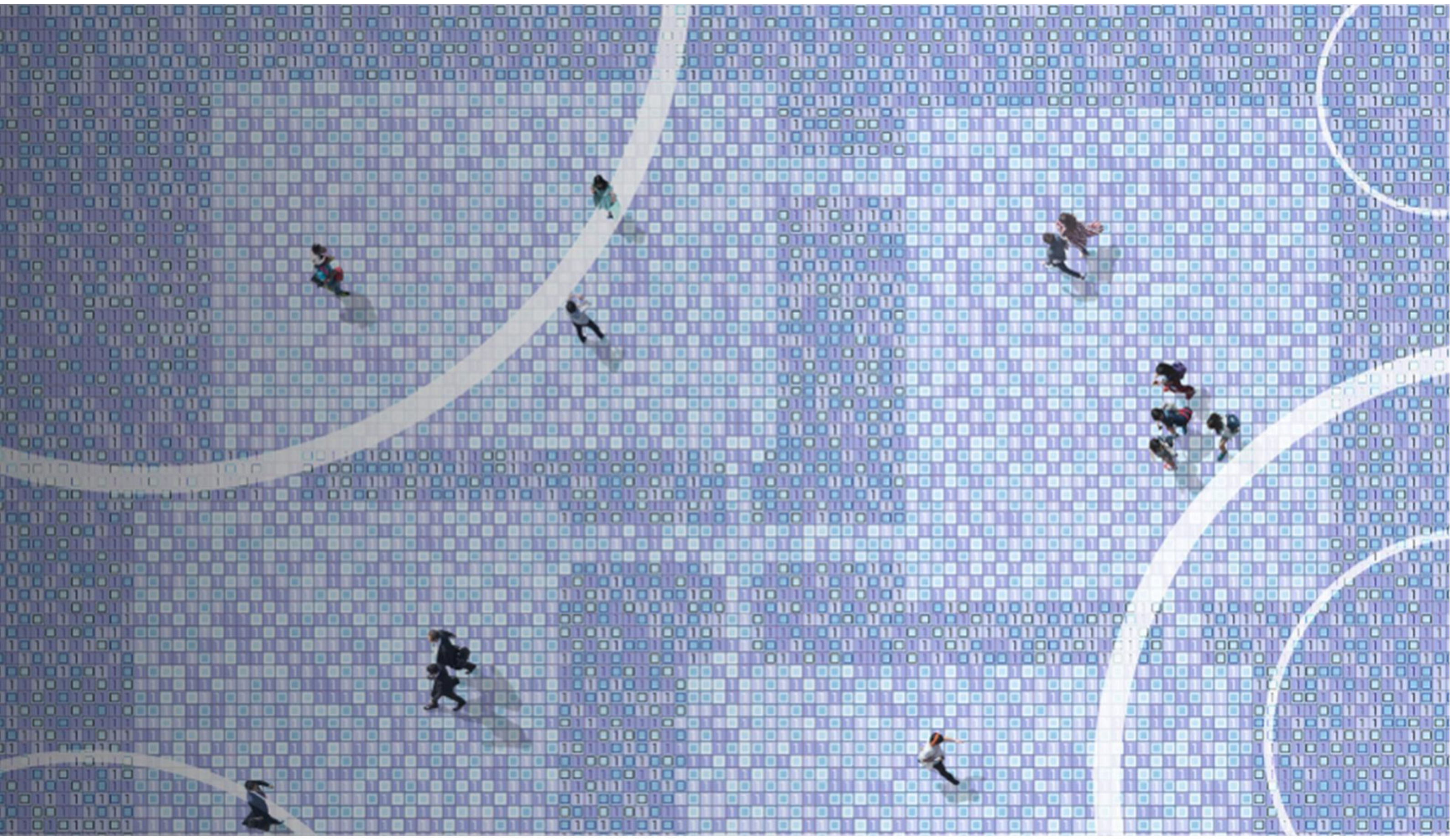
Boosts for online safety: Microtutorial trial

Report

Kantar Public Behavioural Practice: Michael Ratajczak, Rupert Riddle, Yuchen Yang, and Natalie Gold.

Ofcom: Rupert Gill, Amy Hume, and Pinelopi Skotida.



1. Background and objectives

1.1. Regulatory Context

Ofcom has a duty to promote media literacy, including in respect of material available on the internet. Ofcom's approach to media literacy is multi-dimensional and considers a range of aspects including how the design of services can impact on users' ability to participate fully and safely online.

In addition, as of November 2020, Ofcom oversees the regulatory regime which requires UK-established Video Sharing Platform (VSP) providers to include measures and processes in their services that protect users from the risk of viewing harmful content. Measures taken by a provider must be appropriate for the purposes of protecting users and must be effective in achieving this purpose. However, there is limited research in the public domain about their effectiveness.

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, discovering related information after watching an educational video).

However, there can be content on these platforms that is harmful to users. Ofcom is looking to research the effectiveness of different safety measures used by online platforms to safeguard users from harm more broadly and looking to build an evidence base about the effectiveness of specific techniques or interventions.

Users need to be able to make informed decisions about how to use platforms safely. One component of this is having the capability to use platforms safely, including knowing how to report or not report harmful content, as well as skip, deploy parental controls if relevant, and understand the terms and conditions of the platform.

Using behavioural science to foster people's competence to make their own choices is a sometimes called a 'boost'. One type of boost is a 'microtutorial', or short training – in this instance designed to build peoples' skills in using the features of the platform, especially targeting the skills needed to use platforms safely. How a microtutorial is delivered to users will also influence how much people learn and how they use that knowledge.

1.2. Experiment aims and objectives

In the context of online safety, a number of platforms already employ reporting functions which allow VSP users to flag content as inappropriate or harmful. In this research, the focus was on testing the impact of interventions that could support users to submit reports about content that they consider potentially harmful. Specifically, it was of primary interest to investigate what impact different types of microtutorial have on the probability of reporting potentially harmful videos, compared to a control arm with no tutorial. It was also of interest to Ofcom to test which type of microtutorial is most effective at encouraging users to submit reports of potentially harmful video content.

1.3. Research questions

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

RQ.1. Do microtutorials increase the probability of reporting harmful content?

RQ.2. Which microtutorial affects the probability of reporting harmful content the most?

> RQ2a. Does a static microtutorial (significantly) affect the probability of reporting harmful content?
>
> RQ2b. Does a video microtutorial (significantly) affect the probability of reporting harmful content?
>
> RQ2c. Does an interactive microtutorial (significantly) affect the probability of reporting harmful content?

2. Sample and data collection

2.1. Sample

The target population for this study was the UK population, with demographic quotas set based on the adjusted quotas used in previous research with Ofcom. Specifically, the quotas in this trial were based on the relative proportions of respondents in each demographic sub-group who used VSPs at least once in the 12 months prior to participating in the Reporting Online trial.[1] Note that participants who had not used VSPs in the past 12 months (i.e., the 12 months immediately prior to the date they participated in this trial) were included in this trial, unlike previous trials. However, participants who participated in the Reporting,[1] Alerts,[2] and Defaults online trials were not allowed to participate.

A total of 2,862 UK participants, aged between 18 and 69, were recruited for this experiment. Given the exclusion of 7,600 participants who completed the three previous experiments, recruitment of participants using only Kantar's LifePoints panel was not feasible. As a result, additional panels were used for participant recruitment during this study, using Kantar Profiles platform: 2,071 participants were recruited using Kantar's LifePoints panel, 516 were recruited using Qmee, and 275 participants were recruited using Marketblend.[3] There was a risk that participants may sit on more than one panel and could participate in the study more than once. In Kantar's experience this tends to be a very low risk but still a risk. To mitigate for the risk of duplicates, the sample size was increased from 2,800 to 2,862, under the assumption that we would not have more than 60 duplicates, and any identified would be removed. Analysis following the experiment did not identify any duplicates using Kantar Profiles' identification method for duplicate accounts.[4] Thus, the additional observations were treated as normal.

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 2023 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 quotas set at the start of the study, and the quotas achieved

Demographics		Start	Finish
Gender	Male	49%	49%
	Female	51%	50%
	Other	-	<1%
	Prefer not to say	-	<1%
Age	18–24	14%	14%
	25–39	34%	34%
	40–54	30%	30%
	55–69	22%	22%
Ethnicity	White	87%	86%
	Mixed/Multiple Ethnic Groups	2%	2%
	Asian/Asian British	7%	6%
	Black/African/Caribbean/ British	3%	3%
	Other Ethnic Group	1%	1%
	Prefer not to say[5]	-	2%
Socio-economic	ABC1	56%	56%

[1] https://www.ofcom.org.uk/__data/assets/pdf_file/0020/241832/Online-Trials-Appendix-2-Reporting-Mechanisms.pdf

[2] https://www.ofcom.org.uk/__data/assets/pdf_file/0021/241833/Online-Trials-Appendix-1-Alert-Messages.pdf

[3] There are no known differences between the panel populations of Lifepoints, Qmee, and Marketblend. Additionally, participants were recruited using least-fill randomisation. Thus, any difference in panel population would not be expected to impact inference because participants from the three panels should be equally distributed across arms.

[4] Kantar Profiles rely on identification of duplicate accounts using: email addresses, cookies, and a RelevantID (RelevantID® – Imperium). RelevantID gathers a large number of data points from a respondent's computer, such as operating system version, browser version, plug-in, etc., and assigns a relative weight to each data point. The data gathered is put through deterministic algorithms to create a unique digital fingerprint of each computer. The digital fingerprint identifies duplicate respondents who take the same survey more than once from the same machine. RelevantID flags a computer each time a user tries to take a survey, so it is able to detect if multiple email accounts are being used to take surveys from a single computer. In addition, RelevantID has the unique ability to identify multiple panel accounts from different research firms on the same computer. Suspect respondents are flagged in the system and, based on business rules, are either allowed, redirected or completely filtered out of surveys in which they attempt to participate. Thus, a responded could be signed up to different panels, and the panel may not know about it, but they would not be allowed to participate if they used the same email address or if they were identified as a duplicate by RelevantID.

[5] Includes participants who did not agree to be asked this question (n = 40) and those who refused to answer this question when asked (n = 7).

grade	C2DE	44%	44%
Country	England	84%	84%
	Wales	5%	5%
	Scotland	8%	8%
	Northern Ireland	3%	3%

2.2. Data collection

Kantar Public adhered to 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. Participants were notified at the beginning of the study that they may 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 and potentially harmful videos were reviewed by Kantar Public's Profiles' Privacy team and Kantar Public's Global Head of Quality, Information and Security. In addition, this team assessed and documented what data would be collected, how it would be collected and determined that the data would be collected in compliance with Profiles' data protection framework.

2.3. Randomisation

Participants were randomly allocated into one of the experiment's four arms, three of which included microtutorials aimed at boosting users' capability to use platform functions including reporting potentially 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 studies,[6] as well as in clinical trials,[7] and was successfully used to recruit participants in the previous three online behavioural RCTs for Ofcom.

2.4. Incentivisation

Panel participants who were recruited via LifePoints were paid a sum of '*LifePoints'* for completing the experiment. These points can be exchanged for vouchers or cash via PayPal. Participants recruited via Qmee were paid in cash for completing the experiment. We are unable to confirm how participants recruited via Marketblend were incentivised for this study, though most panel providers incentivise participants via cash or vouchers.

2.5. Ethics

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 was not possible to do this in an ecologically valid manner without exposing them to potentially harmful content. However, all content was legal and we mitigated against the risks as follows.

Kantar Public's Behavioural Practice team used the same videos as the Defaults Online Trial. The potentially harmful video content (content that some participants could consider to be harmful) was selected for inclusion by:

1. Searching various VSPs for legal but potentially harmful videos that have been made downloadable by their originators so they could be downloaded directly from the website.
2. Sharing these videos between the Kantar Public's 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 acceptable for inclusion because of the content's lower impact and greater prevalence and hence likelihood of being seen 'for real'. Ofcom's own research indicates that over six in ten (62%) of internet users have experienced at least one instance of potentially harmful behaviour or content online in the last four weeks.[8]

The following steps were taken to mitigate any residual risk of harm to respondents 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

[6] 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.

[7] For example: https://onlinelibrary.wiley.com/doi/full/10.5694/j.1326-5377.2002.tb04955.x

[8] Ofcom Online Experiences Tracker Summary Report (2022).

to participate if they did not want to be exposed to this.

- 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.
- 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.

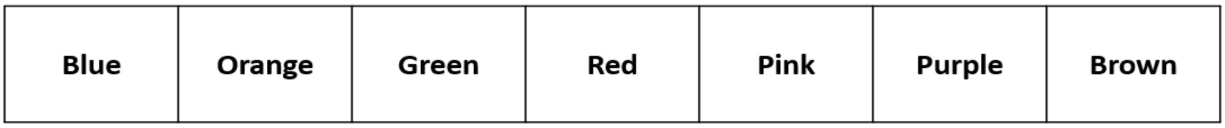
2.5. Disclaimer

Ofcom had no role in identifying, or selecting, the potentially harmful video content selected by Kantar Public for use in this study. 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 adheres to the Market Research Code of Conduct 2019.

2.6. Attention tests

Kantar Profiles conducted a range of quality and validation checks when recruiting panellists.[9] In addition, two attention checks were included in this experiment. First, any respondent, who completed the study in less than 40% of the median completion time of all respondents, was removed. Second, any respondent who failed to correctly answer the attention check question below was excluded from the study. The attention check question specified: "Please select the "Orange" 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:



Blue	**Orange**	**Green**	**Red**	**Pink**	**Purple**	**Brown**

We found that 89 participants failed the attention check during this experiment, and a further 161 were excluded due to completing the study too quickly.

2.7. 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. We identified no significant data capture issues at this stage of data collection.

