	22%
Ethnicity	White	87%	87%
	Mixed/Multiple Ethnic Groups	2%	2%
	Asian/Asian British	7%	6%
	Black/African/Caribbean/Black British	3%	3%
	Other Ethnic Group	1%	<1%
	Prefer not to say[2]	-	1%
	ABC1	56%	56%

[1] When using parallel quotas, the sample will aim to fulfil all required quotas on age, gender, SEG, location and ethnicity. However, those proportions would not be interlocked with each other. This would mean a final sample with the correct proportion of each category, i.e., 49% male, 51% female, 56% SEG ABC1, 44% SEG C2DE etc. However, it is possible that all the male participants might end up in SEG ABC1. This study used parallel quotas instead of interlocked quotas because it would have been very time-consuming and expensive to recruit a balanced sample that uses interlocked quotas.

[2] Includes participants who did not agree to be asked this question (n = 22) and those who refused to answer this question when asked (n = 8).

Socio-economic	C2DE	44%	44%
Location	London	14%	11%
	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%	15%
	South West	8%	9%
	Wales	5%	5%
	Scotland	8%	8%
	Northern Ireland	3%	2 %

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, three of which included interface-based interventions that aimed to reduce the consumption of potentially harmful video content.

To allocate respondents to experimental arms, a method of blocked randomisation was used (least- filled quotas). This method ensured that blocks were filled 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] and was successfully used to recruit participants in the Reporting Mechanisms Trial.

2.4. Incentivisation

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

2.5. Ethics

[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

The purpose of the experimental environment was to replicate the real-world context as closely as possible, to get as close as possible to actual VSP users' behaviour. It would have been difficult, if not impossible, to gain externally valid evidence of the propensity to skip legal but potentially harmful content in an experiment that did not expose participants to actual legal but potentially harmful content. However, exposing participants to much of the content deemed 'harmful' would not have been ethically acceptable, and to attempt to do so without mitigation presented high risk to participants, as well as to Ofcom.

Kantar Public's Behavioural Practice team reused the videos from the previous experiment with Ofcom on the reporting of potentially harmful content (see the Reporting Mechanisms Trial report). In the Reporting Mechanisms Trial, legal but potentially 'harmful' content (content that some participants could consider to be harmful) was selected for inclusion 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 and potential financial fraud.
3. Sharing these videos with the Kantar project team and Kantar Public's Profiles' Privacy team 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.

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 indicated that 70% of VSP users were exposed to potential online harm on the services they used during the past three months).[6]

The following steps were taken to mitigate any residual risk 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. Participants were allowed to refuse to participate if they did not want to be exposed to this, and 189 respondents chose to refuse to participate (Figure 1 shows the participant flow).

In addition to the above, in the study, participants were able to skip any of the video content at any point. This means that they were not required to watch any of the videos if they did not want to.

A debrief screen at the end of the experiment provided web links to support on any of the potential harms included in the content shown in the experiment.

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 tests

The Profiles panel conducted a range of quality and validation checks when recruiting their panellists.[7] In addition, and to keep the quality of data high and remove any skimmers who were attempting to get through the experiment as quickly as possible, an attention check was included in this experiment.

First, any respondent, who completed the study in less than 40% of the median completion time for all respondents, was removed. Second, any respondent who failed to correctly answer the attention check question was excluded from the study. The attention check specified: "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:

[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

[7] More information available at https://www.kantar.com/expertise/research-services/panels-and-audiences/lifepoints-research-panel

The total drop-out rate, due to responders completing the study too quickly or failing the attention check, over the whole study, was 11%. The drop-out rate due to responders failing the attention check was 3%. This was similar to the drop-out rate in the first experiment (Reporting Mechanisms Trial), where the total drop-out rate, due to responders completing the study too quickly or failing the attention check, was 8% and the drop-out rate due to failing the same attention check, only, was 5%.

2.8. Soft launch

To ensure that there were no unforeseen issues with the experimental design and script, an initial soft launch involving 10% of participants was conducted in March. 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. During the soft launch, one participant was excluded from the study due to experiencing a glitch during the experiment. This glitch did not occur for any other participant

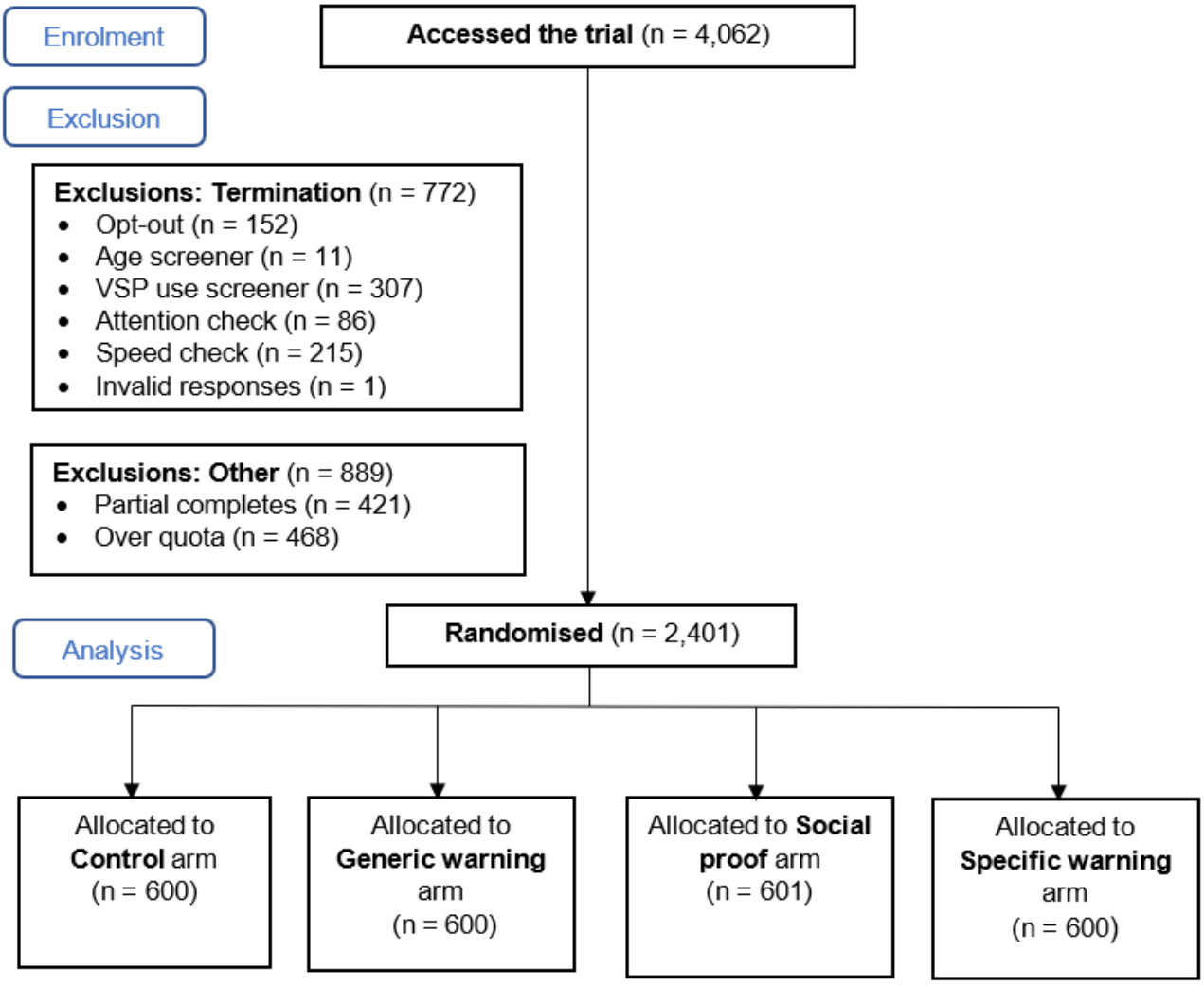


Figure 1. Participant flow diagram[8]

[8] Over quota refers to participants who were sent an invite to participate in the experiment, but whose quotas were full by the time they accessed the experiment.

3. Trial design and flow

Figure 2 describes the trial design and flow of the experiment. First participants are shown the introduction screen and asked about their demographics, then they are randomised into one of the four intervention arms. After randomisation, they are shown the training screen, and after the training screen they are exposed to the VSP interface containing 3 potentially harmful and 3 neutral videos. After having the opportunity to watch all the videos and interact with the platform, they are asked to complete a post-trial survey. Last, they are shown the debrief screen and can exit the experiment.

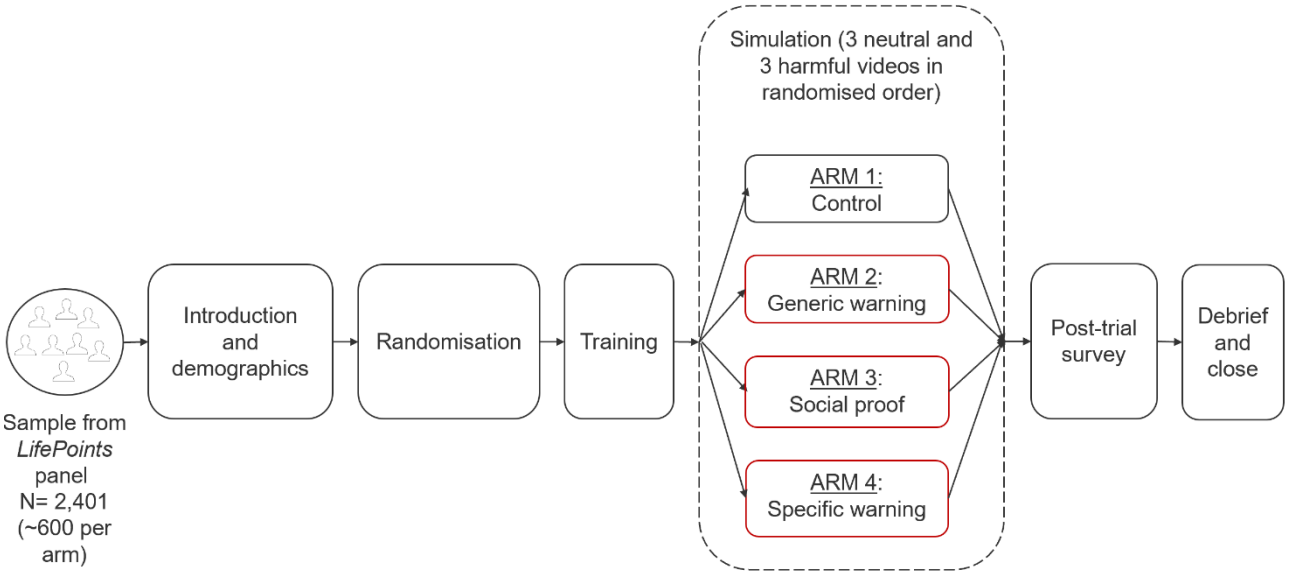


Figure 2. Trial design and flow

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 participation in the study involved. 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 potential harms included in the content shown in the experiment (see Figure 3).

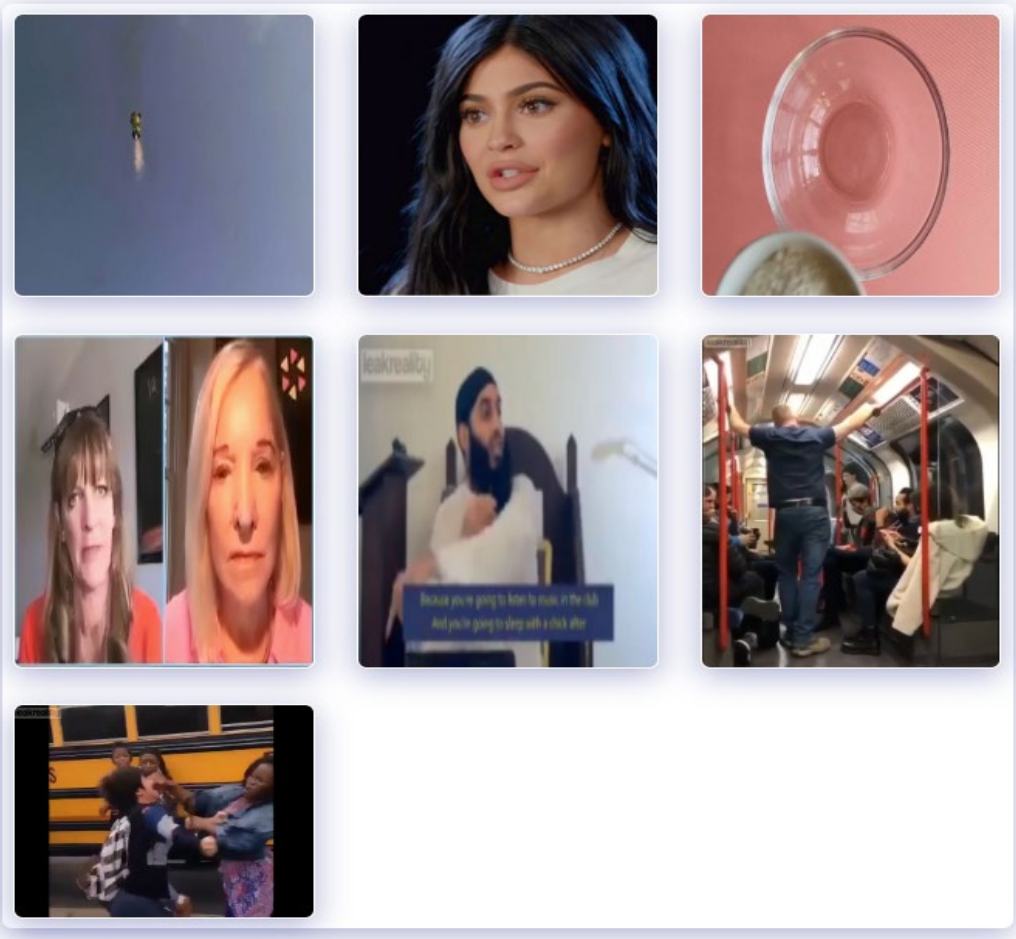
The screen shows what participants would see upon the completion of the study. Specifically, a "thank you" note and guidance and support for where to look for further information regarding the reporting of harmful content. Screenshots of all the videos used in the trial are also shown.