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

3. Trial design and flow

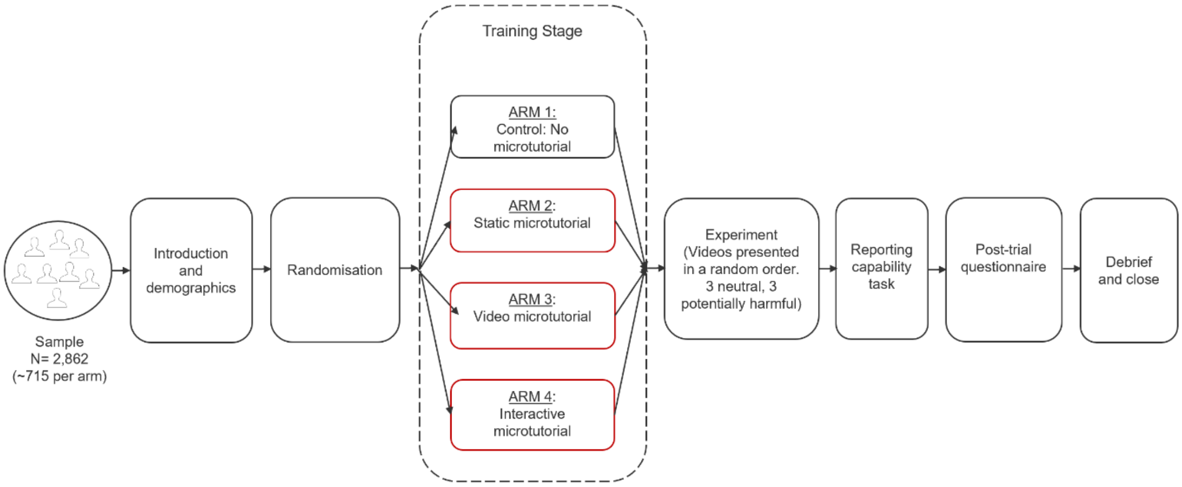

Training Stage
Sample
N= 2,862
(~715 per arm)
Introduction and demographics
Randomisation
ARM 1: Control: No microtutorial
ARM 2: Static microtutorial
ARM 3: Video microtutorial
ARM 4: Interactive microtutorial
Experiment (Videos presented in a random order. 3 neutral, 3 potentially harmful)
Reporting capability task
Post-trial questionnaire
Debrief and close

Figure 1. 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 contain 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, and 186 participants chose to opt out of the experiment at this stage (Figure 2 shows the participant flow).

There was also a debrief screen at the end of the experiment which provided a link (https://saferinternet.org.uk/report-harmful-content) to support on any of the potential harms included in the content shown in the experiment.

3.2. Demographics

On entry to the trial, participants were asked demographic questions to so that recruitment could be monitored against quotas of interest (age, gender, socioeconomic background, location, and ethnicity). Participants were also asked how often they use video sharing platforms.

3.3. Microtutorial stage

Once participants confirmed their demographics, participants were randomly allocated to experimental blocks. Participants in the control condition (Arm 1) received no information about how to use functions in the platform, whereas participants in the three intervention arms were required to complete one of three microtutorials (participants were exposed to a different microtutorial depending on their intervention arm), as specified in Section 4. Note that the Microtutorial was an adaptation of the "Training Stage" in the previous Online Trials on Alerts and Reporting.[10]

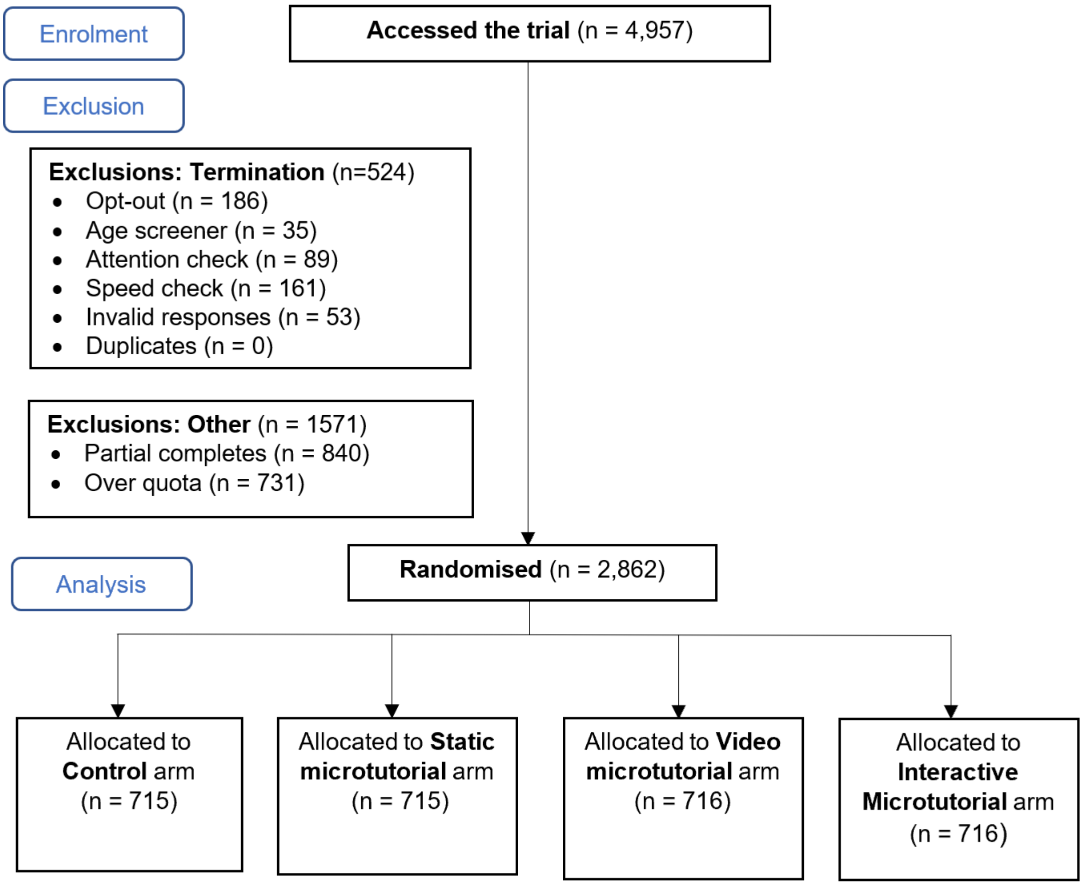
After completing the microtutorial assigned to them, all participants, regardless of arm assignment, were presented with a short prompt before continuing to the main section of the experiment. This prompt is shown in Figure 3.

3.4. Randomised videos

The experimental design aimed to examine whether, after completing a microtutorial, participants were more likely to report potentially harmful video content, compared to participants who did not complete a microtutorial. The interface for this experiment was a variation of a generic VSP simulation (the salience intervention, Arm 2, in the Reporting Online Trial),[11] incorporating features that are common to many platforms but without any familiar branding.

Participants were shown six videos presented in a random order within the simulated VSP interface. Three pieces of content were neutral, and three contained content that users may find potentially harmful.

[10] https://www.ofcom.org.uk/research-and-data/economics-discussion-papers/understanding-the-impact-of-vsp-design-on-user-behaviour
[11] https://www.ofcom.org.uk/__data/assets/pdf_file/0020/241832/Online-Trials-Appendix-2-Reporting-Mechanisms.pdf



Enrolment

Exclusion

Accessed the trial (n = 4,957)

Exclusions: Termination (n=524)
- Opt-out (n = 186)
- Age screener (n = 35)
- Attention check (n = 89)
- Speed check (n = 161)
- Invalid responses (n = 53)
- Duplicates (n = 0)

Exclusions: Other (n = 1571)
- Partial completes (n = 840)
- Over quota (n = 731)

Analysis

Randomised (n = 2,862)

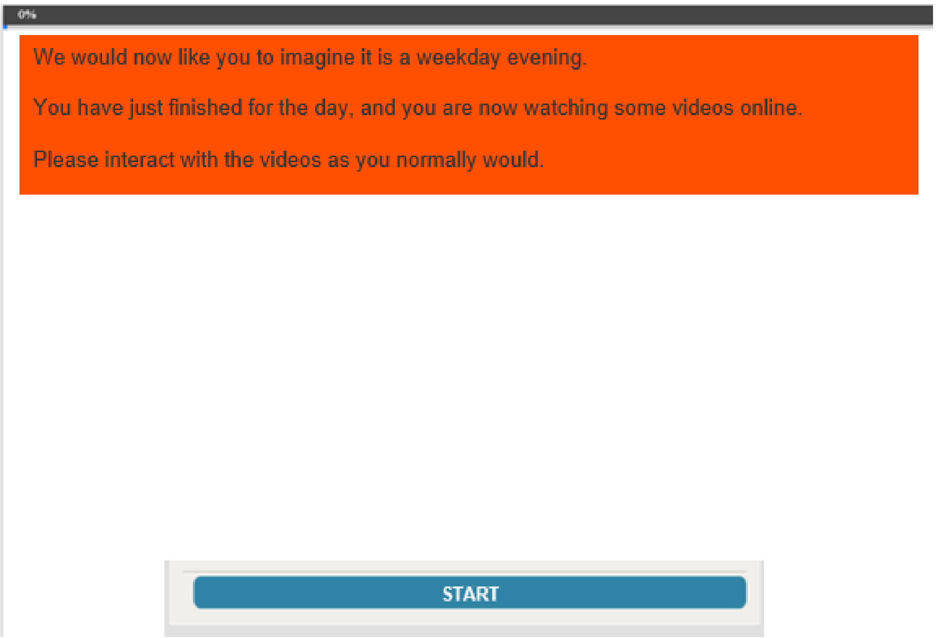
Allocated to **Control** arm (n = 715)

Allocated to **Static microtutorial** arm (n = 715)

Allocated to **Video microtutorial** arm (n = 716)

Allocated to **Interactive Microtutorial** arm (n = 716)

Figure 2. Participant flow diagram[12]



We would now like you to imagine it is a weekday evening.

You have just finished for the day, and you are now watching some videos online.

Please interact with the videos as you normally would.

Figure 3. Experimental Prompt

[12] 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.

Video content:

Neutral One: Optical Illusion: https://odysee.com/@AussieFighter:8/Optical-Illusion-chair:8

Neutral Two: How to do a pullup: https://odysee.com/@LiquidLoans:0/how-to-do-a-pull-up-tutorial:9

Neutral Three: Vegan Matcha Pancakes: https://vimeo.com/23873736[13]

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 / Offensive language (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 participants' attention and to ensure that videos were not over a minute in duration.

3.5. Reporting capability test

After being shown the six videos in the main experimental task, all participants were shown another potentially harmful video, and were prompted to report it.

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

The prompt read: "Please imagine that you have decided to report the following video. Please use the buttons within the interface to complete your report."

The purpose of this was to find out whether participants knew how to report potentially harmful videos. Participants had the option to skip this content rather than reporting it, however if they choose to skip, they were asked to select one of four options to indicate why:

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

3.6. Post-trial questionnaire

Participants were asked about their experiences of using the platform and, where relevant, about attitudes and beliefs relating to the microtutorial they experienced. Responses to these questions were used as secondary outcome measures to help us understand more about the potential impact microtutorials on participants (Section 7.5).

[13] Hyperlinks to three videos (Neutral Three, Potentially Harmful Two, and Potentially Harmful Three) do not work (last checked on the of 22nd of March 2023). Since the work was conducted, Neutral Three: Vegan Matcha Pancakes has been deleted. In addition, https://leakreality.com has been taken down. Thus, it is not possible to provide working hyperlinks to these videos.

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 is a generic version of a VSP, with a pre-experiment prompt (Figure 3). Participants did not receive any instructions about how to interact with the VSP interface. Participants were simply asked to click "start" to begin the main experiment.
2. *Arm 2 – Static microtutorial*: participants saw three static text-based screens of a microtutorial that informed the user about how to interact with functions of the VSP interface, highlighting the "report", "like", "dislike", "share", "skip" and "add comment" features. Participants had to click through each screen of the static microtutorial as shown in Figures 4-6. After completing the static microtutorial, the experiment progressed to the next stage, which was the pre-experiment prompt shown in Figure 3. The interface of the platform, other than the inclusion of the static microtutorial prior to beginning the task, was the same as in Arm 1.
3. *Arm 3 – Video microtutorial*: Participants were required to watch a video that explained how they could interact with the VSP environment they were faced with. This video explained how participants could interact with functions of the interface as shown in Figure 7. Participants could not progress to the next stage of the experiment until after the video ended. After the video ended the participant was able to progress to the next screen by clicking 'continue'. Doing so progressed the user to the pre-experiment prompt shown in Figure 3. The interface of the platform, other than the inclusion of the video microtutorial prior to beginning the task, was the same as in Arm 1 and 2.
4. *Arm 4 – Interactive microtutorial*: Participants had to participate in an interactive microtutorial which explained how to interact with each element on the VSP interface. Specifically, participants were required to complete each highlighted action (e.g., "pausing the video") to simulate performing this action, before moving on to the next stage of the microtutorial as shown in Figures 8 and 9. At each stage of the microtutorial, participants could only complete one action at a time. Once participants completed the interactive microtutorial, the experiment progressed to the next stage -- the pre-experiment prompt is shown in Figure 3. The interface of the platform, other than the inclusion of the interactive microtutorial prior to beginning the task, was the same as in Arm 1, 2, and 3.

4.1. Hypotheses

Hypothesis 1: Having a static microtutorial will increase the probability of reporting potentially harmful video content (compared to the control arm).

- The probability of reporting potentially harmful video content will be significantly higher in an arm with static microtutorial (Arm 2) compared to an arm with no microtutorial (control arm)

Hypothesis 2. Having a video microtutorial will increase the probability of reporting potentially harmful video content (compared to the control arm).

- The probability of reporting potentially harmful video content will be significantly higher in an arm with a video microtutorial (Arm 3) compared to the control arm

Hypothesis 3: Having an interactive microtutorial will increase the probability of reporting potentially harmful video content (compared to the control arm).

- The probability of reporting potentially harmful video content will be significantly higher in an arm with an interactive microtutorial (Arm 4) compared to the control arm

Hypothesis 4. Having a video microtutorial will increase the probability of reporting potentially harmful video content (compared to static microtutorial).

- The probability of reporting potentially harmful video content will be significantly higher in an arm with a video microtutorial (Arm 3) compared to an arm with static microtutorial (Arm 2)

Hypothesis 5. Having an interactive microtutorial will increase the probability of reporting potentially harmful video content (compared to static microtutorial).

- The probability of reporting potentially harmful video content will be significantly higher in an arm with an interactive microtutorial (Arm 4) compared to an arm with static microtutorial (Arm 2)

Hypothesis 6. Having an interactive microtutorial (Arm 4) will increase the probability of reporting potentially harmful video content compared to a video microtutorial (Arm 3).

- The probability of reporting potentially harmful video content will be significantly higher in an arm with an interactive microtutorial (Arm 4) compared to an arm with a video microtutorial (Arm 3)

4.2. Intervention designs

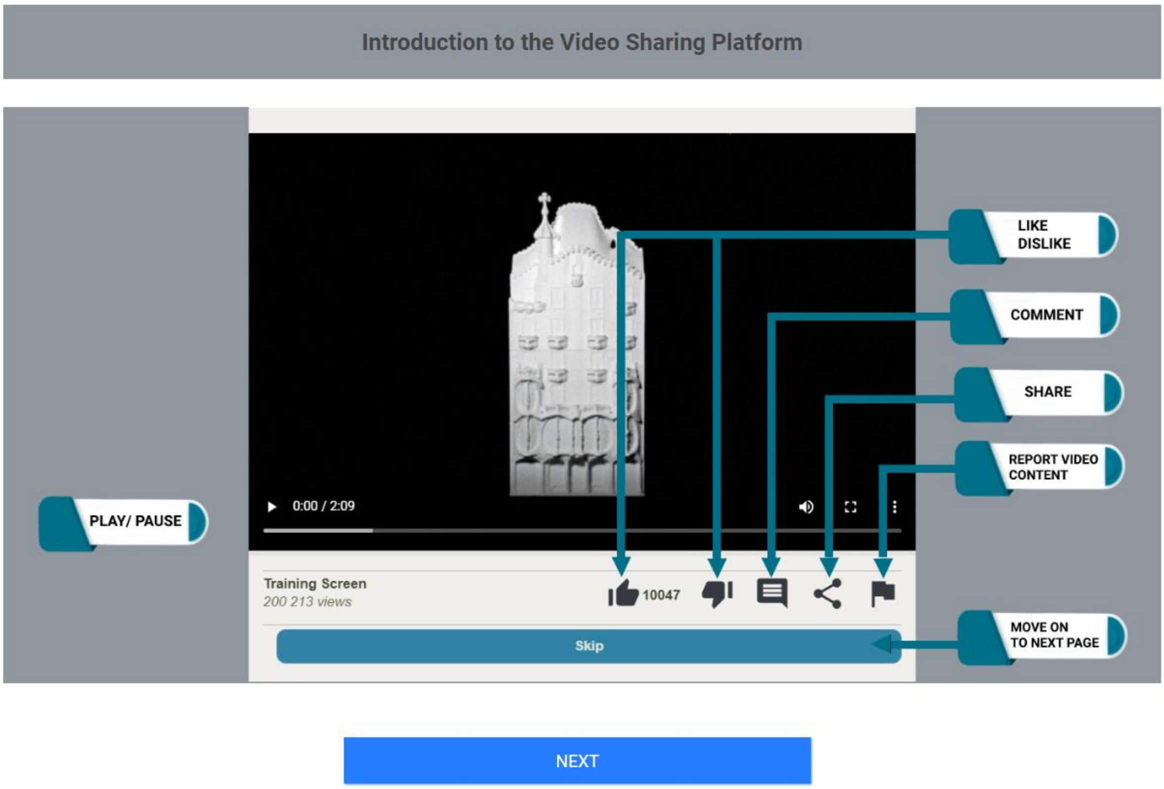


Figure 4. Arm 2 – Static microtutorial: The first screen of the interface participants were shown prior to starting the main experiment. The first screen of the interface was similar to the control arm used in previous online trials. In this experiment, we added two screens (see Figure 5 and Figure 6), following the first screen shown here, to introduce commenting and reporting features of the platform.[14]

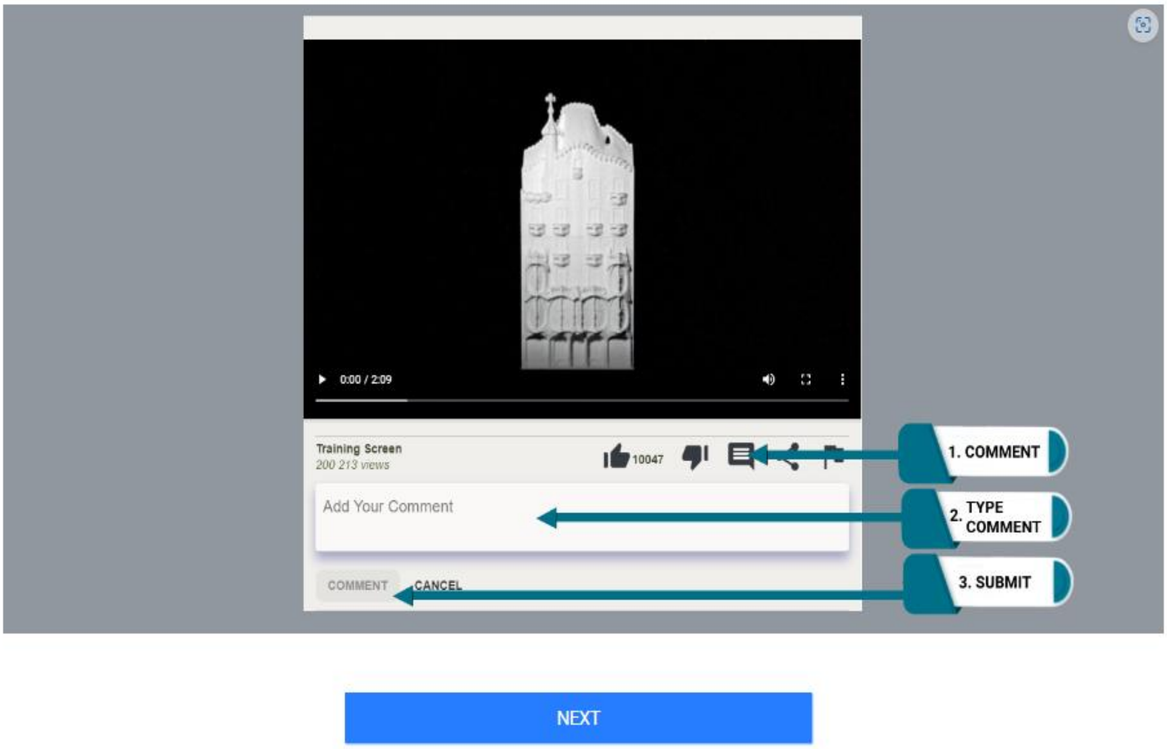


Figure 5. Arm 2 – Static microtutorial: commenting screen.

[14] After completing the respective microtutorial, participants were shown the prompt shown in Figure 3 prior to starting the main experiment. Participants in the control arm were shown the prompt straight away.

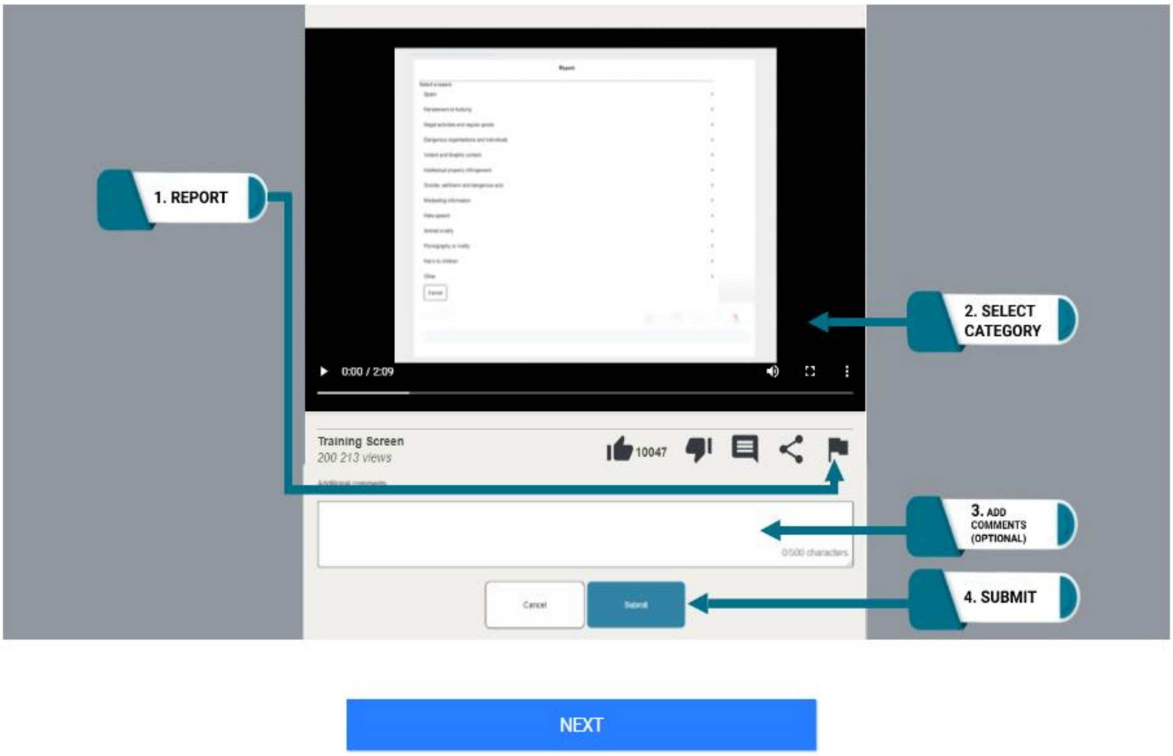


Figure 6: Static microtutorial: reporting screen.

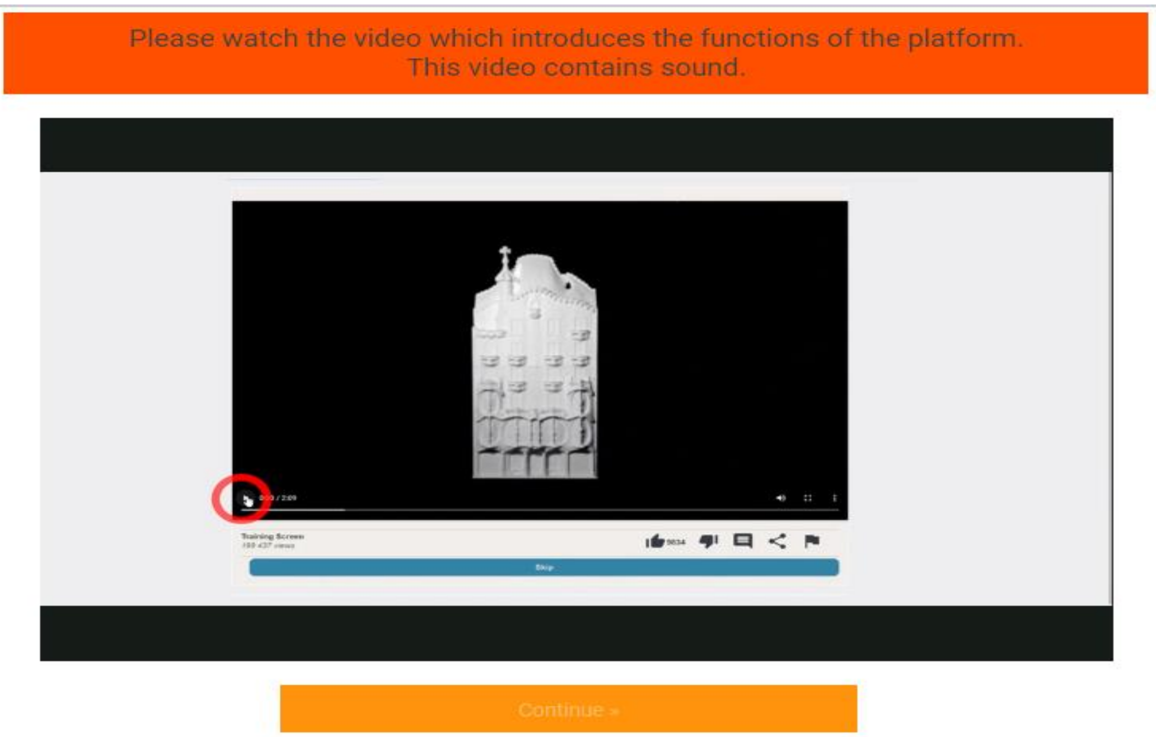
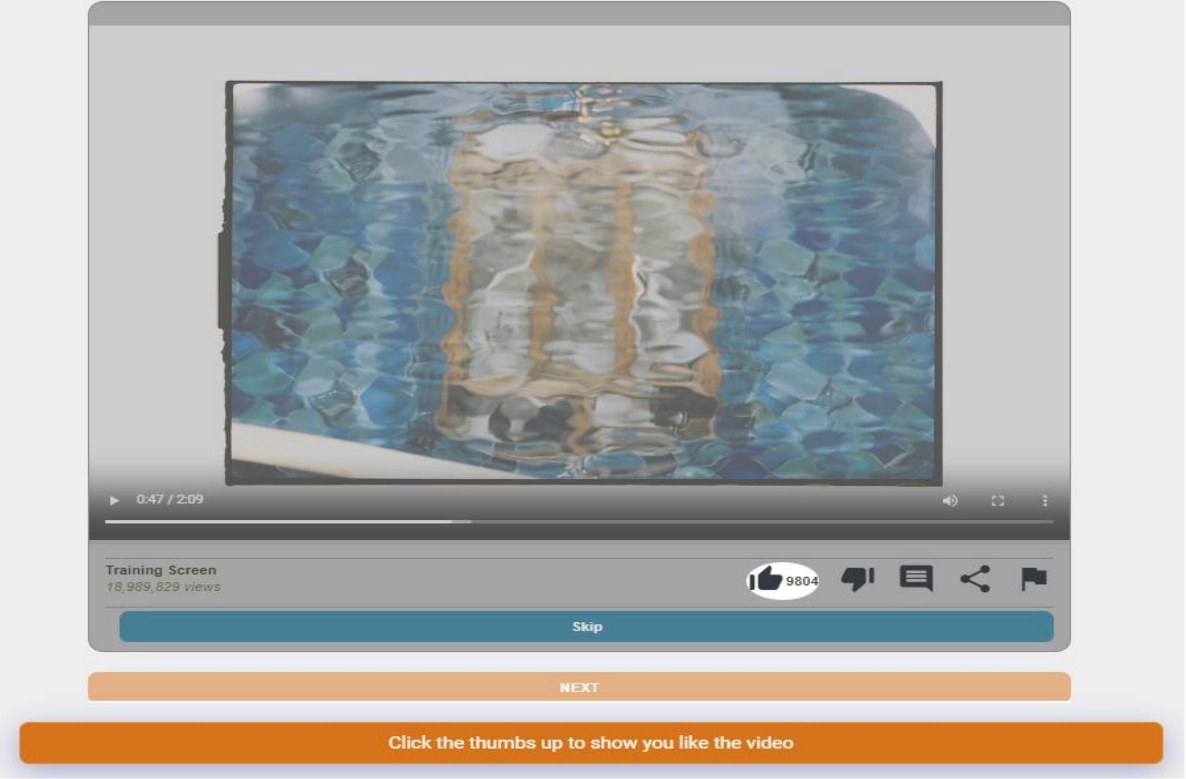
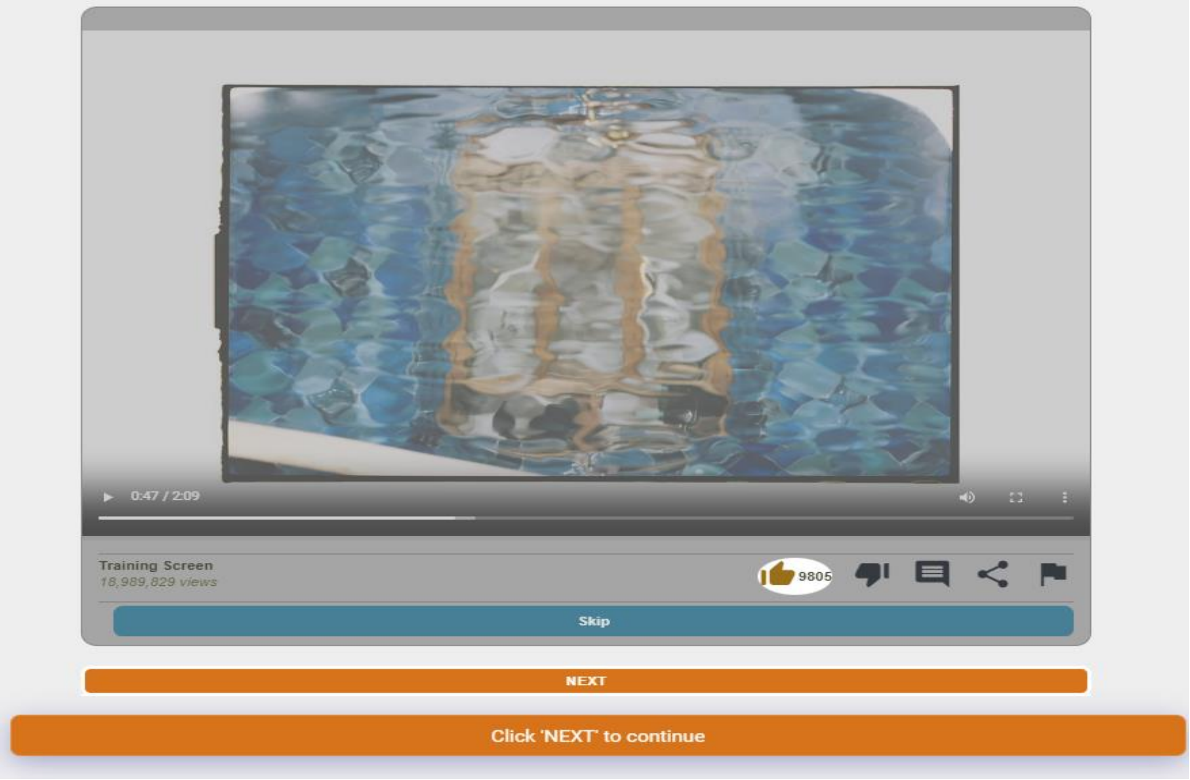

Please watch the video which introduces the functions of the platform.
This video contains sound.