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).



Continue »

Figure 3. Debrief screen

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 (age, gender, socioeconomic background, location, and ethnicity).

Following this, 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 used within the past 12 months. 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 confirmed their demographics, the interface that they would be using in the experiment was introduced. At this stage participants were randomly allocated to experimental blocks, and they had the opportunity to interact with the interface that they were allocated to. First, participants saw a static

screenshot of the interface they were randomly allocated to with instructions for how they could use the buttons available and a short description of how the experiment would proceed.

The interface was a variation of a 'generic' VSP that had previously been found to increase the incidence of reporting of potentially harmful content (salience + prompt intervention, Arm 3, in Reporting Mechanisms Trial), incorporating and varying features that were common to many platforms but without any familiar branding. After users saw the labelled screenshot, they were shown a training video that they were able to interact with by choosing to react (like/dislike),[9] comment or share (indicated by adding in comments or pressing the share button in the interface), report, or skip past to the next piece of content (see the Figure 4 below).

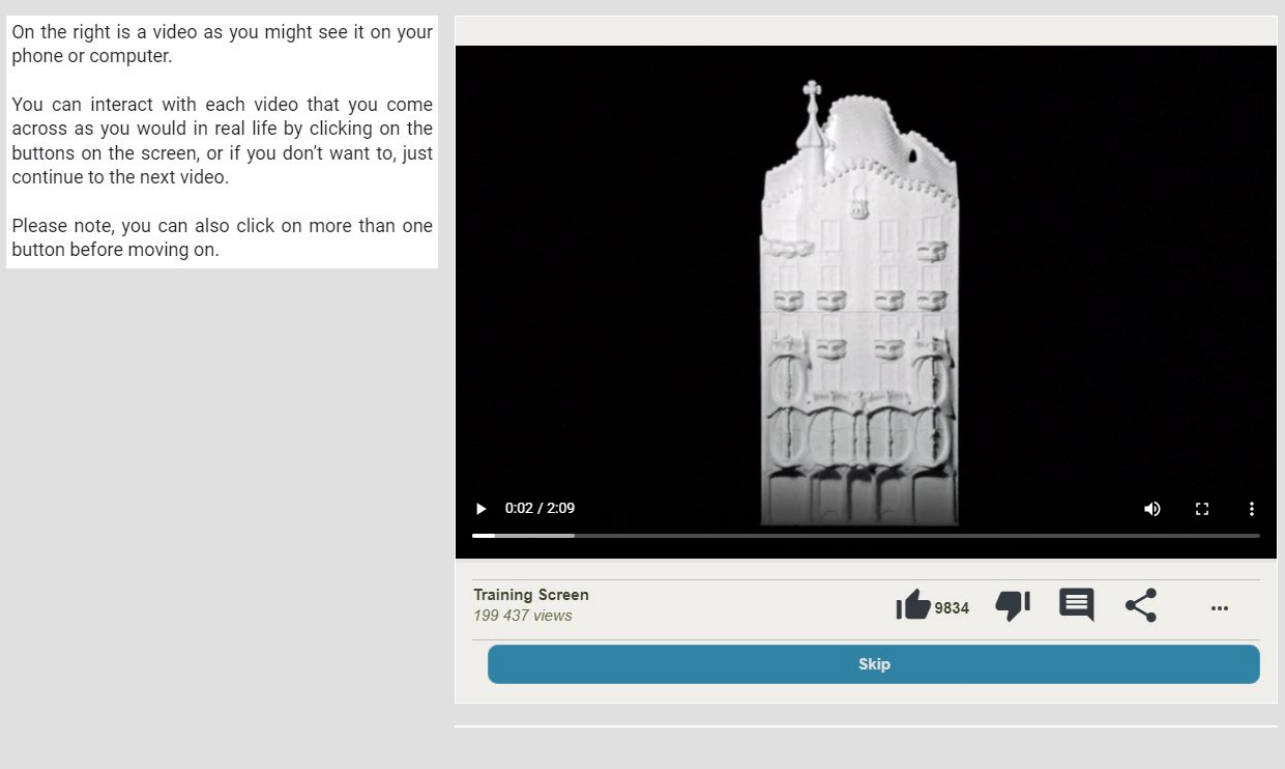


Figure 4. Control interface training screen. The training screen shows the interface that participants have been randomised to, in this case the control interface.

Participants were able to 'play' with this version until they were familiar with how it worked. The video content showed 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": 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

The experimental design aimed to reduce the amount of consumption of the potentially harmful content by encouraging skipping behaviour.

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

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 (of the type that was in the interest of the research outcomes to encourage people to reduce the consumption of). In treatment arms (Arm 2-4), an alert pop-up appeared before the potentially harmful content and covered the whole screen so that participants had to indicate whether they would like to view each potentially harmful video. The videos were either skipped if participants chose to skip each video or auto-played if participants chose to watch them. All three neutral videos were auto-played without any alert messages popped up.

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

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.

3.5. Post-trial survey

The participants were asked survey questions to understand their internet and social media usage as well as their attitudes toward the effectiveness of alert messages and whether they found the alerts annoying and regretted at watching the videos.

The same measures for all trial arms, other than questions asking about the utility of specific alert messages, were used, to ensure that results were comparable across all arms; this allowed us to use the responses to questions as **secondary outcomes** from the trial as well as diagnostics.

Secondary outcomes were used to: determine whether the effect estimates of the interventions were sensitive to how the skipping behaviour was measured; see how comparable the reporting behaviour of participants was to the reporting behaviour of participants in the Reporting Mechanisms Trial; examine whether there were differences in the viewing time of the potentially harmful content between arms; investigate how underlying attitudes were distributed in the sample, and the extent to which these were associated with outcomes from the experiment. The average time to complete the study was 9 minutes and 6 seconds.

4. Interventions

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

1. *Arm 1 – Control*: The control arm included an interface that simulated a VSP. The interface was the salience + prompt intervention used in Arm 3 of the Reporting Mechanisms Trial, without any alert messages informing users that they were about to see potentially harmful content. This interface remained the same for Arms 2, 3 and 4 but with different alert messages based on the treatment.

2. *Arm 2 – Generic warning:*[10,11] Participants saw "This video may contain sensitive material." presented in a pop-up before seeing each potentially harmful video. If users saw the alert message before interacting with the video, they might be more likely to skip the potentially harmful video, thereby reducing their consumption of potentially harmful content, given that this option was available.

   Hypothesis 1: Participants in Arm 2 would be more likely to skip potentially harmful videos than participants in Arm 1.

3. *Arm 3 – High-level descriptive social proof:*[12,13] Participants saw "This video contains material that other viewers on this platform have reported as being sensitive." presented in a pop-up before seeing each potentially harmful video. The use of social proofs typically changes people's behaviour to conform with these social proofs. Consequently, if users were informed that other people found potentially harmful videos offensive, they might be more likely to reduce their consumption of these videos compared to the control. Furthermore, research findings indicated that the use of high-level descriptive social proofs in pop-up messages might encourage users to reduce their consumption of potentially harmful videos to a greater extent than pop-up messages which did not use these proofs.

   Hypothesis 2.i: Participants in Arm 3 would be more likely to skip potentially harmful videos than participants in Arm 1.

   Hypothesis 2.ii: Participants in Arm 3 would be more likely to skip potentially harmful videos than participants in Arm 2.

4. *Arm 4 – Specific content warning:*[14] Participants saw a pop-up describing the content specific to each potentially harmful video before seeing each potentially harmful video.

   Covid-19 Misinformation: "This video contains misinformation."

   Tube Racism: "This video contains scenes of violence."

   Homophobic: "This video contains offensive language."

   It was hypothesised that the use of specific warnings might be more effective at reducing the consumption of potentially harmful video content than generic warnings.[15] However, it was uncertain if the use of specific warnings would be more or less effective than the use of high-level descriptive social proofs in reducing the consumption of the potentially harmful video content.

   Hypothesis 3i: Participants in Arm 4 would be more likely to skip potentially harmful videos than participants in Arm 1.

   Hypothesis 3ii: Participants in Arm 4 would be more likely to skip potentially harmful videos than participants in Arm 2.

   Hypothesis 3iii: Participants in Arm 4 would be more or less likely to skip potentially harmful videos than participants in Arm 3.

---

[10] Desimone, R., & Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annual review of neuroscience, 18*(1), 193-222.
[11] Steenbergh, T. A., Whelan, J. P., Meyers, A. W., May, R. K., & Floyd, K. (2004). Impact of warning and brief intervention messages on knowledge of gambling risk, irrational beliefs and behaviour. *International Gambling Studies, 4*(1), 3–16.
[12] Auer, M. M., & Griffiths, M. D. (2015). Testing normative and self-appraisal feedback in an online slot-machine pop-up in a real-world setting. *Frontiers in psychology, 6*, 339.
[13] Xiao, X., & Borah, P. (2020). Do Norms Matter? Examining Norm-Based Messages in HPV Vaccination Promotion. *Health Communication, 36*(12), 1476–1484.
[14] Ling, C., Gummadi, K. P., & Zannettou, S. (2022). " Learn the Facts About COVID-19": Analyzing the Use of Warning Labels on TikTok Videos. arXiv preprint
[15] Ibid.

.

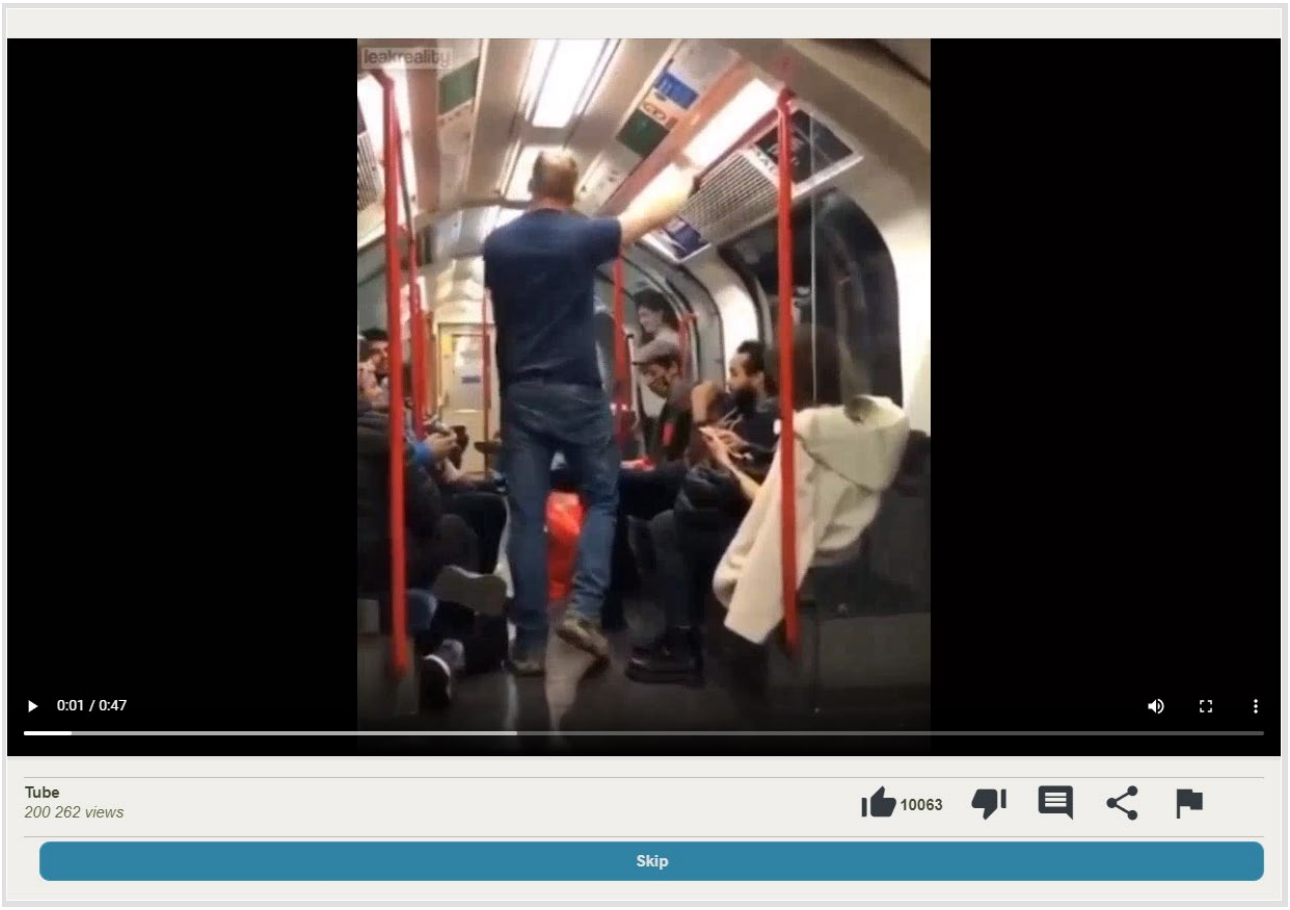


Figure 5. Arm 1 – Control (no alert message)