Figure 7. Arm 3 – Video microtutorial: Participants were shown a video that highlighted interface features for participants, prior to starting the main experiment.



Click the thumbs up to show you like the video

Figure 8. Arm 4 – Interactive microtutorial: Interface that participants were shown prior to starting the main experiment. Participants received text-based prompts indicating that they had to complete an action.



Click 'NEXT' to continue

Figure 9. Arm 4 – Interactive microtutorial: Once participants performed the action, the interface updated to reflect the completion of the action, and the prompt indicated they could progress to the next stage, or action, by clicking “NEXT” to continue.

5. Outcomes

5.1. Primary outcome

In this study, we measured whether a participant had submitted a completed report or not, for each potentially harmful video (i.e., for 3 videos for each participant). For a report to be completed, the participant had to click on the reporting flag, select a reporting category and submit the report. It was not necessary for the participant to include a comment in the report. This variable constituted the primary outcome variable (1.1 in Table 1). Note that two outcome measures were revised following the data collection. Outcome 1.1 was revised because we could not have analysed it in the way we specified it in the trial protocol (see section 6.1.2 for an explanation). Outcome 2.5. was revised to better estimate whether participants did not complete a report.

5.2. Secondary outcomes

Secondary outcomes are listed in Table 2.

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

	Outcome measure
Primary	1.1. (Behavioural) A binary variable indicating whether a user completed a report of each potentially harmful video (binary coded as 1 if a user completed a report of a harmful video or 0 if they did not) Revised to: A count variable of 0 to 3 reports of potentially harmful videos, per participant
Secondary	2.1. (Behavioural) A variable with the sum of all interactions [likes (1) + or dislike (1) + click on share button (1) + click on comment button (1) + click on reporting flag (1)] for each participant, per each video 2.2. (Behavioural) A variable with aggregated count per participant of the number of completed reports 2.3. (Behavioural) A binary variable indicating whether a user completed a report of a potentially harmful video 2.4. (Behavioural) A binary variable in the dataset that is coded as 1 if a user pressed the flag on a potentially harmful video and 0 if they did not 2.5. (Behavioural) A binary variable in the dataset that is coded as 1 if a user selected a category to report a potentially harmful video and 0 if they did not Revised to: A binary variable in the dataset that is coded as 1 if a user pressed the flag on a potentially harmful video but not submitted the report and 0 if they did not press the flag button 2.6. (Behavioural) A binary variable indicating whether a user skipped a potentially harmful video, or a neutral video. Recorded using binary variable [1 if user skipped (by clicking skip button) or 0 if they did not] 2.7. (Behavioural) A binary variable indicating whether a user has reported harmful content into the correct category, coded as 1 if the user reported into the correct category, and 0 for all other categories 2.8. (Behavioural) A binary variable indicating whether a user has disliked each potentially harmful video (1) or not (0) 2.9. (Behavioural) A binary variable indicating whether a user completed a report on the task following the 6 randomised videos where users are prompted to complete a report 2.10. (Behavioural) A binary variable indicating whether a user submitted a report using the correct category on the task following the 6 randomised videos where users are prompted to complete a report 2.11. (Behavioural) A binary variable indicating whether a user completed a report of a neutral video 2.12. (Attitudinal) Responses to reporting prompt question in the prompted reporting task if a user skipped the video without reporting 2.13. (Attitudinal) Responses to attitudinal survey questions

Descriptive metrics	A binary variable for each video for each participant indicating whether a participant completed that action (1) or not (0) for all possible forms of engagement An aggregated count for each feature (e.g., like, share) for each experimental arm The length of time from entering the interface to pressing any button, recorded to two decimal places The length of time from starting a report (pressing the report button) to submitting a report (pressing the submit button) (to two decimal places) The order in which participants were shown each video, irrespective of if the video was played or skipped The length of time participants viewed potentially harmful and neutral videos, per video (to two decimal places) A binary variable indicating whether a participant was shown 3 potentially harmful videos in a row or not (this is expected to be a relatively rare occurrence) The length of time a user spent in the intervention stage (completing the microtutorial) Comments entered by participants as comments on individual videos Comments entered by participants as additional details via reports The number of reports which were made on videos that were categorised as potentially harmful A binary variable which indicates whether a participant ‘opted out’ of the experiment, in order to determine the relative opt-out rate of participants across the entire experiment Number of participants who decided not to continue at the introduction stage Number of participants who dropped out during the study Number of participants who failed the attention check

6. Statistical methods and analysis

6.1. Statistical methods

6.1.1. Primary analysis: Intended approach

The primary outcome was anticipated to be whether a participant has decided to report or not report each potentially harmful video (see 1.1 in Table 2).

Given that this outcome is binary (report vs. not report), we proposed to use a logistic mixed-effects model to examine the differences between the different intervention arms. We favoured this approach over aggregating data and running linear models, or Poisson regression models, because treating each reporting decision as a binary event makes full use of the data. One of the key advantages of treating the reporting behaviour as a binary decision for each video and running mixed-effects logistic regression models, is that these models consider additional uncertainty due to the effect of variation between individuals and due to the effect of variation in the probability of reporting behaviour between different potentially harmful videos. In contrast, aggregating data and running analyses using a linear model, results in a loss of information as it ignores individual variation. Ignoring variation in the probability of reporting between people and between potentially harmful videos could give more precise estimates of the effects, but these estimates could be misleading. This is because the effects would be less likely to reflect the reality (in terms of the magnitude) and the probability of detecting spurious effects (or errors of direction of the effect) would be higher (higher Type I error rate) compared to the approach we proposed to use. An additional motivation for wanting to use mixed-effects models was that we expected reporting behaviour to be different for different videos. A reasonable assumption to make was that some potentially harmful videos would be more likely to be reported by some people than others (on average). Using models with an aggregated outcome measure of the number of reports makes the assumption that every potentially harmful video has equal probability of being reported. Instead, by using a binary indicator of whether a report has been completed per video, and including a random intercept for each video in a mixed-effects model, we would be making a more realistic assumption that some of potentially harmful videos may have a higher probability of reporting than others. Under our preferred approach, the 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 binary variable indicating whether a participant i watching potentially harmful video j completed a report or not. The binary variable is 1 if the video is successfully reported, but 0 if the video is not successfully reported.

β_0 is the predicted value for a baseline category - here arm 1 - whereas β_1, β_2, and β_3, represent deviations in the log-odds of reporting potentially harmful videos of arms 2, 3, and 4, respectively, from arm 1.

u_{1i} is the random intercept of participant i, $u_{1i} \sim N(0, \sigma_1)$ for $i \in \{1, 2, \dots, N\}$ 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 $j \in \{1, 2, 3\}$.

6.1.2. Primary analysis: Revised approach

By comparing the expected distribution of zeros (derived by running 1,000 simulations of scaled residuals using a fitted mixed-effects logistic regression model) against the observed values (from the collected data), we found more zeros in the data than expected (zero-inflation). Thus, we were not able to treat reporting as a binary event. Consequently, the primary outcome data was analysed in the same way as the primary outcome data in the Reporting Online experiment:[15] using a zero-inflated Poisson regression model, treating the primary outcome as a count variable (0 to 3 reports per participant). 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). The interpretation of the estimates produced by the zero-inflation component of such a model may seem counterintuitive, because a negative estimate corresponds to a positive effect of increasing the probability of reporting.

Critically, in the context of this outcome, zero-inflated Poisson regression models make more sense than some other models that are used to dealing with zero-inflation, such as hurdle models. This is because a hurdle model is more constrained than a zero-inflated Poisson model, in that it treats zeros in the data as being structural.[16] Structural zeros in the context of this data refer to zeros from participants who always produce zero (i.e., choose not to report videos) in the simulation and in real-life. In contrast, the zero-inflated Poisson model treats zeros in the data as arising from some

[15] https://www.ofcom.org.uk/__data/assets/pdf_file/0020/241832/Online-Trials-Appendix-2-Reporting-Mechanisms.pdf

[16] The hurdle model is constrained in how it deals with zeros, because it a mixture of a truncated Poisson distribution (modelling counts of 1, 2, 3 in the context of this study) and a binary distribution, where it considers zeros only in the binary component.

participants who never report potentially harmful videos (structural zeros), but also from participants who may report a potentially harmful video in real-life but choose not to report any of the ones in the experiment (referred to as sampling zeros).[17] Evidently, a more realistic assumption to make was that there are structural and sampling zeros in the data, rather than just structural.

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 the estimates of the zero-inflated Poisson model that was fit to the primary outcome data. 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 primary model predicted the observed counts relatively well (because the red line, and dots, fit the pattern visualised by the histogram).

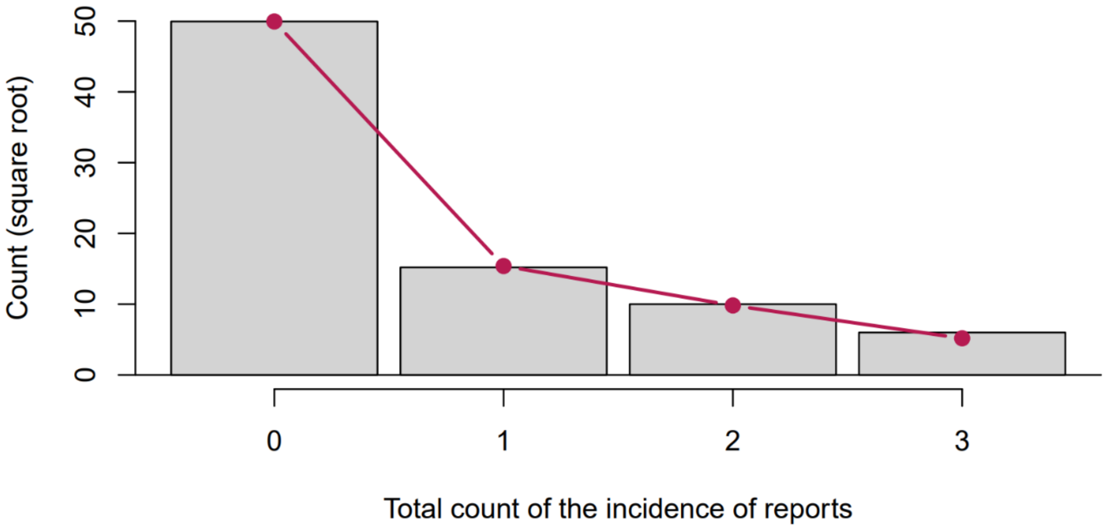


Figure 10. The distribution of counts (square rooted) by the total count of the incidence of reports.

To answer our research questions (see section 1.3), comparisons between arms were performed. When running these multiple comparisons, the Bonferroni correction was utilised to maintain the family-wise error rate.

6.1.3. Secondary analysis

Secondary outcomes 2.1 to 2.11 (in Table 1) were analysed using one of the following methods: mixed-effects logistic regression models, logistic regression models, or zero-inflated Poisson regression models. Secondary outcomes 2.12 and 2.13 (specifically those using responses to Likert scales) were analysed using ordinal regression models.

6.1.4. Descriptive statistics

We report descriptive statistics in this report using figures and tables.

6.2. Statistical power

6.2.1. Overview

We did not run power simulations for this project, due to time constraints. We assumed that a sample size of 2,800 participants (700 per arm) would be sufficient to detect a 10% difference in the probability of reporting potentially harmful videos between arms in this study. This assumption was based on the results of power simulations that we ran for the Defaults Online Trial (where the outcome was the probability of skipping potentially harmful videos). Note that the accuracy of the results of these simulations can be questioned, given that a different outcome measure was used in this study. However, there is a margin for error given that the Defaults Online Trial simulations showed that recruiting 2,800 participants resulted in power that is 8% higher than the conventional threshold, used in power simulations, of 80% at $\propto = 0.05$. The simulations that informed the sample size selection in the Defaults Online Trial are specified below.

6.2.2. Assumptions underlying the power simulations

To run power simulations for logistic mixed-effects models, assumptions about the variance and standard deviation parameters of the random effects are required. To obtain meaningful estimates of power using power simulations, these assumptions should be grounded in prior research. Thus, the estimates of the parameters of variation in the probability of skipping between participants (random intercept for participants, σ_1) and in the probability of skipping between potentially harmful videos

[17] This is because the zero-inflated Poisson model is a mixture of a Poisson (modelling counts of 0, 1, 2, 3 in the context of this study, therefore considering sampling zeros) and binary (considering structural zeros).

(random intercept for potentially harmful videos, σ_2) were based on the estimates found in the Skipping Online Experiment. Effect sizes of 8%, 9% and 10% in the probability of skipping - per intervention arm compared to the control arm - were assumed under different scenarios. As with the previous behavioural online experiments for Ofcom, the minimum detectable effect size of interest was set at 10%.

Table 3 shows the estimates of power under different model assumptions, given 1,000 simulations per scenario. It is unlikely that the estimates under any scenario will be the same as the ones obtained from models that use the collected data. Thus, the estimates of power to detect the effect of the intervention, given the scenarios considered, are not going to be an exact representation of the true effect of the interventions (in the context of our study). However, they are likely to be reasonably close, since we are using the same sampling strategy and potentially harmful videos as we did in the Skipping Online Experiment. In the case of a null effect, the estimates below may 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).

Following the results of the power simulations, a decision was taken to recruit a sample of 2,800 participants, resulting in 700 participants per treatment arm. This was informed by Scenario 12 in Table 3. Scenario 12 shows that, given our assumptions, 2,800 participants were needed to ensure that the trial was sufficiently powered to detect a minimum effect size of 10%. Specifically, under Scenario 12, the power to reject the null hypothesis of there being no difference between arms was 88% (which is higher the conventional threshold, used in power simulations, of 80% at $\propto = 0.05$).

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[18]	Power
1	1600	8%	1.94	0.46	0.1	40%
2	1600	9%	1.94	0.46	0.1	52%
3	1600	10%	1.94	0.46	0.1	60%
4	2000	8%	1.94	0.46	0.1	49%
5	2000	9%	1.94	0.46	0.1	63%
6	2000	10%	1.94	0.46	0.1	73%
7	2400	8%	1.94	0.46	0.1	55%
8	2400	9%	1.94	0.46	0.1	71%
9	2400	10%	1.94	0.46	0.1	79%
10	2800	8%	1.94	0.46	0.1	66%
11	2800	9%	1.94	0.46	0.1	78%
12	2800	10%	1.94	0.46	0.1	88%

[18] The parameter was added in power simulations to be conservative (which is useful in the context where we can expect the interface not to work equally well for every participant), but it is not something that can be estimated by the model. Observation level variability parameter is not required in logistic models - the probability parameter captures this from observational variability in the draw from the binomial distribution (the probability parameter captures both location of the expectation of the mean and how much variance there will be). Thus, σ_3 is adding noise that cannot be modelled by the logistic model. It reflects variance that is above and beyond observation level noise than would be expected by the logistic model.

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 42 (Table 4).

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	42	47.6	55.7
Arm 2 – Static	41	48.5	52
Arm 3 – Video	41	49.4	58.7
Arm 4 – Interactive	40	50.3	57.4

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

The distribution of device operating system used to complete the experiment was relatively balanced across each treatment arm (Table 5). Please refer to Appendix B, in Section 10, for the demographic breakdown of participants by device type.

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

Device operating system	Arm 1 Control (%)	Arm 2 Static (%)	Arm 3 Video (%)	Arm 4 Interactive (%)
Android	41.0%	38.2%	37.0%	38.0%
iOS	26.4%	30.2%	24.4%	26.7%
Windows	24.2%	22.1%	28.2%	26.5%
macOS	4.9%	6.0%	6.7%	5.2%
iPadOS	1.0%	1.3%	2.1%	0.8%
ChromeOS	1.8%	1.5%	1.1%	2.2%
Linux	0.7%	0.7%	0.4%	0.6%
Unknown	0%	0%	0%	0%

7.2. Primary Outcome Analysis

7.2.1. Headline results

All microtutorials increased the likelihood of reporting potentially harmful videos, compared to not having a microtutorial. The interactive microtutorial was found to be most effective. It increased the likelihood of reporting potentially harmful videos more than any other microtutorial. These findings were robust to sensitivity analyses that adjusted for devices on which the experiment was completed and whether the potentially harmful videos were played by participants (section 7.2.3).

7.2.2. Reporting of potentially harmful videos

Using all observations, we found that all the interventions had a significant effect on the number of completed reports when watching potentially harmful content versus the control arm (Table 6). Figure 11 shows that 96% of participants in the control did not report any potentially harmful videos, 91% in Arm 2 – Static, 84% in Arm 3 – Video, and 77% in Arm 4 – Interactive.[19] Specifically, estimates derived from odds ratios reported in Table 6 show that completing the static microtutorial decreased the odds of not reporting potentially harmful content by 70% ($(0.30 - 1) \times 100 = -70$) compared to participants in the control condition.[20] Similarly, participants who completed the video and interactive microtutorial had odds 79% and 89% lower of not reporting potentially harmful content, respectively.

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 reporting was not significantly different for participants who were shown three potentially harmful videos in a row compared to those who did not.

[19] These estimates were calculated using observed values (from the collected data). These are reported as they are typically easier to understand, however they do not directly relate to model estimates and cannot be calculated using estimated, reported, Odds Ratios.

[20] 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.

Table 6. Model-based estimates of not reporting potentially harmful videos (zero-inflated Poisson model)

	Odds Ratios	95% CI	*z*-value	*P*
Intercept	13.54	7.64 – 24.00	8.925	< 0.001
Arm 2 – Static	0.30	0.14 – 0.62	-3.201	0.001
Arm 3 – Video	0.21	0.11 – 0.41	-4.707	< 0.001
Arm 4 – Interactive	0.11	0.06 – 0.21	-6.637	< 0.001

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

Adjusting for multiple comparisons, the odds of not reporting potentially harmful videos remained significantly lower for participants in all three treatment arms compared to the control arm (Table 7). Additionally, participants who completed the interactive microtutorial (Arm 4) had significantly lower odds of not reporting potentially harmful videos relative to participants who completed the static microtutorial (64%) and participants who completed the video microtutorial (50%). Thus, interactive microtutorial was most effective at increasing the probability of reporting potentially harmful videos.

Table 7. 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*
Arm 1 – Control vs Arm 2 – Static	0.30	0.11 – 0.78	-3.201	0.008
Arm 1 – Control vs Arm 3 – Video	0.21	0.09 – 0.49	-4.707	< 0.001
Arm 1 – Control vs Arm 4 – Interactive	0.11	0.05 – 0.25	-6.637	< 0.001
Arm 2 – Static vs Arm 3 – Video	0.72	0.35 – 1.49	-1.156	1
Arm 2 – Static vs Arm 4 – Interactive	0.36	0.17 – 0.77	-3.456	0.003
Arm 3 – Video vs Arm 4 – Interactive	0.50	0.29 – 0.89	-3.079	0.012

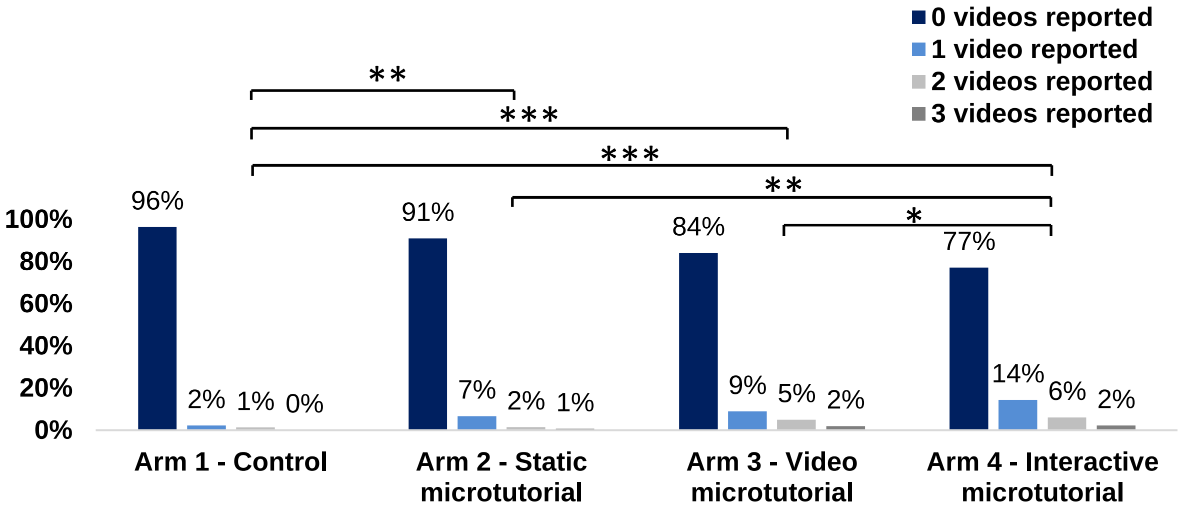


Figure 11. Percentage of the number of reports completed by participants, by arm (multiple comparisons adjusted; * p < 0.05; ** p < 0.01; *** p < 0.001; not sensitive to controlling for device type; modelled using a zero-inflated Poisson model). Note that the percentages for Arm 2 - Static microtutorial add up to 101% due to rounding.

7.2.3. Sensitivity analyses

Whilst user testing the VSP interface, it became apparent that for a minority of participants in certain circumstances, videos would not begin playing automatically.[21] This issue did not affect the microtutorial stage of the trial but may have impacted user experience of the video interface. Thus, we

[21] The issue mainly affected participants who accessed the trial on a device running iOS. However, other Apple devices, running macOS and iPadOS, were also affected. We believe that this is a design choice by Apple, and that this cannot be overridden in this online environment. However, 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.

conducted sensitivity analyses to determine whether the reported effects were affected by this issue. To identify potential instances in which videos did not correctly auto-play, observations in which a video was skipped with a watch time of 0 seconds were identified. The primary analysis was then repeated on two sensitivity datasets; a dataset which included only observations from participants who recorded a non-zero watch time on all videos (n = 2,330)[22] and a dataset which included only observations from participants who recorded a non-zero watch time on at least one video (n = 2,842).[23]

After restricting the dataset to only include observations from participants who recorded a non-zero watch time on at least one video (n = 2,842), 96% of participants in the control did not report any potentially harmful videos, 91% in Arm 2 – Static, 84% in Arm 3 – Video, and 77% in Arm 4 – Interactive (Figure 12).[24] In a replication of the primary analysis, all the interventions were found to have a significant effect on the number of completed reports when watching potentially harmful content relative to the control arm (Table 8). Specifically, estimates derived from odds ratios reported in Table 8 indicate that completing the static microtutorial decreased the odds of not reporting potentially harmful content by 71% compared to participants in the control condition. Similarly, participants who completed the video and interactive microtutorial had odds 79% and 89% lower of not reporting potentially harmful content, respectively.

Table 8. Model-based estimates of not reporting potentially harmful videos (zero-inflated Poisson model)

	Odds Ratios	95% CI	*z*-value	*P*
Intercept	13.52	7.63 – 23.97	8.919	< 0.001
Arm 2 – Static	0.29	0.14 – 0.62	-3.226	0.001
Arm 3 – Video	0.21	0.11 – 0.40	-4.724	< 0.001
Arm 4 – Interactive	0.11	0.05 – 0.20	-6.677	< 0.001

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

In a replication of the primary analysis, after adjusting for multiple comparisons, the odds of not reporting potentially harmful videos remained significantly lower for participants in all three treatment arms (See Table 9). Additionally, participants who completed the interactive microtutorial (Arm 4) had significantly lower odds of not reporting potentially harmful content relative to participants who completed the static microtutorial (64%) and participants who completed the video microtutorial (50%).

Table 9. Estimates of the odds of not reporting 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 – Static	0.29	0.11 – 0.77	-3.226	0.008
Arm 1 – Control vs Arm 3 – Video	0.21	0.09 – 0.49	-4.724	< 0.001
Arm 1 – Control vs Arm 4 – Interactive	0.11	0.04 – 0.25	-6.677	< 0.001
Arm 2 – Static vs Arm 3 – Video	0.72	0.35 – 1.50	-1.137	0.647
Arm 2 – Static vs Arm 4 – Interactive	0.36	0.17 – 0.76	-3.466	0.003
Arm 3 – Video vs Arm 4 – Interactive	0.50	0.28 – 0.88	-3.116	0.011

Figure 12 shows the percentage of participants reporting the number of potentially harmful videos after restricting the dataset to include participants who recorded a non-zero watch time on at least one video. Significant differences, as estimated by the sensitivity model, are shown using horizontal lines.

[22] Devices used by excluded participants: Android = 140, Windows = 65, iOS = 254, iPadOS = 10, MacOS = 55, ChromeOS = 4, Linux = 4.
[23] Devices used by excluded participants: Android = 1, Windows = 3, iOS = 9, iPadOS = 2, MacOS = 5.
[24] Mean observed probabilities were calculated using observed values (from the collected data). These are reported as they are typically easier to understand, however they do not directly relate to model estimates and cannot be calculated using estimated, reported, Odds Ratios.

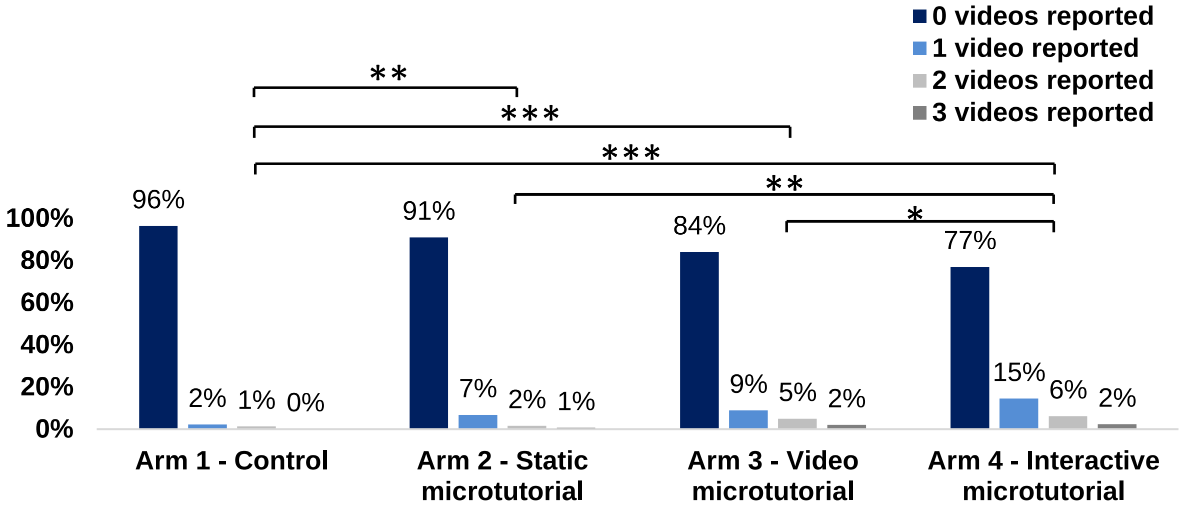


Figure 12. Percentage of the number of reports completed by participants, by arm (multiple comparisons adjusted; * p < 0.05; ** p < 0.01; *** p < 0.001; not sensitive to controlling for device type; modelled using a zero-inflated Poisson model)

After restricting the dataset to only include observations from participants who recorded a non-zero watch time on all videos (n = 2,330), the mean observed probability of reporting potentially harmful videos was 2% in the control arm, 5% in Arm 2 – Static, 9% in Arm 3 – Video, and 13% in Arm 4 – Interactive.[25] In a second replication of the primary analysis, all the interventions were found to have a significant effect on the number of completed reports when watching potentially harmful content relative to the control arm (Table 10). Specifically, estimates derived from odds ratios reported in Table 8 indicate that completing the static microtutorial decreased the odds of not reporting potentially harmful content by 74% compared to participants in the control condition. Similarly, participants who completed the video and interactive microtutorial had odds 82% and 91% lower of not reporting potentially harmful content, respectively.