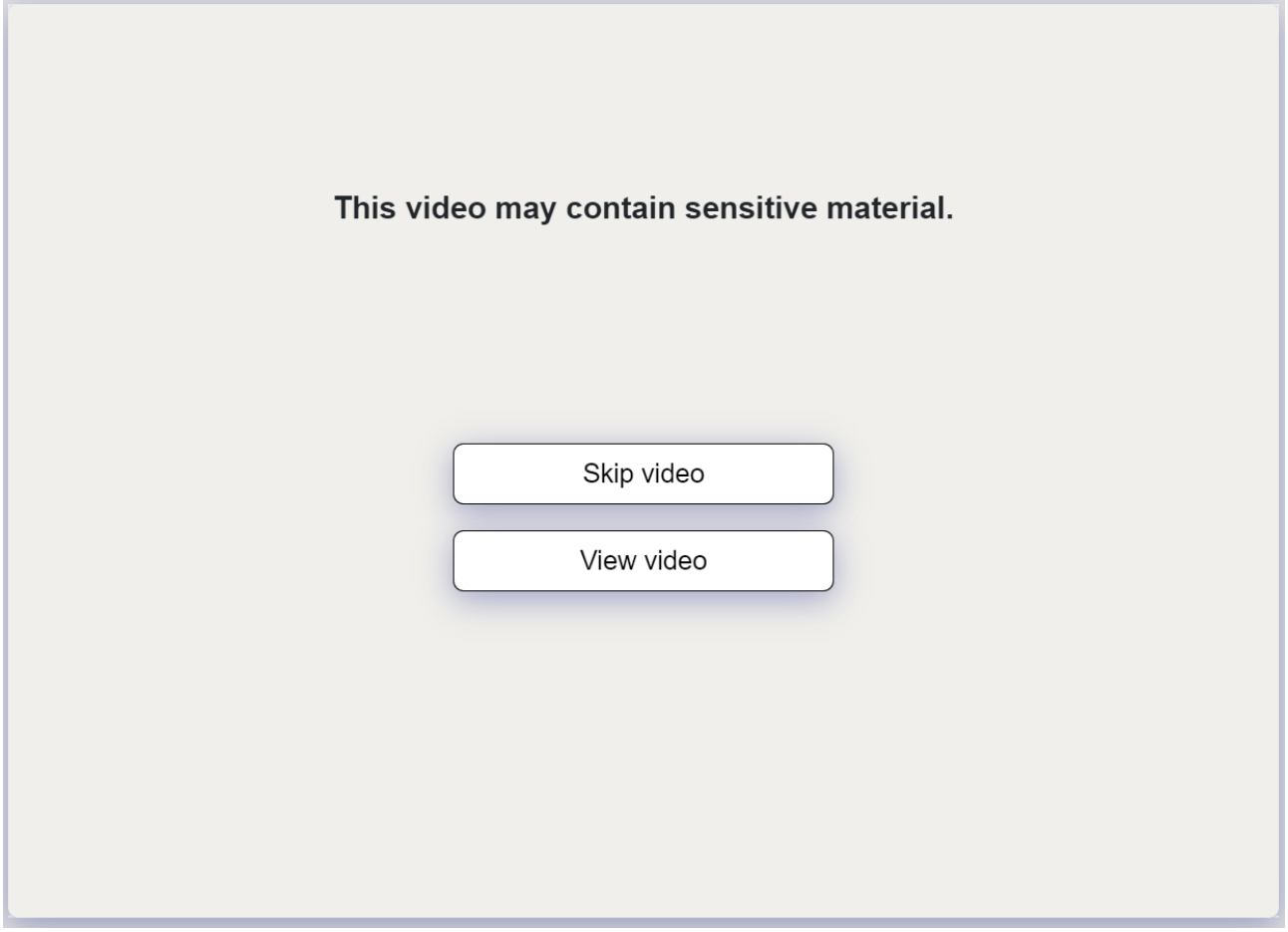


Figure 6. Arm 2 – Generic warning

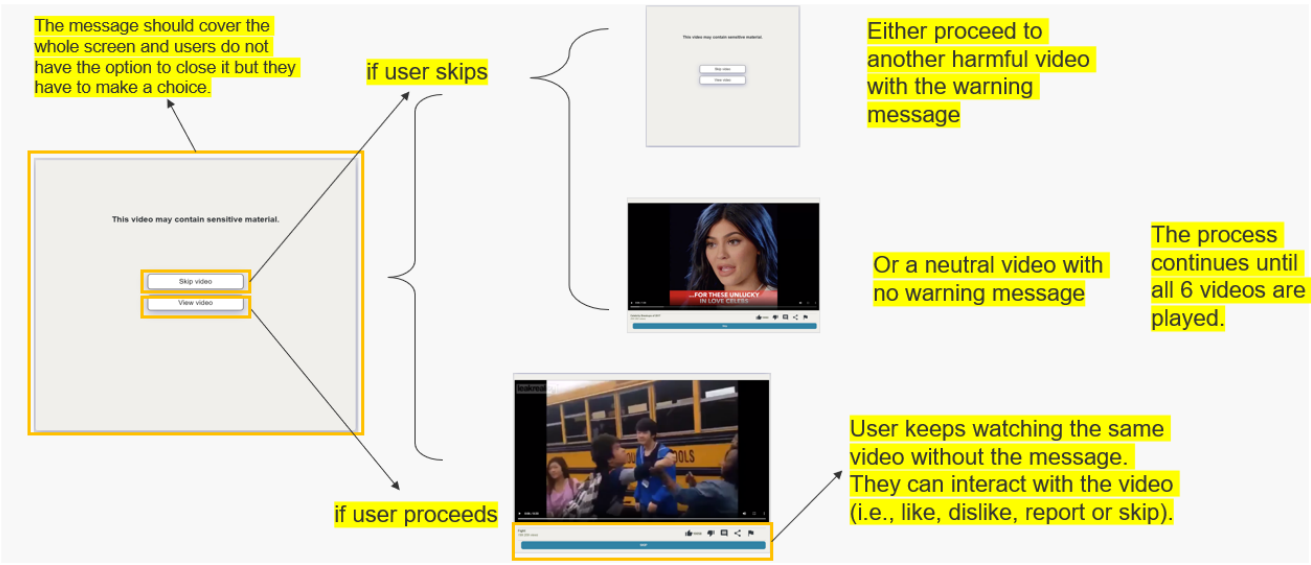


Figure 7. Arm 2 – Generic warning (skipping flow)

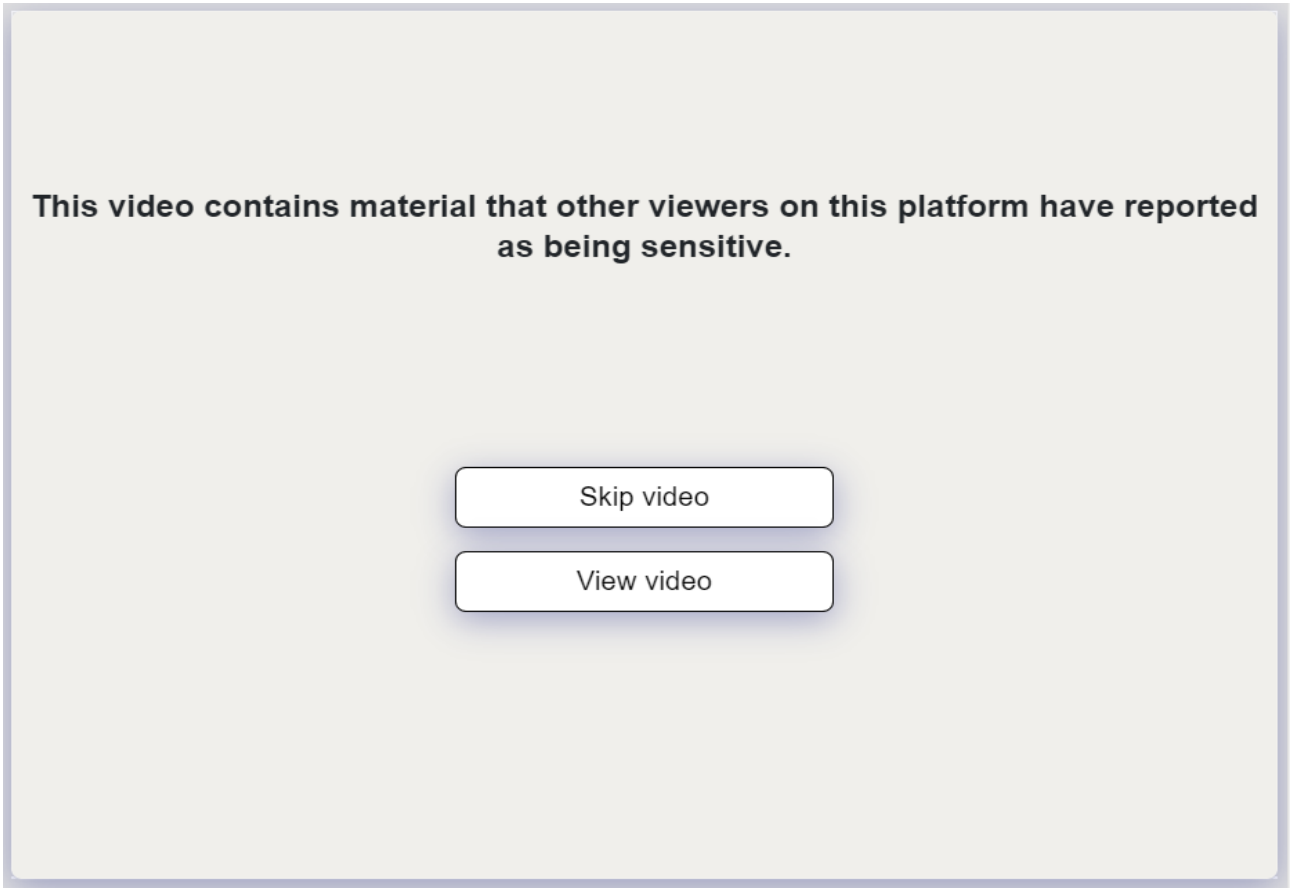


Figure 8. Arm 3 – High-level descriptive social proof

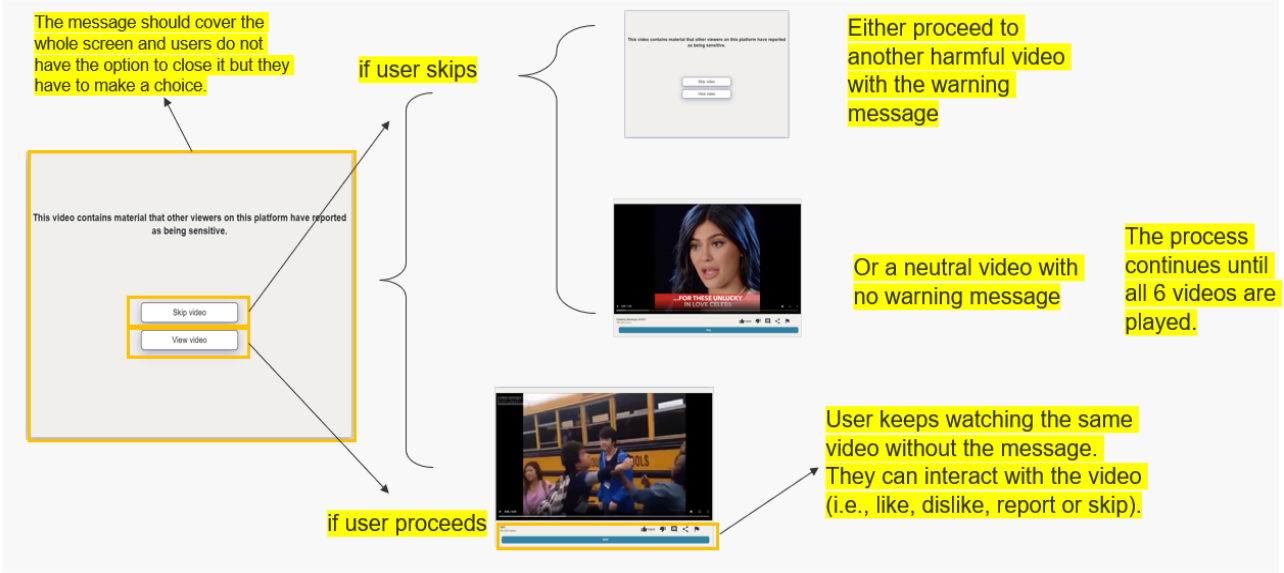


Figure 9. Arm 3 – High-level descriptive social proof (skipping flow)