Table 10. Model-based estimates of not reporting potentially harmful videos (zero-inflated Poisson model)

	Odds Ratios	95% CI	*z*-value	*P*
Intercept	14.74	8.29 – 26.19	9.170	< 0.001
Arm 2 – Static	0.26	0.12 – 0.55	-3.541	< 0.001
Arm 3 – Video	0.18	0.09 – 0.34	-5.228	< 0.001
Arm 4 – Interactive	0.09	0.04 – 0.17	-7.093	< 0.001

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

In another replication of the primary analysis, adjusting for multiple comparisons, the odds of not reporting potentially harmful videos remained significantly lower for participants in all three treatment arms (Table 11). Additionally, participants who completed the interactive microtutorial (Arm 4) had significantly lower odds of not reporting potentially harmful content relative to participants who completed the static microtutorial (66%) and participants who completed the video microtutorial (50%).

Table 11. 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*
Arm 1 – Control vs Arm 2 – Static	0.26	0.10 – 0.69	-3.541	0.002
Arm 1 – Control vs Arm 3 – Video	0.18	0.08 – 0.41	-5.228	< 0.001
Arm 1 – Control vs Arm 4 – Interactive	0.09	0.04 – 0.21	-7.093	< 0.001
Arm 2 – Static vs Arm 3 – Video	0.68	0.32 – 1.42	-1.338	1

[25] Mean observed probabilities were calculated using observed values (from the collected data). These are reported as they are typically easier to understand, however they do not directly relate to model estimates and cannot be calculated using estimated, reported, Odds Ratios.

Arm 2 – Static vs Arm 4 – Interactive	0.34	0.16 – 0.73	-3.587	0.002
Arm 3 – Video vs Arm 4 – Interactive	0.50	0.27 – 0.91	-2.933	0.020

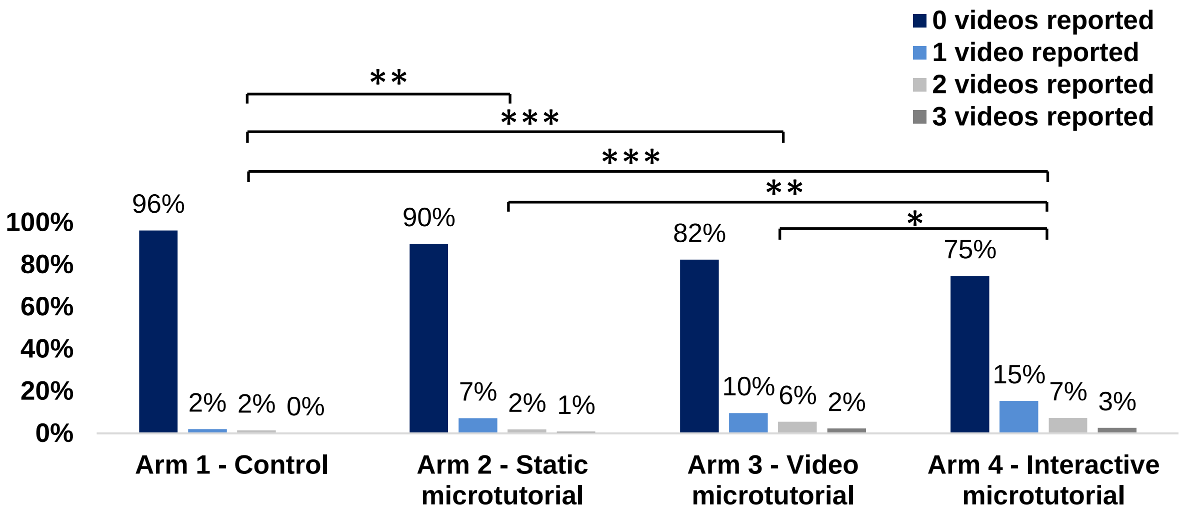


Figure 13. Percentage of the number of reports completed by participants, by arm (multiple comparisons adjusted; * p < 0.05; ** p < 0.01; *** p < 0.001; not sensitive to controlling for device type; modelled using a zero-inflated Poisson model)

7.3. Secondary Outcome Analyses

7.3.3. Reporting of neutral videos

No inference on over-reporting between arms can be made as it is not known whether the reported pattern would replicate due to the low count of reports of neutral videos. Specifically, 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 reporting neutral content was very low (Arm 1 – Control = 0; Arm 2 – Static = 1; Arm 3 – Video = 6; Arm 4 – Interactive = 10).

Consequently, although the number of submitted reports of neutral content appears higher in Arm 4 relative to other arms, 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.

7.3.4. Accurate categorisation during reporting of potentially harmful content

There were no significant differences in the accuracy of the categories that users selected for the potentially harmful content across treatment arms (Table 12). A completed report was classified as accurate if the category the respondent selected during the reporting process matched a pre-specified content label. In this experiment, the first potentially harmful video (Video 4) was labelled as ‘Misinformation’, the second potentially harmful video (Video 5) was labelled as ‘Harassment’, and the third potentially harmful video (Video 6) was labelled as ‘Hate speech’.

We analysed the accuracy of reporting categories using a mixed-effects logistic regression model. Prior to running the model, we excluded observations of videos that were not reported. The mean observed probability of the reporting category being accurate, given that a participant had completed a report of a potentially harmful videos was 54% in the control arm, 56% in Arm 2 – Static, 56% in Arm 3 – Video, and 54% in Arm 4 – Interactive. Figure 12 shows the percentage of participants correctly categorising potentially harmful videos during a report.

Table 12. Model-based estimates of accuracy of categorisation of potentially harmful content

	Odds Ratio	95% CI	*z*-value	*P*
Intercept	1.32	0.29 – 6.04	0.358	0.720
Arm 2 – Static	1.56	0.62 – 3.91	0.954	0.340
Arm 3 – Video	1.30	0.56 – 3.02	0.612	0.541
Arm 4 – Interactive	1.21	0.53 – 2.76	0.455	0.649

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

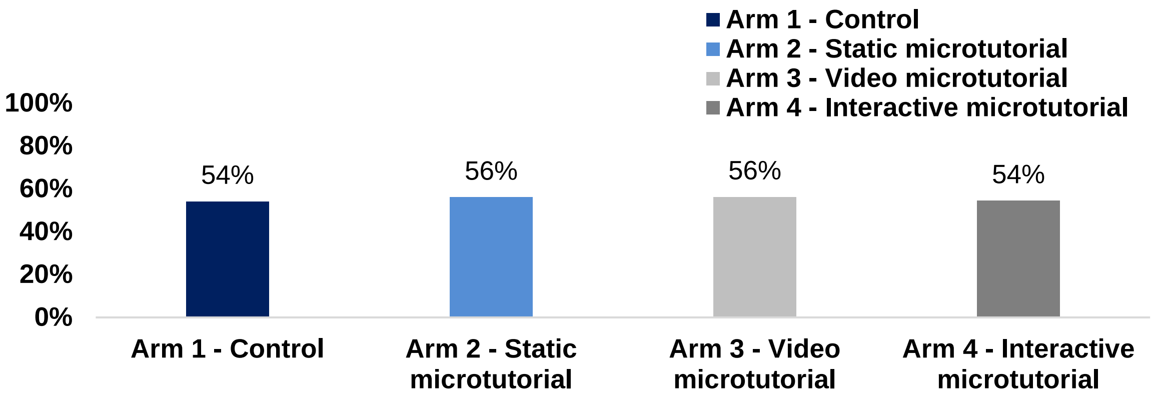


Figure 14. Proportion of completed reports that accurately classified the content of potentially harmful videos, by arm (modelled using a zero-inflated Poisson model)

7.3.5. Pressing the flag button while viewing potentially harmful videos

We found that participants shown microtutorials were significantly more likely to press the reporting flag when watching potentially harmful videos compared to those not shown microtutorials (Table 13). Note that pressing the reporting flag is not the same as completing a report (primary outcome measure). In Arm 1 – Control, 51 participants pressed the flag button, whereas 127, 248, and 290, pressed it in Arm 2 – Static, Arm 3 – Video, and Arm 4 – Interactive, respectively.

We reran the primary outcome model on the count of pressed reporting flags of potentially harmful videos, by participant. Estimates derived from odds ratios reported in Table 13 show that completing the static microtutorial decreased the odds of not pressing the flag button when watching potentially harmful content by 63% compared to participants in the control condition. Similarly, participants who completed the video and interactive microtutorial had odds 77% and 85% lower of not pressing the flag button when watching potentially harmful content, respectively. Thus, we replicated the findings of the primary outcome model.

Table 13. Model-based estimates of not pressing the flag button while watching potentially harmful videos (zero-inflated Poisson model)

	Odds Ratios	95% CI	*z*-value	*P*
Intercept	11.26	7.03 – 18.01	10.093	< 0.001
Arm 2 – Static	0.37	0.21 – 0.66	-3.387	< 0.001
Arm 3 – Video	0.23	0.13 – 0.38	-5.533	< 0.001
Arm 4 – Interactive	0.15	0.09 – 0.25	-7.036	< 0.001

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

Adjusting for multiple comparisons, the odds of not pressing the flag button when watching potentially harmful videos remained significantly lower for participants in all three treatment arms compared to the control arm (Table 14). Additionally, participants who completed the interactive microtutorial (Arm 4) had significantly lower odds of not pressing the flag button while watching potentially harmful videos relative to participants who completed the static microtutorial (64%) and participants who completed the video microtutorial (50%). Thus, interactive microtutorial was most effective at increasing the probability of pressing the flag button while watching potentially harmful videos.

Table 14. Estimates of the odds of not pressing the flag button while watching 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 – Static	0.37	0.18 – 0.78	-3.387	0.004
Arm 1 – Control vs Arm 3 – Video	0.23	0.11 – 0.45	-5.533	< 0.001
Arm 1 – Control vs Arm 4 – Interactive	0.15	0.07 – 0.29	-7.036	< 0.001
Arm 2 – Static vs Arm 3 – Video	0.61	0.36 – 1.03	-2.420	0.093
Arm 2 – Static vs Arm 4 – Interactive	0.39	0.23 – 0.68	-4.402	< 0.001
Arm 3 – Video vs	0.65	0.41 – 1.03	-2.413	0.095

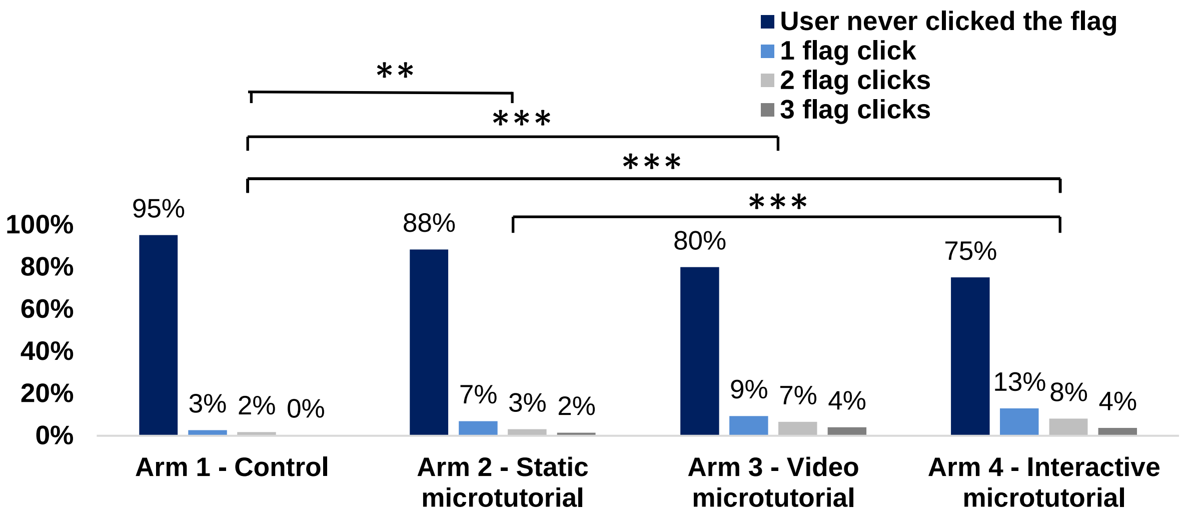


Figure 15. Proportion of times participants pressed the flag button whilst watching potentially harmful videos, by arm (multiple comparisons adjusted; *** p < 0.001; modelled using a zero-inflated Poisson model)

7.3.6. Incomplete reports of potentially harmful videos

We found no significant differences in the number of incomplete reports between arms. In Arm 1 – Control, 14 participants started but did not submit a report, whereas 42, 70, and 51, started but did not complete a report in Arm 2 – Static, Arm 3 – Video, and Arm 4 – Interactive, respectively. We reran the primary outcome model on the count of incomplete reports of potentially harmful videos, by participant. The differences between arms are not significant, because the level of uncertainty associated with such low counts is high (Table 15).

Table 15. Model-based estimates of the odds of not submitting a report of potentially harmful videos, after pressing the flag button

	Odds Ratio	95% CI	*z*-value	*P*
Intercept	6.66	0.77 – 57.51	1.725	0.085
Arm 2 – Static	0.97	0.10 – 9.43	-0.026	0.979
Arm 3 – Video	0.50	0.05 – 4.77	-0.598	0.550
Arm 4 – Interactive	0.29	0.02 – 3.53	-0.968	0.333

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

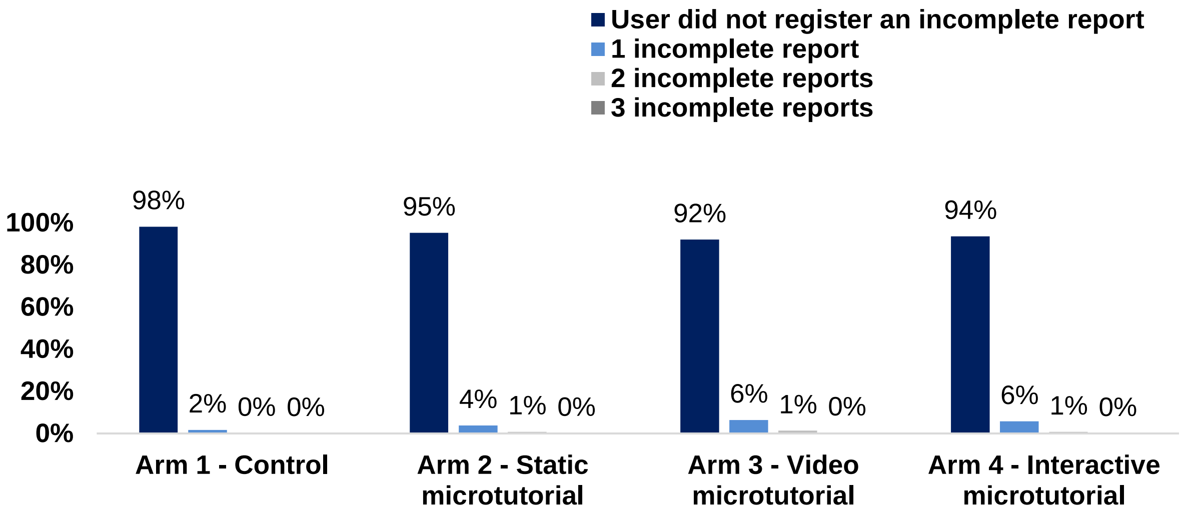


Figure 16. Proportion of times participants registered incomplete reports of potentially harmful videos, by arm (modelled using a zero-inflated Poisson model)

7.3.7. Interacting with potentially harmful videos

Participants who were shown the video microtutorial "interacted"[26] more frequently with potentially harmful videos than participants who were shown the static microtutorial or not shown any microtutorial. Those shown the interactive microtutorial had significantly more interactions with potentially harmful videos than those not shown any microtutorial. In addition, all microtutorials increased the likelihood of participants interacting with potentially harmful videos compared to the control, video and interactive microtutorials significantly more than static. In Arm 1 – Control, there were 643 interactions with potentially harmful videos, whereas 1,032, 1,445, and 1,401, interactions it in Arm 2 – Static, Arm 3 – Video, and Arm 4 – Interactive, respectively.

We ran a zero-inflated mixed-effects truncated Poisson regression model on the count of interactions at the video level. Table 16 shows the count and zero-inflated components of the model. The count component of the model shows that for participants shown video or interactive microtutorials the expected number of interactions with potentially harmful videos was higher by 80% and 66% respectively, compared to those not shown a microtutorial. The zero-inflated component of the model shows that completing the static microtutorial decreased the odds of not interacting with potentially harmful content by 46% compared to participants in the control condition. Similarly, participants who completed the video and interactive microtutorial had odds 66% and 65% lower of not interacting with potentially harmful content, respectively. (Figure 17 shows the proportion of times participants interacted with potentially harmful videos, by arm, using estimates from the zero-inflated component of the model.)

Table 16. Model-based estimates of interacting with potentially harmful videos

Count component

	Risk Ratio	95% CI	*z*-value	*P*
Intercept	0.21	0.16 – 0.27	-11.905	< 0.001
Arm 2 – Static	1.25	0.93 – 1.68	1.452	0.146
Arm 3 – Video	1.80	1.37 – 2.38	4.154	< 0.001
Arm 4 – Interactive	1.66	1.25 – 2.19	3.526	< 0.001

Zero-inflation component

	Odds Ratio	95% CI	*z*-value	*P*
Intercept	2.93	2.66 – 3.23	21.68	< 0.001
Arm 2 – Static	0.54	0.47 – 0.61	-9.33	< 0.001
Arm 3 – Video	0.34	0.30 – 0.39	-16.32	< 0.001
Arm 4 – Interactive	0.35	0.31 – 0.40	-15.98	< 0.001

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

The differences reported in Table 16 were not sensitive to adjusting for multiple comparisons (Table 17). One additional difference was found in the count component of the model: for participants shown the video microtutorial, the expected number of interactions with potentially harmful videos was 44% higher compared to those shown the static microtutorial. Two additional differences were found in the zero-inflation component. Participants who completed the video and interactive microtutorial had odds 36% and 35% lower of not interacting with potentially harmful content respectively, compared to those shown the static microtutorial.

Table 17. Estimates of interacting with potentially harmful videos (p values and CIs corrected for multiple comparisons using the Bonferroni correction)

Count component

Comparison	Risk Ratio	95% CI	*z*-value	*P*
Arm 1 – Control vs Arm 2 – Static	1.25	0.83 – 1.87	1.452	0.879
Arm 1 – Control vs Arm 3 – Video	1.80	1.24 – 2.62	4.154	< 0.001
Arm 1 – Control vs Arm 4 – Interactive	1.66	1.14 – 2.42	3.526	0.003

[26] An interaction was any of the following: Like, Dislike, Comment, Share, Flag click. Like and Dislike were mutually exclusive options.

Arm 2 – Static vs Arm 3 – Video	1.44	1.06 – 1.97	3.146	0.010
Arm 2 – Static vs Arm 4 – Interactive	1.33	0.97 – 1.81	2.390	0.101
Arm 3 – Video vs Arm 4 – Interactive	0.92	0.70 – 1.21	-0.817	1

Zero-inflation component

Comparison	Odds Ratio	95% CI	*z*-value	*P*
Arm 1 – Control vs Arm 2 – Static	0.54	0.45 – 0.64	-9.33	< 0.001
Arm 1 – Control vs Arm 3 – Video	0.34	0.29 – 0.41	-16.32	< 0.001
Arm 1 – Control vs Arm 4 – Interactive	0.35	0.29 – 0.42	-15.98	< 0.001
Arm 2 – Static vs Arm 3 – Video	0.64	0.54 – 0.75	-7.315	< 0.001
Arm 2 – Static vs Arm 4 – Interactive	0.65	0.55 – 0.77	-6.953	< 0.001
Arm 3 – Video vs Arm 4 – Interactive	1.02	0.87 – 1.20	0.366	1

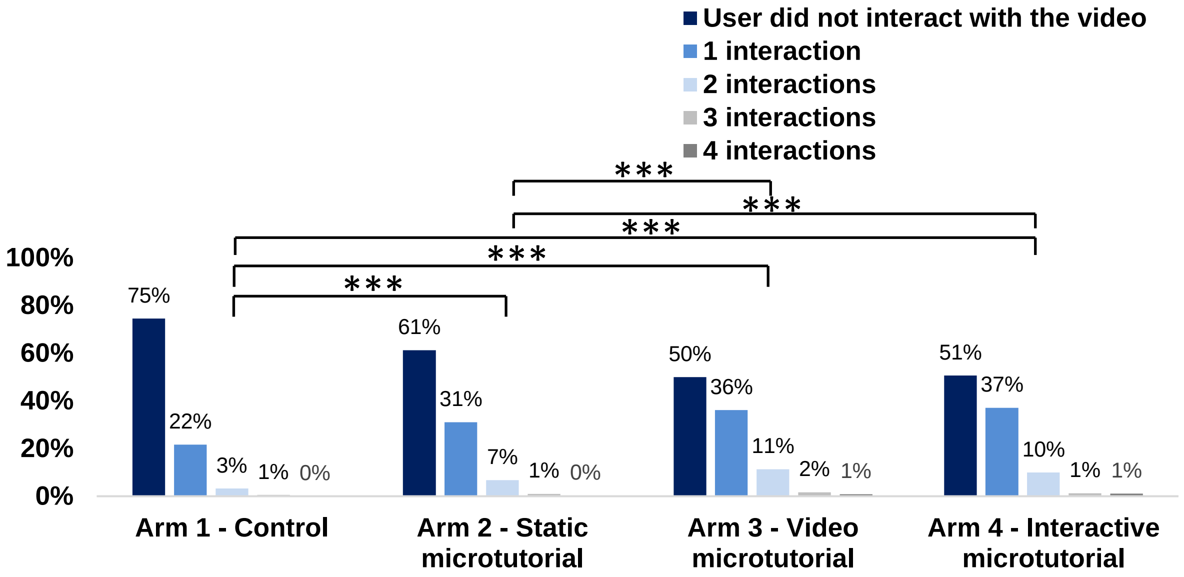


Figure 17. Proportion of times participants interacted with potentially harmful videos, by arm (multiple comparisons adjusted; * p < 0.05; ** p < 0.01; *** p < 0.001; modelled using a zero-inflated mixed-effects truncated Poisson regression model)

7.3.8. Disliking of potentially harmful videos

We found that the odds of disliking potentially harmful content were significantly higher in all treatment arms relative to the control condition (Table 18). The mean observed probability of disliking potentially harmful videos was 15% in the control arm, 24% in Arm 2 – Static, 34% in Arm 3 – Video, and 30% in Arm 4 – Interactive. The model used for inference was a mixed-effects logistic regression.

Table 18. Model-based estimates of the odds of disliking of potentially harmful videos

	Odds Ratio	95% CI	*z*-value	*P*
Intercept	0.05	0.04 – 0.07	-21.652	< 0.001
Arm 2 – Static	2.41	1.77 – 3.28	5.573	< 0.001
Arm 3 – Video	5.60	4.10 – 7.64	10.860	< 0.001
Arm 4 – Interactive	4.36	3.20 – 5.93	9.329	< 0.001

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

Table 19 shows that the differences were not sensitive to correcting for multiple comparisons. In addition, the odds of disliking potentially harmful videos were significantly higher in Arm 3 – Video and

Arm 4 – Interactive compared to Arm 1 – Control. There were no significant differences between Arm 3 – Video and Arm 4 – Interactive.

Table 19. Estimates of the odds of disliking 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 – Static	2.41	1.61 – 3.61	5.573	< 0.001
Arm 1 – Control vs Arm 3 – Video	5.60	3.73 – 8.42	10.860	< 0.001
Arm 1 – Control vs Arm 4 – Interactive	4.36	2.90 – 6.53	9.329	< 0.001
Arm 2 – Static vs Arm 3 – Video	2.33	1.59 – 3.42	5.663	< 0.001
Arm 2 – Static vs Arm 4 – Interactive	1.81	1.23 – 2.65	3.983	< 0.001
Arm 3 – Video vs Arm 4 – Interactive	0.78	0.54 – 1.13	-1.737	0.494

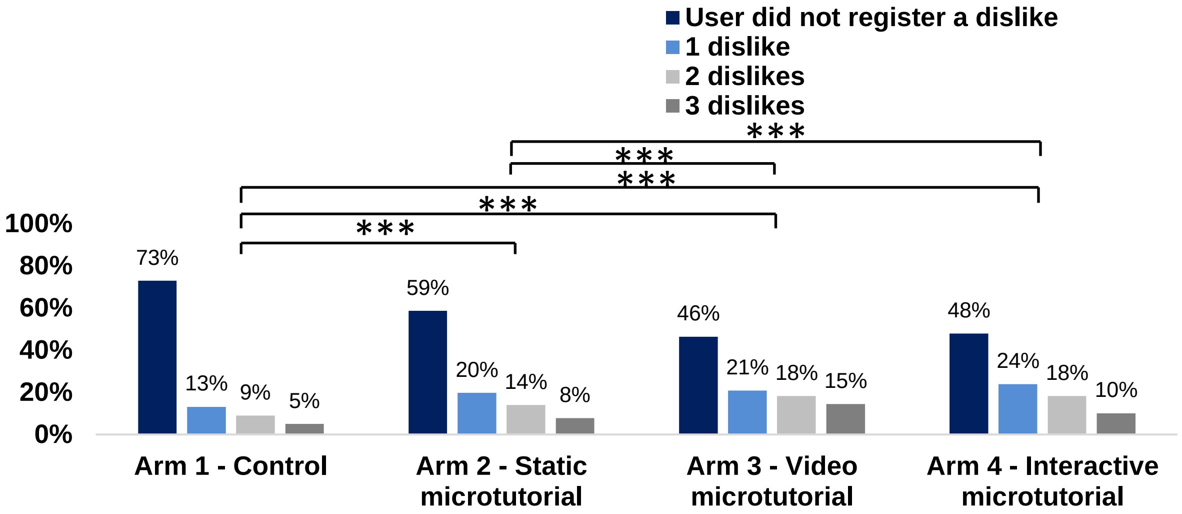


Figure 18. Proportion of times participants pressed the dislike button whilst watching potentially harmful videos, by arm (multiple comparisons adjusted; *** p < 0.001; modelled using a mixed-effects logistic regression model)

7.3.9. Skipping of potentially harmful videos

We found no significant differences in participants odds of skipping potentially harmful videos between treatment arms (Table 20). The mean observed probability of skipping potentially harmful videos was 56% in the control arm, 57% in Arm 2 – Static, 57% in Arm 3 – Video, and 59% in Arm 4 – Interactive. Figure 19 shows the percentage of skipped potentially harmful videos, by arm. The model used for inference was a mixed-effects logistic regression.

Table 20. Model-based estimates of the odds of skipping of potentially harmful content

	Odds Ratio	95% CI	*z*-value	*P*
Intercept	1.61	1.06 – 2.44	2.227	0.026
Arm 2 – Static	1.07	0.79 – 1.45	0.468	0.640
Arm 3 – Video	1.30	0.96 – 1.77	1.718	0.086
Arm 4 – Interactive	1.32	0.97 – 1.78	1.786	0.074

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

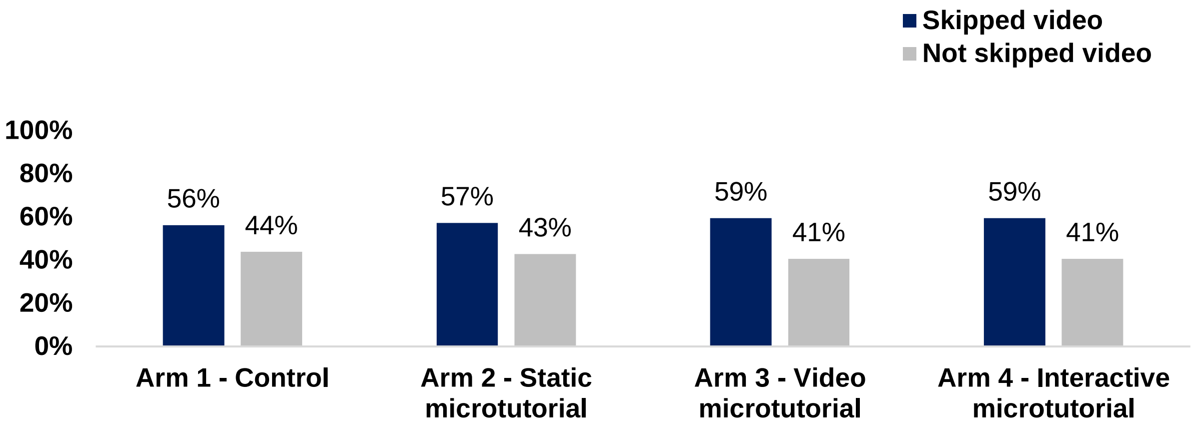


Figure 19. Percentage of skipped and not skipped potentially harmful videos, by arm (modelled using a mixed-effects logistic model)

7.3.10. Skipping of neutral videos

We found that participants who were shown interactive and static microtutorials were more likely to skip the neutral videos compared to those who were not shown any microtutorial. We also found that that participants who were shown the interactive microtutorial were more likely to skip the neutral videos compared to those who were shown the video microtutorial.