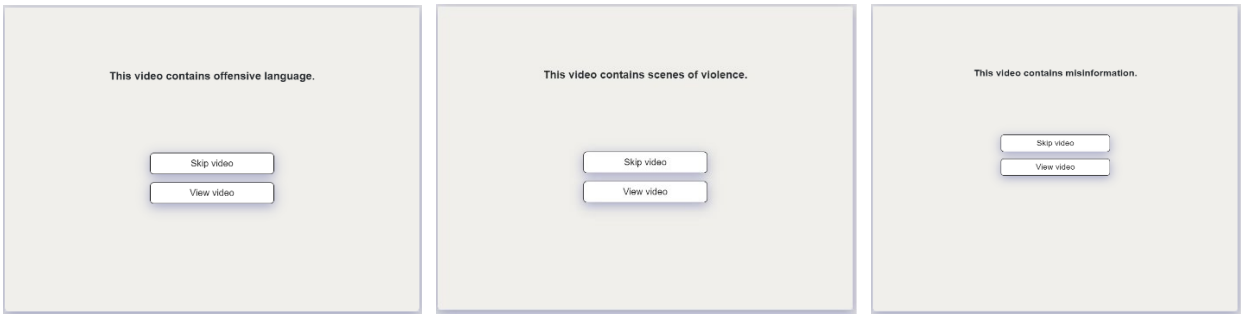


Figure 10. Arm 4 – Specific content warning

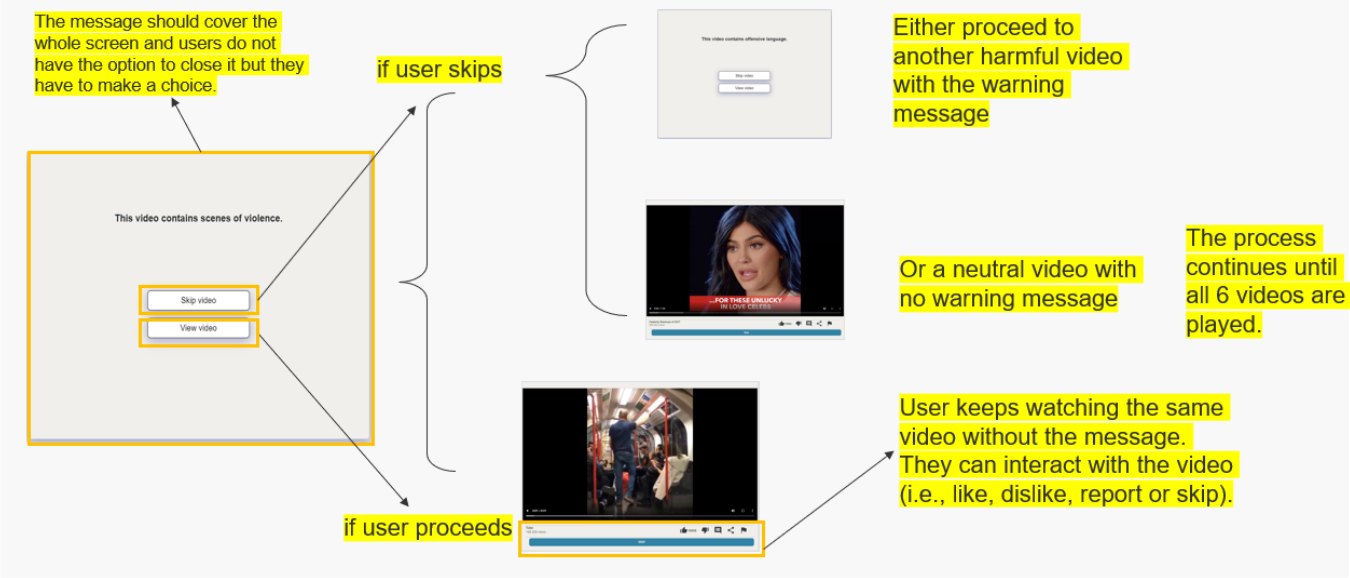


Figure 11. Arm 4 – Specific content warning (skipping flow)

5. Outcomes

5.1. Primary outcome

In this study, we measured whether a participant decided to skip (1) or not skip (0) each potentially harmful video (out of 3) at the pop-up alert stage or at the stage of watching the videos (if they decide to proceed). This binary variable constituted the primary outcome variable (1.1 in Table 2). In the control and intervention arms, videos that were not engaged with and not watched to the end by participants were considered to be skipped (1), and all other videos were classed as not skipped (0). This included videos that, for example, were 'liked' but not watched to the end. On discussion with Ofcom, we agreed that videos that were skipped at any point before finishing, regardless of whether participants engaged with them (e.g., by liking them), should be coded as skipped (1). Thus, we have derived a new primary outcome measure using the collected data to account for this.

Revised outcome measure: whether a participant decided to skip (1) or not skip (0) each potentially harmful video (out of 3) at the pop-up alert stage or at the stage of watching the videos (if they decide to proceed). In the control and intervention arms, if a participant opted to skip a video at any point prior to 1 second before each video ended, their response was coded as skipping (1), otherwise (if they watched the video at least until 1 second prior to each video ending) it was coded as not skipping (0).[16][17]

5.2. Secondary outcomes

The first secondary outcome (2.1 in Table 1) was originally whether a participant decided to skip (1) or not skip (0) each video (including neutral content) at either the pop-up alert stage, where relevant, or at the stage of watching the videos. In the current revision, this outcome was revised to whether a participant decided to skip (1) or not skip (0) each neutral video at either the pop-up alert stage, where relevant, or at the stage of watching the videos. The decision to revise this outcome was made because it is of interest to examine whether interventions targeting potentially harmful content affect skipping of neutral videos.

The second secondary outcome was a binary variable indicating whether a user submitted a complete report of potentially harmful content when viewing each potentially harmful video (2.2 in Table 2). This outcome was measured to see how comparable the reporting behaviour of participants was compared to the reporting behaviour of participants in the Reporting Mechanisms Trial. In addition, it allowed us to see whether there were any differences in the incidence of reporting between the different intervention arms in this study.

The third secondary outcome was the length of time participants viewed each potentially harmful video (2.3 in Table 2). Measuring this outcome allowed us to examine whether participants in the intervention arms spent a different amount of time, compared to participants in the control arm, on watching the potentially harmful videos.

Last, we captured responses to attitudinal questions at the end of the study (2.4 in Table 2). This allowed us to see whether there were any differences in self-reported attitudes to the alert messages between the intervention arms.

Table 2. The list of outcome measures and descriptive metrics used in the study.

	Behavioural	Attitudinal
Revised Primary Outcome	1.1. Revised to: Whether a participant decided to skip (1) or not skip (0) each potentially harmful video (out of 3) at the pop-up alert stage or at the stage of watching the videos (if they decide to proceed). In the control and intervention arms, if a participant opted to skip a video at any point prior to 1 second before each video ended, their response was coded as skipping (1), otherwise (if they watched the video at least until 1 second prior to each video ending) it was coded as not	

[16] We decided to use 1 second as a threshold to allow for differences in connection speed between participants.

[17] For example, if a video was exactly 48.49 seconds long, then anyone who pressed 'Continue' when the watch time was lower than 47.49 seconds would be classed as skipping.

	skipping (0)	
Secondary	2.1. Revised to: whether a participant decided to skip or not skip each neutral video at either the pop-up alert stage or at the stage of watching the potentially video 2.2. A binary variable indicating whether a user submitted a complete report of potentially harmful content when viewing each potentially harmful video 2.3. The length of time participants viewed each potentially harmful video	Belief that a given content was actually harmful/worth reporting 2.4. Responses to attitudinal questions regarding: -- Whether the warning messages were useful to participants. -- Whether participants found the warning messages annoying. -- Whether participants regretted watching each potentially harmful video they chose to see after being exposed to a pop-up alert
Descriptive metrics	Other engagement: Number of likes, dislikes, shares, and comments on potentially harmful content and neutral videos Length of viewing time for both harmful and neutral video posts A binary variable indicating whether a participant was shown 3 potentially harmful videos in a row or not	

6. Statistical methods and analysis

6.1. Statistical methods

Primary analysis

The primary outcome for the analysis was whether a participant decided to skip or not skip each potentially harmful video at the pop-up alert stage or at the stage of watching the potentially harmful video (if they decided to proceed) (see 1.1 in Table 2).

Given that this outcome was binary (skip vs. not skip) a logistic mixed-effects model was proposed to examine the differences between the different intervention arms. A logistic mixed-effects model is a logistic model containing both fixed and random effects. One of the key advantages of using this form of model is that it considers additional variability in the data (because it allows the addition of random effects to a logistic regression model). Random effects model additional uncertainty due to variation in individual responses (random intercept for participants) and in the potentially harmful video content (random intercept for potentially harmful videos).

The motivation for including random intercepts for each video and each participant was that the skipping behaviour was expected to be different for different videos (based on average times from the Reporting Mechanisms Trial, reasonable differences in the probability of skipping between videos were expected), and between different participants. In other words, it was not assumed that every potentially harmful video had equal probability of skipping, nor that this probability was the same for every individual. Instead, it was assumed that some of these videos had a lower or higher probability of skipping than others and that these probabilities varied by person. Thus, a basic proposed model specification was:

$$Y_{ij} \sim Bernoulli\left(Y_{ij}^{0}\right), Y_{ij} \in \{0,1\}, Y_{ij}^{0} = Prob(Y_{ij} = 1)$$

$$Logit\left(Y_{ij}^{0}\right) = \beta_0 + \beta_1 Arm_i^2 + \beta_2 Arm_i^3 + \beta_3 Arm_i^4 + u_{1i} + u_{2j}.$$

In the equation above, Y_{ij} is a binary variable indicating whether a participant i watching potentially harmful video j presses the skip button or not. The binary variable is 1 if the video is skipped, but 0 if the video is not skipped.

β_0 is the estimated coefficient for a baseline category - in other words a baseline chance for skipping a potentially harmful video in Arm 1 - whereas β_1, β_2, and β_3, represent deviations in the log-odds of skipping potentially harmful videos of Arms 2, 3, and 4, respectively, from Arm 1. Note that in the equation above, the logit link function changes how we interpret the model estimates (intercepts, effects, and variances). The logit function is the log of the odds of getting a 1 (skipping potentially harmful video, in this context). For example, if the β_1 value were 2.20 in log-odds, then this would mean that the log-odds of skipping potentially harmful video are 2.20 higher in Arm 2 compared to Arm 1. Log-odds can also be converted to probabilities, assuming β_1 = 2.20 log-odds, the probability of skipping a potentially harmful video is (1/(1+exp(-2.20))*100 = 90% higher in Arm 2 compared to Arm 1.

u_{1i} is the random intercept of participant i, $u_{1i} \sim N(0, \sigma_1)$ for $i \in \{1, 2, \ldots, N\}$ where N is the number of participants, and u_{2j} is the random intercept of potentially harmful video j, $u_{2j} \sim N(0, \sigma_2)$ for $j \in \{1, 2, 3\}$.

To answer our first research question (see Section 1.3), using this model, we wanted to test the following hypothesis:

$H_0^1: \beta_0 = \beta_k; H_1^1: \beta_0 \neq \beta_k \; where \; k \in \{1, 2, 3\}$.

If the inclusion of the random intercept for videos will lead to non-convergence, this random effect will be dropped. A simpler analysis, without random effects, will be considered if the inclusion of both random intercepts of participants and potentially harmful videos led to non-convergence. By default, both random effects, in the context of this study, are required to make full use of the data.

In addition, to answer our second research question (see Section 1.3), comparisons between the treatment and control arms will be performed. When running these multiple comparisons, the Bonferroni correction will be utilised to maintain the family-wise error rate.

Secondary analysis

Secondary outcome 2.1 (in Table 2) was intended to be analysed in the same way as the primary outcome (1.1 in Table 2), to determine whether the effect estimates of the interventions are sensitive to how the skipping behaviour is measured.

If we will not be able to analyse the Secondary outcome 2.2 (in Table 2) analysed in the same way as the primary outcome due to the inflation of zeros, Secondary outcome 2.2 will be analysed in the same as the primary outcome in the Reporting Mechanisms Trial: using a zero-inflated Poisson regression model.

Secondary outcome 2.3 (in Table 2) will be analysed using a similar model to the primary outcome (1.1 in Table 2). The difference in the modelling approach is that for Secondary outcome 2.3 the model will be a robust linear mixed-effects model (because the outcome is continuous),[18] rather than a logistic mixed-effects model that will be used for the binary outcome variable that constitutes the primary outcome in this study.

Secondary outcome 2.4 (in Table 2) will be analysed using Kruskal-Wallis test to compare the responses of the participants in the intervention arms to those in the control arm. The Kruskall-Wallis test is a non-parametric equivalent to a one-way analysis of variance. Since the test is non-parametric it does not assume normal distribution of the residuals. Consequently, it can be used to compare three or more groups on an ordinal outcome variable, such as Likert-scale-type responses to attitudinal questions (for example, Secondary outcome 2.4 in Table 2).

6.2. Statistical power

To run power simulations for logistic mixed-effects models, assumptions about the variance and standard deviation parameters of the random effects are required. In the context of the proposed study, the estimates to be specified concerned the variation in the probability of skipping between participants and variation in the probability of skipping between potentially harmful videos.

To obtain meaningful estimates of power using power simulations, these assumptions would typically be grounded in prior research. However, it was not possible to identify any prior research involving logistic mixed-effects models within this research context. As a result, an assumption was made that the effect size of the interventions aimed at reducing the consumption of the potentially harmful content would be similar to that reported in Reporting Mechanisms Trial study on the reporting of potentially harmful content. It was also assumed that this effect would be detected using mixed-effects models. As a result, the study aimed for n=600 participants per arm giving a total sample size of 2,400. The expectation was that the trial would be sufficiently powered to detect effects of similar size to those observed in Reporting Mechanisms Trial study.

Nonetheless, in parallel power simulations were carried out using the assumption that the sample size was fixed using the sample size from the previous experiment. Parameter estimates for the logistic mixed-effects models in the power simulations were then developed based on “informed” estimates. In this case, “informed” estimates were used because no previous research that could have been used as a source for parameter estimates has been found.

Effect sizes of 3%, 5% and 8% in the probability of skipping - per intervention arm compared to the control arm - were used under different scenarios. The scenarios contained different estimates of the parameters of variation in the probability of skipping between participants (random intercept for participants) and in the probability of skipping between potentially harmful videos (random intercept for potentially harmful videos).

Table 3 shows the estimates of power under different model assumptions, given 100 simulations per scenario. It should be noted that it is unlikely that the estimates under any scenario would be the same as the ones obtained using models given the collected data. Thus, the estimates of power to detect the effect of the intervention, given the scenarios considered, are unlikely to be an accurate representation of the true effect of the interventions. However, in the case of a null effect, the estimates below would provide evidence as to where the null effect might have come from (for example, small effect of the intervention or large variation due to individual differences). In Table 3, σ_1 is the random effect associated with variation between users, σ_2 is the random effect associated with variation between videos, and σ_3 is a noise estimate. Power refers to the percentage of the time that a significant difference was detected given a specified scenario. Note that σ_3 was added in power simulations to be conservative (which is useful in a context where we might

[18] Originally, we reported the results of a linear mixed-effects model. However, when performing the re-analysis, we revised the model choice of this outcome measure to a robust linear mixed-effects model. The inference was not sensitive to the modelling approach change.

that expect the interface would not work equally well for every participant), but it is not something that can be modelled by a logistic model given collected data.

Table 3. Power to detect an effect of specified size, by scenario

Scenario	Sample Size (3 videos each)	Effect	σ_1	σ_2	σ_3	Power
1	2400	3%	1.1	0.424	0.1	23%
2	2400	5%	1.1	0.424	0.1	62%
3	2400	8%	1.1	0.424	0.1	89%
4	2400	3%	0.55	0.424	0.1	33%
5	2400	5%	0.55	0.424	0.1	75%
6	2400	8%	0.55	0.424	0.1	97%
7	2400	3%	1.1	0.203	0.1	22%
8	2400	5%	1.1	0.203	0.1	56%
9	2400	8%	1.1	0.203	0.1	87%
10	2400	3%	0.55	0.203	0.1	34%
11	2400	5%	0.55	0.203	0.1	80%
12	2400	8%	0.55	0.203	0.1	100%

7. Results

In this section, we have abbreviated the names of the intervention arms in tables. Arm 2 – Generic warning is referred to as Arm 2 – GW; Arm 3 High-level descriptive social proof is referred to as Arm 3 – HLDSP; and Arm 4 – Specific content warning is referred to as Arm 4 SCW.

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 (Table 4).[19]

Table 4. Split of participants by age, gender, and socio-economic group (SEG), variables across trial arms

	Age (Median)	Gender (% Male)	SEG (% ABC1)
Arm 1 Control	40	49.2	58.5
Arm 2 GW	40	47.3	54.2
Arm 3 HLDSP	41	50.1	55.2
Arm 4 SCW	41	48.8	56.0

Note: ABC1 refers to upper middle class (A), middle class (B), and lower middle class (C1)

There were some differences in the types of devices that participants completed the experiment on, between arms. However, the distribution of iOS and macOS devices was relatively balanced across each arm (Table 5).[20]

Table 5. Split of participants by device operating system, by arm

Device operating system	Arm 1 Control (%)	Arm 2 GW (%)	Arm 3 HLDSP (%)	Arm 4 SCW (%)
Other	35.7	32.7	36.3	39.0
iOS (iPhone/iPad)	22.0	24.5	24.1	24.7
Windows	33.0	33.7	30.0	26.8
macOS	8.8	8.3	8.7	9.2
Linux	0.3	0.5	0.5	0.2
Unknown	0.2	0.3	0.5	0.2

7.2. Primary Outcome Analysis

7.2.1. Headline results

Pop-up alerts employing high-level descriptive social proof were found to be effective at increasing the probability of skipping potentially harmful videos, compared to not having a pop-up. This finding was robust to sensitivity analyses that adjusted for devices on which the experiment was completed and whether the potentially harmful videos were played by participants.

[19] Note that perfect balance does not have to be achieved for mixed-effects models to work well. This is because groups with less data will automatically be shrunk towards overall mean values. Consequently, mixed-effects models are well suited to unbalanced designs.
[20] Emphasis has been placed on iOS and macOS devices because the simulation of the VSP was most likely not to work as intended on these devices (see Section 7.2.3).

Pop-up alerts employing generic warnings or specific content warnings were also effective, but only under ideal conditions that were not representative of all participants' experience of interacting with the simulated VSP (due to differences in auto-play functionality between devices and operating systems). Thus, we think that they are less likely to work in the real-world context than pop-ups employing high-level descriptive social proof.

7.2.2. Skipping of potentially harmful videos

The mean observed probability of skipping potentially harmful videos was 69% in the control arm, 72% in Arm 2 – Generic warning, 75% in Arm 3 – High-level descriptive social proof, and 74% in Arm 4 – Specific content warning.[21] Table 6 shows that the odds of skipping the potentially harmful videos were significantly higher in Arm 3 – High-level descriptive social proof and Arm 4 – Specific content warning compared to the control arm. The whole model, including random effects, accounted for 44% of the variance in the probability of skipping potentially harmful videos (pseudo delta R^2 was 0.44).[22]

Table 6. Model-based estimates of the odds of skipping of potentially harmful videos

	Odds Ratios	95% CI	*z*-value	*P*
Intercept	3.90	1.33 – 11.43	2.479	0.013
Arm 2 GW	1.24	0.95 – 1.62	1.584	0.113
Arm 3 HLDSP	1.68	1.28 – 2.20	3.786	< 0.001
Arm 4 SCW	1.42	1.09 – 1.85	2.563	0.010

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

Adjusting for multiple comparisons, the odds of skipping potentially harmful videos were higher in Arm 3 – High-level descriptive social norm compared to the control arm (see Table 7). There were no other significant differences between arms.

There were slight differences in the proportion of device types used to complete the experiment between the control and intervention arms (see Table 5). Consequently, we re-ran the primary model with a device type covariate to control for any differences due to differences in proportions of device types used to complete the experiment, between control and intervention arms. A model with all device types did not converge.[23] For this model to converge, 48 observations from uncommon devices / operating systems were removed from the analysis.[24, 25] The reported effects were not sensitive to controlling for the device type.

In addition, the reported effects were not sensitive to the inclusion of a variable indicating whether participants watched three videos in a row or not; the probability of skipping was not significantly different for participants who were shown three potentially harmful videos in a row compared to those who did not.

Table 7. Estimates of the odds of skipping of potentially harmful videos (p values and CIs corrected for multiple comparisons using the Bonferroni correction)

Comparison	Odds Ratios	95% CI	*z*-value	*P*
Arm 1 Control vs	1.24	0.87 – 1.76	1.584	0.680

[21] Mean observed probabilities were calculated using observed values (from the collected data). The observed probabilities were reported as they are typically easier to understand than estimated (or model-based) probabilities. Note that the observed probabilities do not directly relate to the estimated odds ratios that are in Tables 6 and 7.
[22] Note that random effects models consider participant and video variability, thus the reported estimates are not driven by one particular video or by one particular participant.
[23] The implications of a model not converging are that the estimates of such a model cannot be trusted. This is because such a model's estimates are likely to be inaccurate (not precise), unreliable (not consistent), and/or biased (distorted).
[24] By rarely used devices, we mean 'Unknown' (constituting 21 observations, therefore 7 participants) and 'Linux' (constituting 27 observations, therefore 9 participants) devices. The observations relating to these devices were removed, because the inclusion of them led to lack of model convergence. This is because effect estimates for some arms could not have been reliably estimated. For example, only 1 participant in Arm 4 completed the experiment using a 'Linux' device or an 'Unknown' device.
[25] Note that, prior to analysing the data, we also grouped together participants running different versions of the Windows operating system into a "Windows" category. We did this because we had no reason to believe that the experiment would run differently for participants running different versions of a Windows operating system on desktop computers or laptops.

Arm 2 GW				
Arm 1 Control vs Arm 3 HLDSP	1.68	1.18 – 2.39	3.786	< 0.001
Arm 1 Control vs Arm 4 SCW	1.42	0.999 – 2.01	2.563	0.062
Arm 2 GW vs Arm 3 HLDSP	1.36	0.95 – 1.93	2.209	0.163
Arm 2 GW vs Arm 4 SCW	1.14	0.80 – 1.63	0.982	1*
Arm 3 HLDSP vs Arm 4 SCW	0.84	0.59 – 1.20	-1.228	1*

* Approximately (rounding error)

Figure 12 shows the percentage of participants skipping potentially harmful videos, by arm. The significant difference, as estimated using the primary outcome model, is shown using a horizontal line.

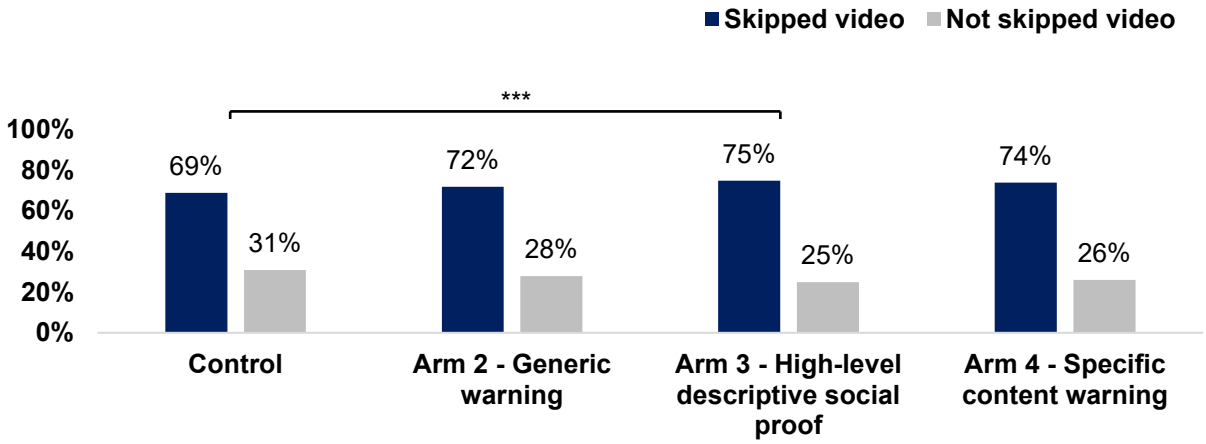


Figure 12. Percentage of skipped and not skipped potentially harmful videos, by arm (multiple comparisons adjusted; *** p < 0.001; not sensitive to controlling for device type; modelled using a mixed-effects logistic model)

7.2.3. Sensitivity check

Whilst user testing a separate online trial, it became apparent that for a minority of participants in certain circumstances, videos would not begin playing automatically.[26] This issue did not affect the alert message stage of the trial but may have impacted user experience of the video interface. Thus, we conducted sensitivity analyses to determine whether the reported effects were affected by this issue. Specifically, we excluded participant observations in which a potentially harmful video was skipped via the interface (after the alert message) with a watch time of 0 seconds.[27] In total, this led to the exclusion of 200 observations. We re-ran the primary model without these 200 observations to examine intervention effects under conditions where all participants see videos.

In addition, we also re-ran the primary model with a device type covariate to control for any differences due to differences in proportions of device types / operating systems used to complete the experiment between control and intervention arms. For this model to converge, 47 observations from uncommon devices / operating systems were removed from the analysis (see Footnotes 22 and 23).[28]

[26] The issue mainly affected participants who accessed the trial on a device running iOS (the operating system for iPhones and iPads, n = 117 observations). However, other Apple devices were also affected. We believe that this is a design choice by Apple, and that this cannot be overridden in this online environment.

[27] Device type or operating system was not included as criteria for the exclusion of observations, as this was also an issue for a minority of participants using other devices / operating systems, such as Windows (n = 32 observations), macOS (n = 24 observations), unknown device type (n = 26 observations), and Linux (n = 1 observation).

[28] The number of observations removed from this dataset was 47, as opposed to the 48 observations that were removed from the dataset used in the analysis reported in Section 7.2.2. This is because the dataset used for sensitivity analysis did not have the 200

The observed probability of skipping potentially harmful videos in this adjusted primary model was 67% in the control arm, 72% in Arm 2 – Generic warning, 75% in Arm 3 – High-level descriptive social proof, and 73% in Arm 4 –Specific content warning. Table 8 shows that the odds of skipping the potentially harmful videos were significantly higher in Arm 2 – Generic warning compared to the control arm, in Arm 3 – High-level descriptive social proof compared to the control arm, and in ARM 4 – Specific content warning compared to the control arm.

Table 8. Sensitivity check model-based estimates of the odds of skipping of potentially harmful videos

	Odds Ratios	95% CI	*z*-value	*P*
Intercept	3.29	1.08 – 9.99	2.099	0.036
Arm 2 GW	1.47	1.12 – 1.94	2.764	0.006
Arm 3 HLDSP	1.98	1.50 – 2.61	4.828	< 0.001
Arm 4 SCW	1.69	1.28 – 2.22	3.716	< 0.001

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

Adjusting for multiple comparisons, we replicated the significant differences reported in Table 8 (see Table 9). Note that the reported findings were not sensitive to controlling for the device type or whether participants watched three videos in a row or not. Thus, Arm 2 – Generic warning and Arm 4 – Specific content warning pop-ups increased the probability of skipping potentially harmful videos, but only amongst the sample of participants who, at least partially, watched potentially harmful videos. In other words, these pop-ups may work under ideal study conditions, but are less likely to work on average than pop-ups employing high-level descriptive social proof (Section 7.2.2.).

Table 9. Estimates of the odds of skipping of potentially harmful videos (p values and CIs corrected for multiple comparisons using the Bonferroni correction)

Comparison	Odds Ratios	95% CI	*z*-value	*P*
Arm 1 Control vs Arm 2 GW	1.47	1.03 – 2.11	2.764	0.034
Arm 1 Control vs Arm 3 HLDSP	1.98	1.38 – 2.85	4.828	< 0.001
Arm 1 Control vs Arm 4 SCW	1.69	1.17 – 2.42	3.716	0.001
Arm 2 GW vs Arm 3 HLDSP	1.34	0.94 – 1.93	2.109	0.210
Arm 2 GW vs Arm 4 SCW	1.14	0.80 – 1.64	0.968	1*
Arm 3 HLDSP vs Arm 4 SCW	0.85	0.59 – 1.22	-1.143	1*

* Approximately (rounding error)

Figure 13 shows the percentage of participants skipping potentially harmful videos, by arm. Significant differences, as estimated using a model without 200 observations, are shown using horizontal lines.

observations in which a potentially harmful video was skipped via the interface with a watch time of 0 seconds. Out of these 200 observations, 1 was a participant who completed the experiment using either ‘Unknown’ or ‘Linux’ device.

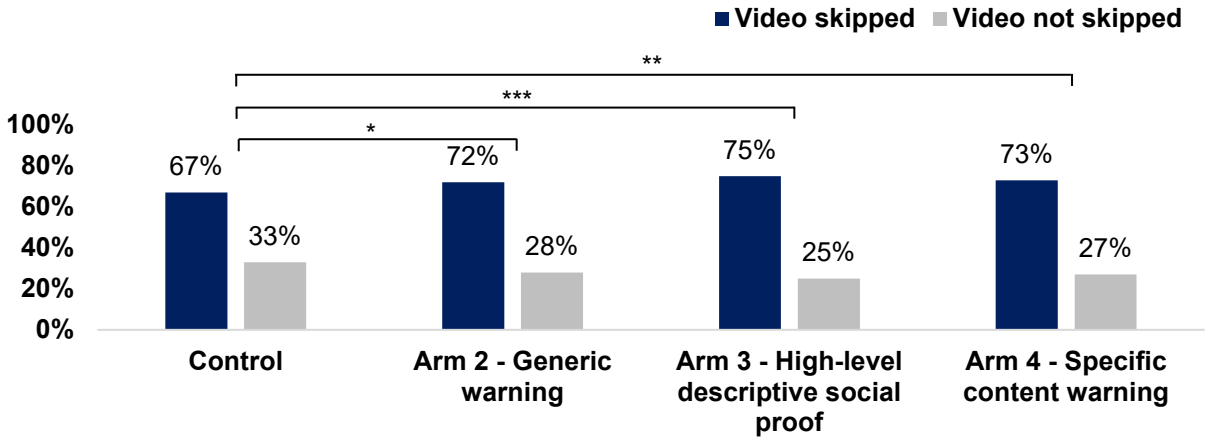


Figure 13. Percentage of skipped and not skipped potentially harmful videos, by arm – sensitivity with some observations removed (multiple comparisons adjusted; * p < 0.05; ** p < 0.01; *** p < 0.001; modelled using a mixed-effects logistic model)

7.3. Secondary Outcome Behavioural Analysis

7.3.1. Skipping of neutral videos

As part of our sensitivity analysis, the same model as in Section 7.2.2 was run on the skipping data of neutral videos. We found no differences in the probability of skipping neutral content across all four arms of the trial.

The observed probability of skipping potentially harmful videos was 70% in the control arm, 70% in Arm 2 – Generic warning, 71% in Arm 3 – High-level descriptive social proof, and 69% in Arm 4 –Specific content warning. There were no significant differences in the odds of skipping neutral videos between the intervention arms and the control arm (Table 10). The whole model, including random effects, accounted for 40% of the variance in the probability of skipping neutral videos.

Table 10. Model-based estimates of the odds of skipping of neutral videos

	Odds Ratios	95% CI	*z*-value	*P*
Intercept	3.71	1.74 – 7.90	3.391	0.001
Arm 2 GW	1.01	0.78 – 1.30	0.067	0.947
Arm 3 HLDSP	1.07	0.83 – 1.38	0.505	0.614
Arm 4 SCW	0.96	0.75 – 1.24	-0.287	0.774

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

Adjusting for multiple comparisons, there were no significant differences in the odds of skipping neutral videos between any arms (see Table 11). Thus, pop-ups that aimed to reduce the probability of viewing potentially harmful videos had no impact on the probability of skipping neutral videos. Figure 14 visualises this by showing the percentage of participants skipping neutral videos, by arm.

Table 11. Estimates of the odds of skipping of neutral videos (p values and CIs corrected for multiple comparisons using the Bonferroni correction)

Comparison	Odds Ratios	95% CI	*z*-value	*P*
Arm 1 Control vs Arm 2 GW	1.01	0.72 – 1.41	0.067	1*
Arm 1 Control vs Arm 3 HLDSP	1.07	0.76 – 1.49	0.505	1*
Arm 1 Control vs Arm 4 SCW	0.96	0.69 – 1.35	-0.287	1*
Arm 2 GW vs Arm 3 HLDSP	1.06	0.76 – 1.48	0.438	1*

Arm 2 GW vs Arm 4 SCW	0.95	0.68 – 1.33	-0.355	1*
Arm 3 HLDSP vs Arm 4 SCW	0.90	0.65 – 1.26	-0.792	1*

* Approximately (rounding error)

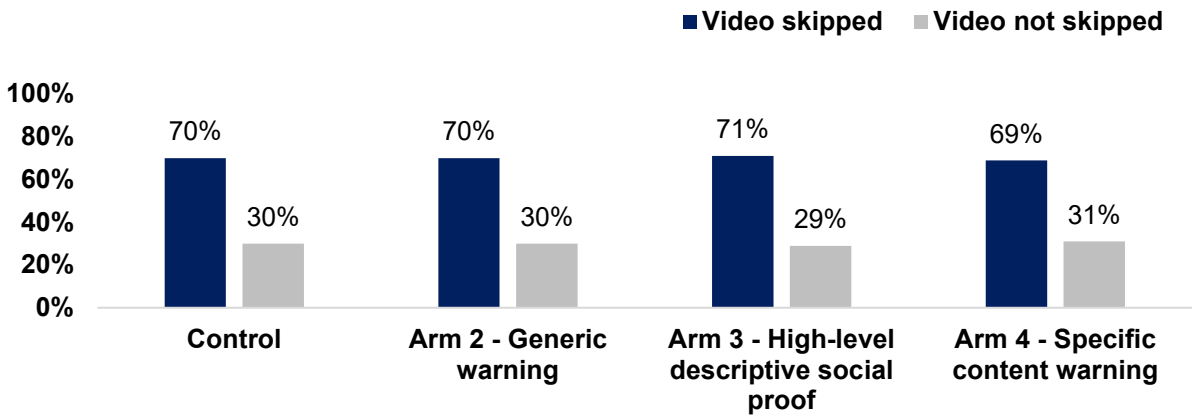


Figure 14. Percentage of skipped and not skipped neutral videos, by arm (multiple comparisons adjusted; modelled using a mixed-effects logistic model)

7.3.2. Reporting of potentially harmful videos

Participants in Arm 1 – Control completed the most reports of potentially harmful videos (n = 230), followed by participants in Arm 2 – Generic warning (n = 186), Arm 4 – Specific content warning (n = 147), and Arm 3 – High-level descriptive social proof (n = 126). Estimates derived from odds ratios reported in Table 8 show that participant in Arm 3 – High-level descriptive social proof had 112% $((2.12 - 1) \times 100 = 112)$ higher odds of not reporting potentially harmful content compared to participants in the control condition.[29] Similarly, participants in Arm 4 – Specific content warning had 98% higher odds of not reporting than participants in the control arm. (Reporting of potentially harmful videos was analysed using a zero-inflated Poisson regression model.)[30]

Table 12. Model-based estimates of not reporting potentially harmful videos (zero-inflated Poisson model)

	Odds Ratios	95% CI	*z*-value	*P*
Intercept	1.50	1.10 – 2.04	2.566	0.010
Arm 2 GW	1.10	0.69 – 1.76	0.396	0.692
Arm 3 HLDSP	2.12	1.33 – 3.38	3.142	0.002
Arm 4 SCW	1.98	1.27 – 3.09	3.022	0.003

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

The differences reported in Table 12 were not sensitive to the adjustment for multiple comparisons (Table 13). Table 13 also shows that there were no other significant differences between arms.

Table 13. Estimates of the odds of not reporting of potentially harmful videos (p values and CIs corrected for multiple comparisons using the Bonferroni correction)

Comparison	Odds Ratios	95% CI	*z*-value	*P*

[29] Note that 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 count of zero.
[30] The logistic mixed-effects model with random intercepts for participants and videos did not converge for this outcome, likely because of the large number of non-reports.

Arm 1 Control vs Arm 2 GW	1.10	0.59 – 2.04	0.396	1*
Arm 1 Control vs Arm 3 HLDSP	2.12	1.15 – 3.91	3.142	0.010
Arm 1 Control vs Arm 4 SCW	1.98	1.11 – 3.55	3.022	0.015
Arm 2 GW vs Arm 3 HLDSP	1.92	0.999 – 3.71	2.565	0.062
Arm 2 GW vs Arm 4 SCW	1.80	0.96 – 3.37	2.416	0.094
Arm 3 HLDSP vs Arm 4 SCW	0.94	0.50 – 1.7	-0.270	1*

* Approximately (rounding error)

Neutral videos

The counts of reports of neutral videos were too low to analyse. Neutral videos were reported only 12 times, by 11 participants (one participant reported 2 neutral videos). Specifically, 1 neutral video was reported in Arm 1 – Control, 4 videos were reported in Arm 2 – Generic warning and Arm 3 – High-level descriptive social proof, and 3 videos were reported in Arm 4 – Specific content warning. Overall, there is no evidence to show that interventions aimed at increasing the probability of skipping of potentially harmful videos, changed the probability of reporting neutral videos.

7.3.3. Time spent watching potentially harmful videos

On average, participants spent longer watching the videos in the intervention arms compared to the control arm (see Figure 15). Figure 15 also shows that participants spent more time watching Video 4 – Misinformation than Video 5 – Violence or Video 6 – Offensive language.

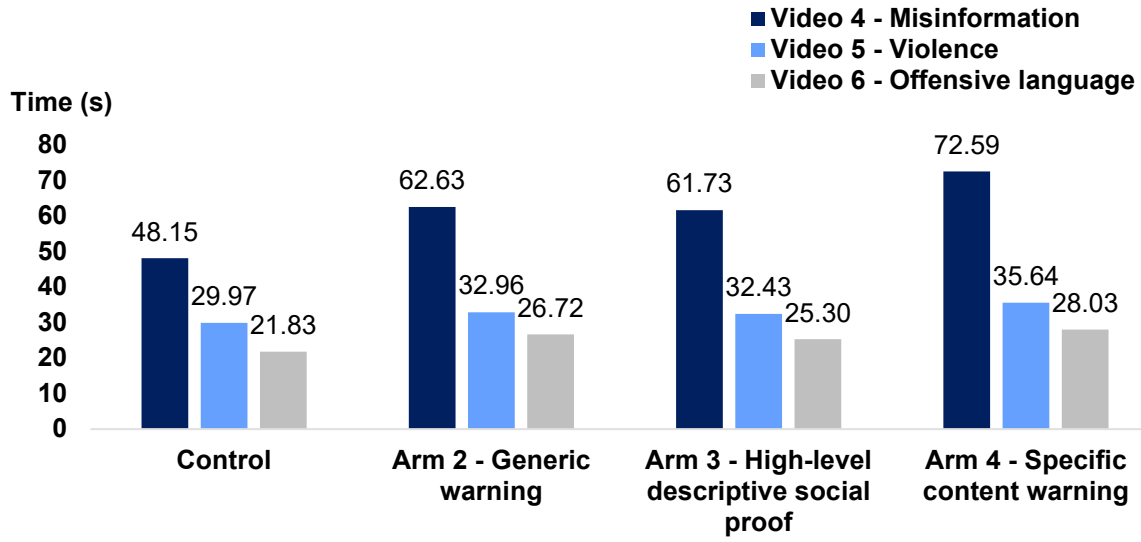


Figure 15. The time participants spent watching each potentially harmful video, in different arms (excluding participants who skipped the potentially harmful video at the pop-up alert stage)

Excluding participants who skipped the videos at the pop-up alert stage, the length of viewing time across potentially harmful videos was found to be higher in the intervention arms compared to the control arm (Table 14). Critically, this effect was not sensitive to the inclusion of age and education as covariates in the model. Thus, it was not merely an artifact of excluding participants who skipped the potentially harmful video at the pop-up alert stage. (The time spent watching potentially harmful videos was analysed using a robust mixed-effects logistic regression model.)

Table 14 Model-based estimates of time spent watching potentially harmful videos

	Estimate	95% CI	*z*-value	*P*
Intercept	27.59	4.37	6.317	0.023

Arm 2 GW	5.12	1.20	4.280	< 0.001
Arm 3 HLDSP	4.27	1.22	3.497	< 0.001
Arm 4 SCW	8.19	1.25	6.554	< 0.001

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

Table 15 shows that, adjusting for multiple comparisons, participants in Arm 4 – Specific content warning spent significantly longer watching potentially harmful videos compared to participants in Arm 3 – High-level descriptive social proof. None of the previously described significant differences between intervention arms and the control arm (reported in Table 14) were sensitive to adjusting for multiple comparisons. Participants in Arm 4 – Specific content warning also did, on average, spend longer watching the potentially harmful videos than did participants in Arm 2 – Generic warning, but this difference was not significant.

Table 15. Estimates of time (in seconds) to spent watching potentially harmful videos (p values corrected for multiple comparisons using the Bonferroni correction)

Comparison	Estimate	95% CI	*z*-value	*P*
Arm 1 Control vs Arm 2 GW	5.12	1.20	4.280	< 0.001
Arm 1 Control vs Arm 3 HLDSP	4.27	1.22	3.497	0.003
Arm 1 Control vs Arm 4 SCW	8.19	1.25	6.554	< 0.001
Arm 2 GW vs Arm 3 HLDSP	0.85	1.28	-0.665	1*
Arm 2 GW vs Arm 4 SCW	3.07	1.30	-2.355	0.111
Arm 3 HLDSP vs Arm 4 SCW	3.92	1.33	-2.954	0.018

* Approximately (rounding error)

Figure 16 shows that the potentially harmful video viewing time was longer in every intervention arm compared to the control arm.

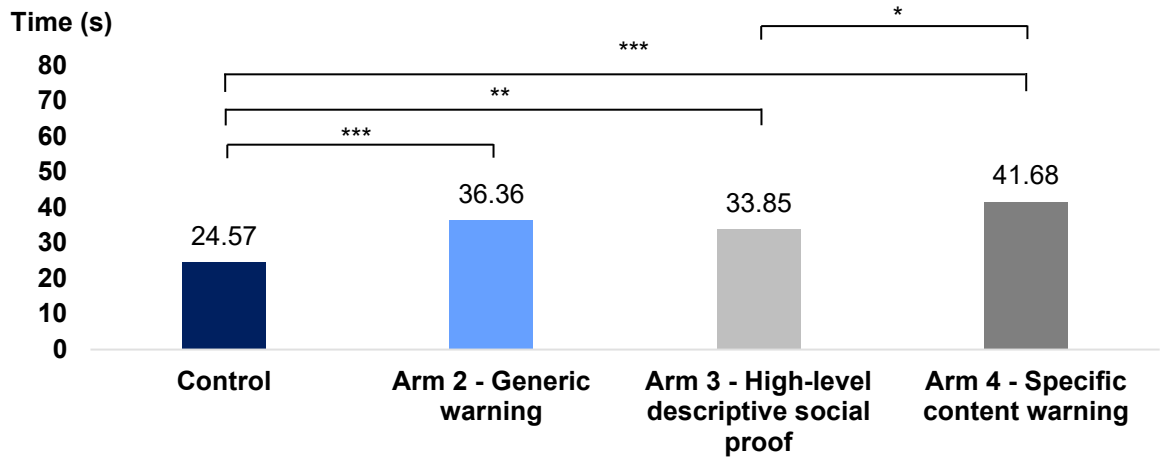


Figure 16. Median time (in seconds) spent watching potentially harmful videos, per arm (multiple comparisons adjusted; * p < 0.05; ** p < 0.01; *** p < 0.001; modelled using a robust mixed-effects linear regression model)

7.4. Responses to attitudinal questions[31]

A Kruskal-Wallis test was conducted to identify if there were significant differences between treatment arms in their perceptions of usefulness of the warning ('*I found the warning messages I just saw useful to me*') (Table 16).

Table 16. Descriptives, '*I found the warning messages I just saw useful to me*'.

	Mean	*Median*	*SE*
Arm 1 Control	-	-	-
Arm 2 GW	3.59	4	0.05
Arm 3 HLDSP	3.66	4	0.04
Arm 4 SCW	3.78	4	0.04

This test highlighted the presence of significant differences in perceptions between arms ($\chi^2 = 8.850$, $p = 0.012$), so Dunn's post-hoc test was conducted (corrected for multiple comparisons using the Benjamini-Hochberg method).

Notably, rates of agreement were significantly higher in Arm 4 – Specific content warning than in Arm 2 – Generic warning (Table 17).

Table 17. Benjamini-Hochberg multiple comparison estimates

	z-value	*P*	*P-adjusted*
Arm 2 GW vs Arm 3 HLDSP	-0.782	0.434	0.434
Arm 2 GW vs Arm 4 SCW	-2.879	0.004	0.012
Arm 3 HLDSP vs Arm 4 SCW	-2.101	0.036	0.054

A significant difference in perceptions of the 'annoyingness' of warnings was also detected ($\chi^2 = 7.072$, p=0.029) (Table 18).

Table 18. Descriptives, '*I found the warning messages I just saw annoying'*.

	Mean	*Median*	*SE*
Arm 1 Control	-	-	-
Arm 2 GW	2.44	2	0.05
Arm 3 HLDSP	2.59	3	0.05
Arm 4 SCW	2.44	2	0.05

Rates of agreement were significantly higher in Arm 3 – High-level descriptive social proof than in Arm 2 – Generic warning (Table 19).

Table 19. Benjamini-Hochberg multiple comparison estimates

	z-value	*P*	*P-adjusted*
Arm 2 GW vs Arm 3 HLDSP	-2.245	0.025	0.037
Arm 2 GW vs Arm 4 SCW	0.108	0.914	0.914
Arm 3 HLDSP vs Arm 4 SCW	2.356	0.018	0.055

[31] Excludes 'Don't know' responses.

Those who were warned about potentially harmful video content but chose to watch the video anyway were asked the extent to which they regretted their choice. To analyse this outcome measure, a variation of a mixed-effects ordinal regression model, cumulative link mixed-effects model (CLMM), was used.[32]

Using CLMM, there were no significant differences in the odds of regretting viewing over all potentially harmful videos (Table 20).

Table 20. Model-based estimates of reported regret of choosing to watch potentially harmful videos

	Odds Ratios	95% CI	*z*-value	*P*
1|2	0.17	0.10 – 0.27	-7.203	< 0.001
2|3	0.83	0.51 – 1.33	-0.785	0.432
3|4	5.32	3.28 – 8.62	6.766	< 0.001
4|5	32.24	19.53 –53.23	13.576	< 0.001
Arm 3 HLDSP	1.13	0.83 – 1.55	0.792	0.429
Arm 4 SCW	0.77	0.56 – 1.06	-1.594	0.111

Note: 1|2 to 4|5 are intercepts, or thresholds, of the ordinal variable regret (1 to 5); the Odds Ratios for Arms 3 and 4 are evaluated against Arm 2

Adjusting for multiple comparisons there were no significant differences in the odds of regret in watching potentially harmful video videos (Table 21).

Table 21. Estimates of reported regret of choosing to watch potentially harmful videos (p values and CIs corrected for multiple comparisons using the Bonferroni correction)

Comparison	Odds Ratios	95% CI	*z*-value	*P*
Arm 2 GW vs Arm 3 HLDSP	1.13	0.78 – 1.65	0.792	1*
Arm 2 GW vs Arm 4 SCW	0.77	0.52 – 1.13	-1.594	0.333
Arm 3 HLDSP vs Arm 4 SCW	1.47	0.998 – 2.18	2.333	0.059

* Approximately (rounding error)

7.5. Engagement with videos

Using data from all participants - including those who have skipped before seeing the content - positive engagement with potentially harmful videos (liking, sharing, commenting) was consistently lower in the treatment arms than in the control arm (see Table 22). However, negative engagement (disliking) was also lower in these arms.

[32] Using ordinal models to model ordinal data enables more accurate estimation of the effects than any model which assumes metric or categorical responses. A cumulative model assumes that the observed ordinal outcome variable represents the categorization of a latent continuous variable. It models this categorization by assuming that there are several thresholds at which the outcome variable is partitioned. This categorization is commonly used to model Likert-scale data, when ordered labels are used to collect judgements about a potentially continuous latent variable. CLMM, a cumulative model with fixed and random effects, was appropriate for two reasons. First, there was a crossed random effects structure: each participant was asked to rate their regret for every potentially harmful video they chose to watch. Second, the outcome variable (regret) can be considered an ordinal scale whereby there is ordering of the levels (from 1 to 5) and an upper (5) and lower (1) limit for each writing task (CLMMs allow to account for the potential ceiling and floor effects imposed by these limits, in a way that standard analyses do not).

Table 22. Engagement with potentially harmful videos, by arm

	Liked	*Shared*	*Commented*	*Disliked*
Arm 1 Control	0.28	0.04	0.08	0.75
Arm 2 GW	0.14	0.02	0.07	0.61
Arm 3 HLDSP	0.17	0.02	0.06	0.51
Arm 4 SCW	0.12	0.02	0.04	0.51

* Approximately (rounding error)

In line with what would be expected given the result outlined in 7.5, excluding participants who skipped the potentially harmful content at the pop-up alert stage, those in the intervention arms viewed the potentially harmful videos for longer (Table 23, see Figure 15 in Section 7.5 for breakdown by each potentially harmful video); view time of neutral videos was more similar (Table 24). Note that the videos were trimmed to be engaging in the first 20-45 seconds of viewing, but the longest potentially harmful video was 3 minutes and 8 seconds long.

Table 23. Mean and median viewing time (in seconds) of potentially harmful videos, by arm

	Mean	SD	Median
Arm 1 Control	33.32	39.95	24.57
Arm 2 GW	40.57	43.88	36.36
Arm 3 HLDSP	39.31	43.16	33.85
Arm 4 SCW	44.85	45.72	41.68

Table 24. Mean and median viewing time (in seconds) of neutral videos, by arm

	Mean	SD	Median
Arm 1 Control	25.39	26.48	17.86
Arm 2 GW	25.41	27.55	16.84
Arm 3 HLDSP	25.73	27.88	16.34
Arm 4 SCW	26.26	27.78	17.86

7.6. Exploratory analyses of skipping behaviour

We also conducted additional exploratory analyses. We did this for two reasons. First, to investigate at what point in the user journey participants skipped the potentially harmful videos in the intervention arms. Second, to examine whether skipping behaviour varied by engagements, such as liking or disliking a potentially harmful video. However, this research was not powered to perform these exploratory analyses. As such, any conclusions drawn from the reporting in this section should not be interpreted as representative of the population of panel members who are VSP users.

Participants who were not exposed to an alert message skipped 1,249 potentially harmful videos. Those exposed to generic warnings skipped 1,294, those exposed to high-level descriptive social proofs skipped 1,359, whereas those exposed to specific content warnings skipped 1,323. Figure 17 shows that the majority of skips by participants who were exposed to generic warning and high-level descriptive social proof occurred at the interface stage. In contrast, for participants in the specific content warning arm, the proportion of skips at the pop-up alert stage was nearly the same as the proportion of skips at the interface stage. Thus, although the number of skips of potentially harmful videos was highest amongst participants exposed to high-level descriptive social proof, it appears that the specific content warning was most effective at encouraging participants to skip at the pop-up alert stage.

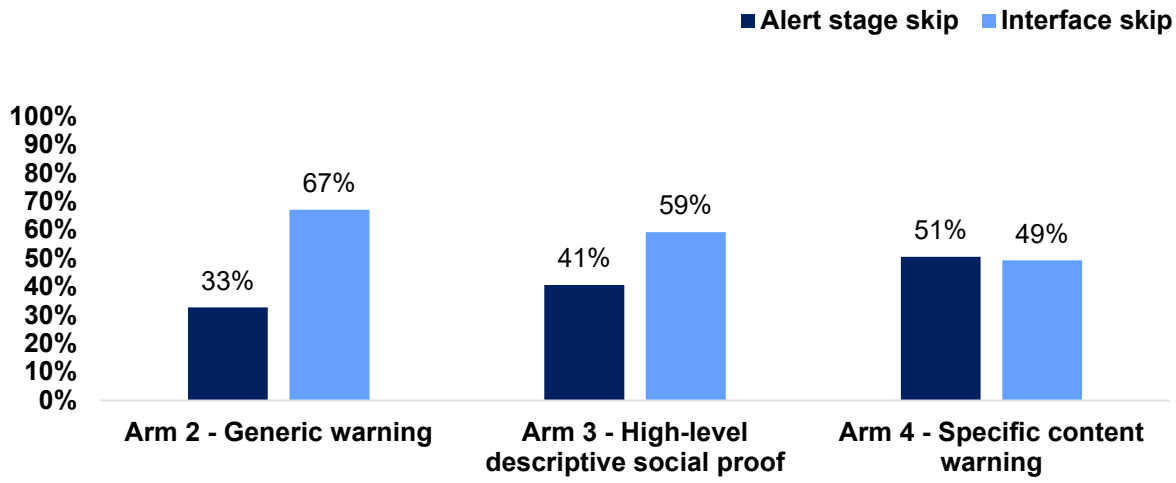


Figure 17. The proportion of skips at the alert stage (of all skips), per intervention arm

Figures 18 shows that Video 4 – Misinformation, the longest of the three potentially harmful videos (188 seconds), was most likely to be skipped at the interface stage. In comparison, Video 5 – Violence was least likely to be skipped at the alert stage. Note that Video – 5 Violence was longer (47 seconds) than Video 6 – Offensive language (42 seconds), but the difference in the length of the videos is much smaller than the difference between these two videos and Video 4 – Misinformation. This suggests that length of a video is likely to be a contributing factor as to whether the video is skipped, but that other factors that we did not consider (for example, preferences of the sample on average), also may matter.

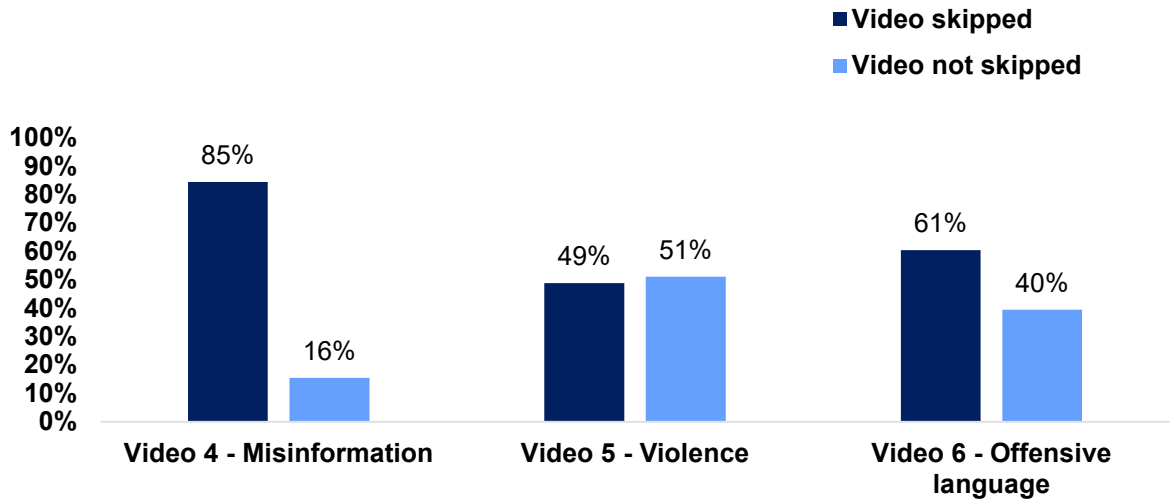


Figure 18. The proportion of videos that were skipped (n = 5,555), per video (excluding observations of videos skipped at the alert stage)

Figures 19 and 20 show that the probability of skipping potentially harmful videos at the interface stage did not vary, in a meaningful way, depending on whether participants liked or disliked potentially harmful videos. Across the three potentially harmful videos, 60% of the potentially harmful videos were skipped by participants who first liked them at the interface stage, whereas 62% were skipped by participants that first disliked them. These estimates are comparable to the 64% of potentially harmful videos skipped at the interface stage, of all potentially harmful videos that were skipped (Figure 18 shows this split by video). Thus, Figures 18-20 show that there is variation in the probability of skipping between videos. However, there is no evidence that there is a difference in the probability of skipping potentially harmful videos depending on whether these videos are liked or disliked. Note that the number of videos that were liked prior to being

skipped is very low (431 out of 7,203), and the number of videos that were disliked prior to being skipped is also low relative to the whole sample (1,421). So, Figures 18 and 19 show suggestive evidence at best.

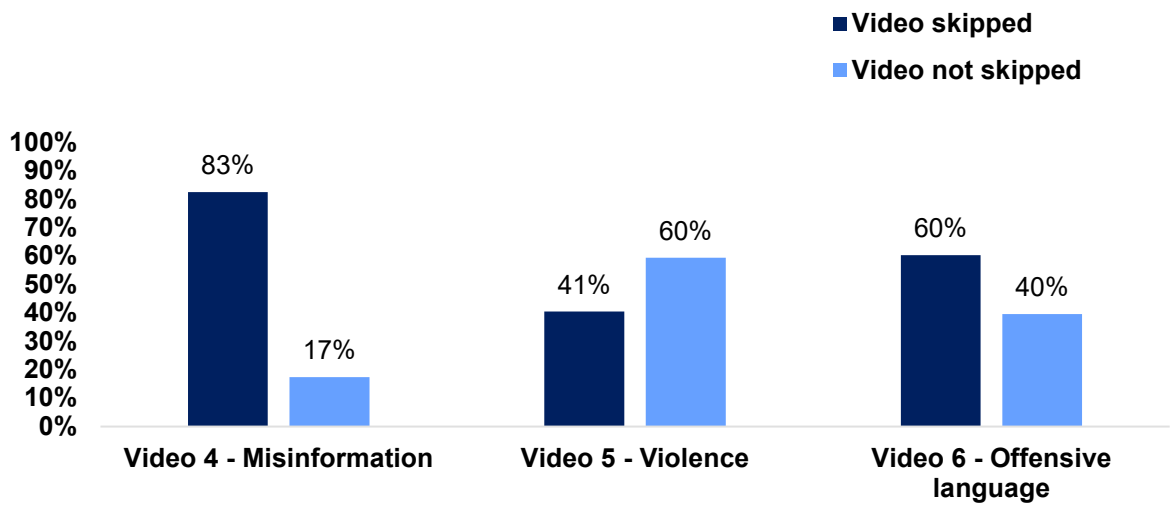


Figure 19. The proportion of videos that were liked and skipped (n = 431), per video (excluding observations of videos skipped at the alert stage)

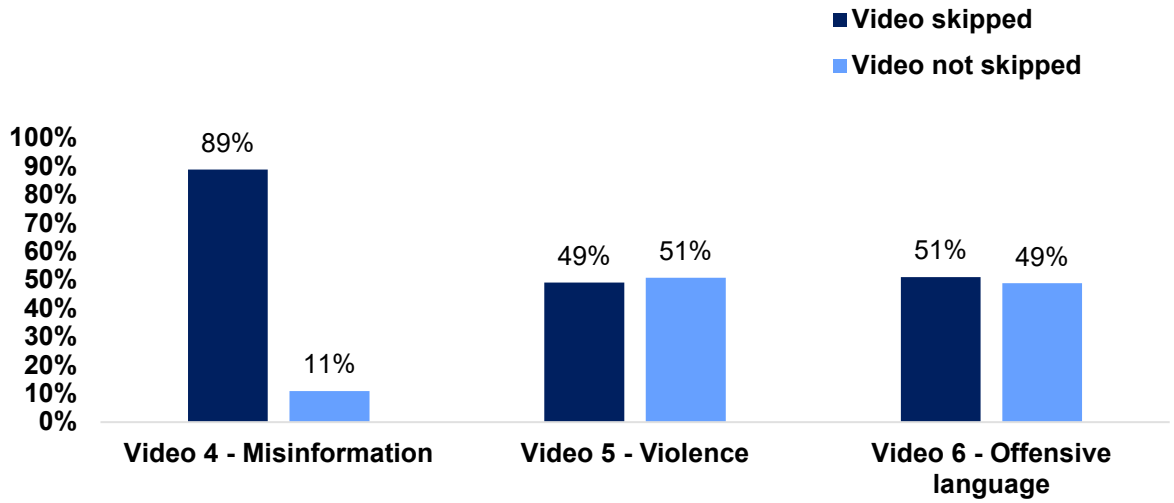


Figure 20. The proportion of videos that were disliked and skipped (n = 1,421), per video (excluding observations of videos skipped at the alert stage)

8. Comments

Adding pop-up alerts with high-level descriptive social proof message before each potentially harmful video significantly reduced the likelihood of users watching these videos. This finding was robust to sensitivity analyses that adjusted for devices on which the experiment was completed and whether the potentially harmful videos were played by participants. Pop-up alerts employing generic warnings or specific content warnings were also effective, but only under ideal conditions that were not representative of all participants' experience of interacting with the simulated VSP (due to differences in auto-play functionality between devices and operating systems). Thus, it is plausible that they are less likely to work in the real-world context than pop-ups employing high-level descriptive social proof.

Critically, we found that none of the pop-up alerts had any effect on the probability of skipping of neutral videos. Consequently, there is no evidence to suggest that the pop-up alerts used in this study suppressed engagement with neutral content. High-level descriptive social proof and specific content warning pop-up alerts were found to decrease the probability of reporting potentially harmful videos, but this is confounded by the fact that fewer participants exposed to these pop-up alerts chose to watch potentially harmful videos.

Participants who were exposed to pop-up alerts and chose to watch potentially harmful videos were found to watch these videos for longer than participants who were not exposed to them. This may be because participants who would have stopped watching quickly when they saw the content, were alerted to this by the pop-up alerts and skipped. In contrast, those not exposed to pop-up alerts, but who would have skipped before viewing potentially harmful videos given a pop-up alert, could only skip at the interface stage. It is plausible that those who chose to watch potentially harmful videos at the pop-up alert stage were more interested in such videos—and would therefore spend longer watching them—than those who decided to skip. Consequently, the samples of participants watching potentially harmful videos in the intervention arms and the control arm might have been fundamentally different. Specifically, the sample of participants in the control arm consisted of a mixture of participants, some of whom were interested in watching potentially harmful videos but some of whom were not, and therefore skipped relatively quickly at the interface stage. In contrast, the sample of participants in the intervention arms likely consisted of a higher proportion of participants who were interested in seeing potentially harmful videos, relative to participants in the control group.

Finally, our exploratory analyses suggest that whether someone has liked or disliked a video is not a good indicator of whether or not they will skip it. Instead, people may be making skipping decisions based on other properties of the video, like its length. However, this evidence is speculative and further research would have to be carried out to examine its validity.