The mean observed probability of skipping potentially harmful videos was 48% in the control arm, 54% in Arm 2 – Static, 52% in Arm 3 – Video, and 58% in Arm 4 – Interactive. Table 21 shows that the odds of skipping neutral content were significantly higher in all treatment arms relative to the control condition. (Mixed-effects logistic regression model was used for inference.)

Table 21. Model-based estimates of the odds of skipping neutral videos

	Odds Ratio	95% CI	*z*-value	*P*
Intercept	0.86	0.39 – 1.91	-0.358	0.720
Arm 2 – Static	1.51	1.15 – 1.99	2.958	0.003
Arm 3 – Video	1.36	1.03 – 1.79	2.196	0.028
Arm 4 – Interactive	2.11	1.60 – 2.78	5.300	< 0.001

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

Adjusting for multiple comparisons, we found that participants who completed the interactive microtutorial (Arm 4) were significantly more likely to skip neutral videos than participants in the control condition and those who completed the video microtutorial (Table 22). Additionally, participants who completed the static microtutorial were significantly more likely to skip neutral content than those in the control condition. However, there was no significant difference between participants who completed the video microtutorial and those in the control condition.

Table 22. Estimates of the odds of 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*
Arm 1 – Control vs Arm 2 – Static	1.51	1.06 – 2.17	2.958	0.019
Arm 1 – Control vs Arm 3 – Video	1.36	0.95 – 1.95	2.196	0.169
Arm 1 – Control vs Arm 4 – Interactive	2.11	1.47 – 3.03	5.300	< 0.001
Arm 2 – Static vs Arm 3 – Video	0.90	0.63 – 1.29	-0.766	1
Arm 2 – Static vs Arm 4 – Interactive	1.39	0.97 – 2.00	2.368	0.107
Arm 3 – Video vs Arm 4 – Interactive	1.55	1.08 – 2.23	3.130	0.011

Figure 20 shows the percentage of participants skipping neutral videos, by arm. Significant differences, as estimated using the primary outcome model, are shown using horizontal lines.

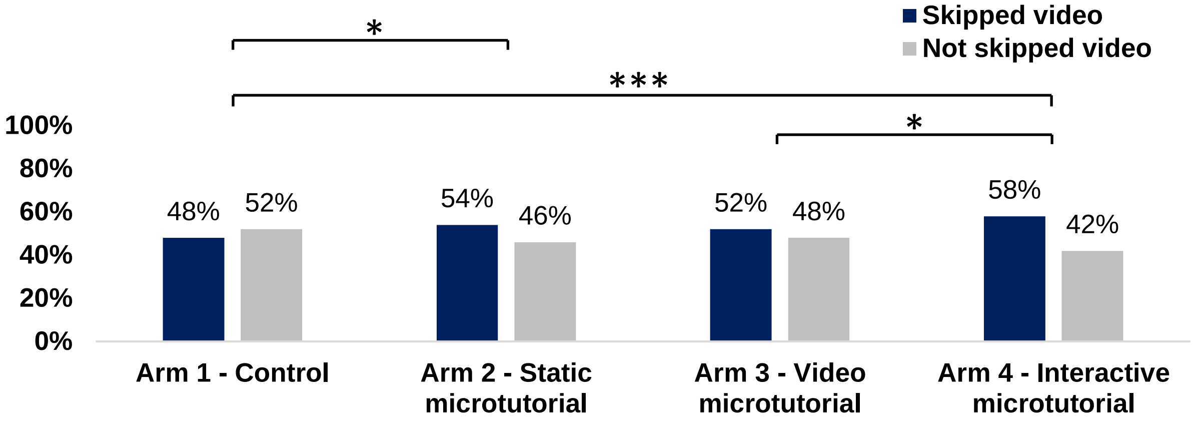


Figure 20. Percentage of skipped and not skipped neutral videos, by arm (multiple comparisons adjusted; * p < 0.05; *** p < 0.001; modelled using a mixed-effects logistic model)

7.3.11. Time spent watching potentially harmful videos

We found no significant treatment effect on the median or mean length of time participants spent viewing potentially harmful videos. Table 23 provides the mean, standard deviation and median values for participant viewing time across each arm. We analysed the length of time participants spent viewing potentially harmful videos using a mixed-effects robust linear regression model. Table 24 shows the estimates of the model.

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

Arm	Mean	SD	Median
Arm 1 – Control	27.82	20.09	29.72
Arm 2 – Active choice	27.01	20.13	26.99
Arm 3 – Auto-skip	26.86	20.04	26.59
Arm 4 – Auto-play	26.51	20.23	26.11

Table 24. Model-based estimates of time spent watching potentially harmful videos

	Estimates	95% CI	*z*-value	*P*
Intercept	27.83	22.82 – 32.84	10.879	< 0.001
Arm 2 – Static	-0.68	-2.76 – 1.41	-0.636	0.525
Arm 3 – Video	-0.94	-3.02 – 1.15	-0.882	0.378
Arm 4 – Interactive	-1.16	-3.25 – 0.92	-1.091	0.275

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

7.4. Reporting Capability Task Analysis

7.4.1. Reporting during the reporting capability task

All participants were prompted to complete a report on the next video that was shown. Participants shown the microtutorials were more likely to complete a report of a potentially harmful video when prompted, compared to those not shown a microtutorial. Some microtutorials were more effective than others. Specifically, participants shown the interactive microtutorial were more likely to complete a report when prompted compared to participants shown the static or video microtutorials, whilst participants shown the video microtutorial were more likely to complete the report than those shown the static microtutorial.

We analysed the likelihood of participants completing a report of a potentially harmful video in the reporting capability task using a logistic regression model. Unlike in the analysis of the primary outcome, it was not necessary to use zero-inflated Poisson models, as there was no evidence of zero-inflation. This is likely due to participants receiving a direct prompt to report the video during the reporting capability task.

The mean observed probability of completing a report during the reporting capability task was 41% in the control arm, 48% in Arm 2 – Static, 51% in Arm 3 – Video, and 60% in Arm 4 – Interactive. Table 25 shows that the odds of completing a report was significantly higher in all treatment arms than in the control.

Table 25. Model-based estimates of reporting during the reporting capability task

	Odds Ratios	95% CI	*z*-value	*P*
Intercept	0.68	0.59 – 0.79	-5.018	< 0.001
Arm 2 – Static	1.36	1.10 – 1.68	2.872	0.004
Arm 3 – Video	1.54	1.25 – 1.90	4.052	< 0.001
Arm 4 – Interactive	2.17	1.75 – 2.68	7.172	< 0.001

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

Adjusting for multiple comparisons, we found that participants in Arm 4 – Interactive were also significantly more likely to complete reports during the reporting capability task than participants in Arm 2 – Static or Arm 3 – Video (Table 26).

Table 26. Estimates of the odds of 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*
Arm 1 – Control vs Arm 2 – Static	1.36	1.03 – 1.79	2.872	0.025
Arm 1 – Control vs Arm 3 – Video	1.54	1.17 – 2.03	4.052	< 0.001
Arm 1 – Control vs Arm 4 – Interactive	2.17	1.64 – 2.86	7.172	< 0.001
Arm 2 – Static vs Arm 3 – Video	1.13	0.86 – 1.49	1.190	1
Arm 2 – Static vs Arm 4 – Interactive	1.59	1.21 – 2.10	4.363	< 0.001
Arm 3 – Video vs Arm 4 – Interactive	1.41	1.07 – 1.85	3.186	0.009

Figure 21 shows the percentage of participants completing a report during the reporting capability task. Significant differences, as estimated by the model, are shown using horizontal lines.

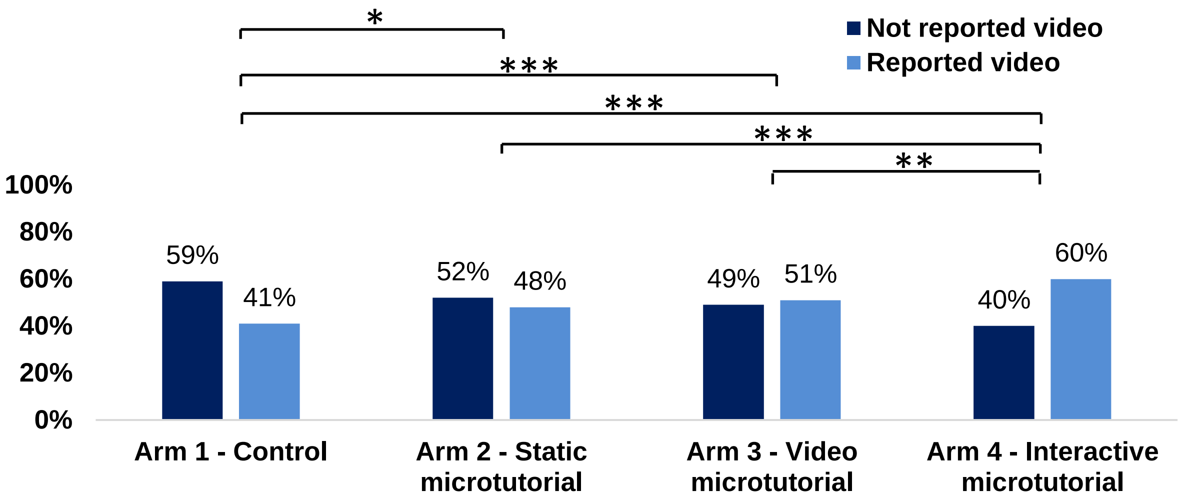


Figure 21. Percentage of reported and not reported potentially harmful videos, by arm (multiple comparisons adjusted; * p < 0.05; ** p < 0.01; *** p < 0.001; modelled using a mixed logistic regression model)

7.4.2. Reasons for not reporting during the reporting capability task

A higher proportion of participants in the control condition indicated that not knowing how to complete a report (41%) was the primary reason for not completing a report during the reporting capability task relative to participants in all three treatment arms (Arm 1 – Static: 36%, Arm 2 – Video: 22%, Arm 4 – Interactive: 20%). Figure 22 shows the distribution of responses participants gave in response to the follow-up reporting capability question about why they chose not to report the potentially harmful video when prompted. Figure 23 shows the raw count of responses. Note that we did not ask participants to give a reason when they selected “Other”.

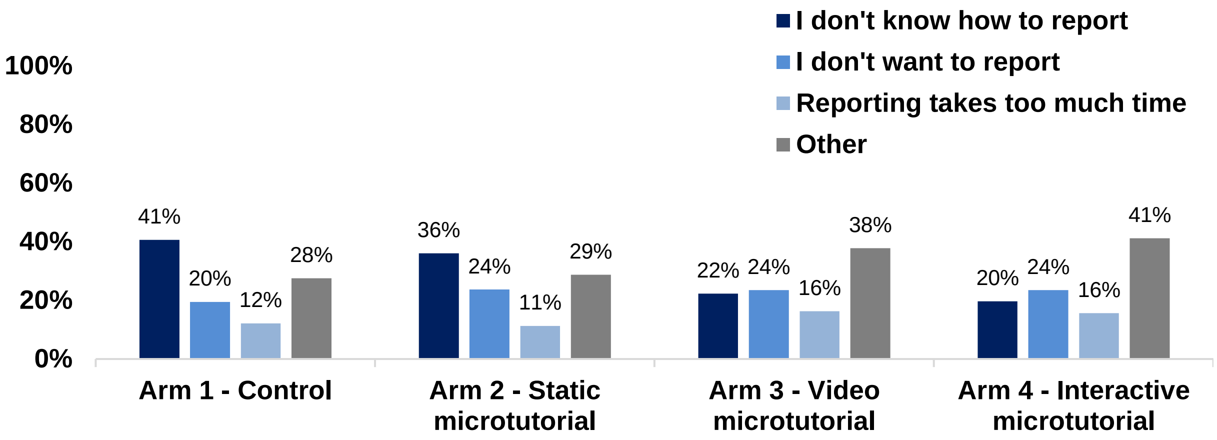


Figure 22. Participant responses to the follow-up reporting capability question

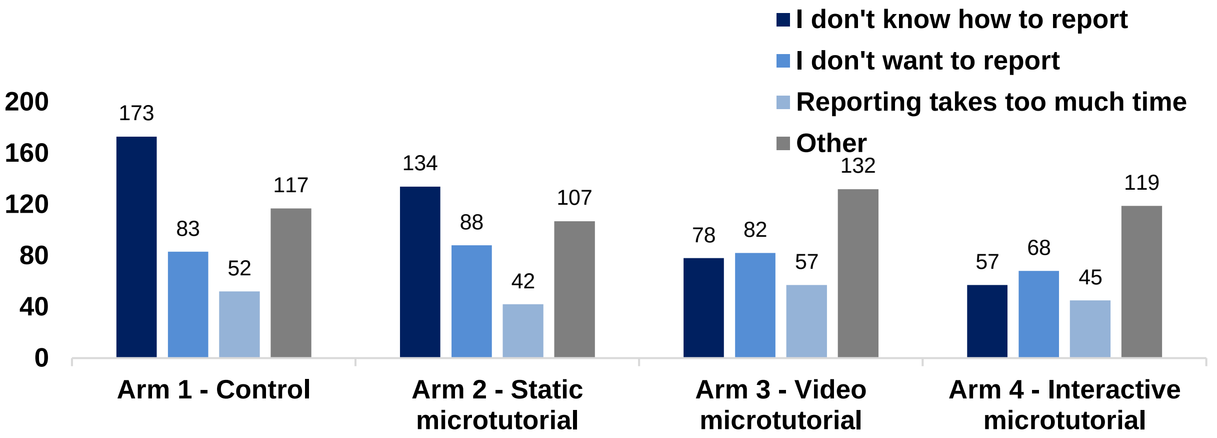


Figure 23. Participant responses to the follow-up reporting capability question (raw count)

7.4.3. Accuracy of reporting during the reporting capability task

We found no significant differences between arms in the accuracy of the reporting of the potentially harmful videos during the reporting capability task. As previously specified, a completed report was classified as accurate if the category the respondent selected during the reporting process matched a pre-specified content label. The video shown in the reporting capability task was classified as ‘Violent content’.

The mean observed probability of the reporting category being accurate, given that a participant had completed a report during the reporting capability task was 40% in the control arm, 43% in Arm 2 – Static, 45% in Arm 3 – Video, and 41% in Arm 4 – Interactive. Table 27 shows that there were no significant differences across treatment arms, and Figure 24 shows the percentage of participants correctly categorising the potentially harmful content shown during the reporting capability task.

Table 27. Model-based estimates of accuracy of categorisation of potentially harmful content during the reporting capability task

	Odds Ratio	95% CI	*z*-value	*P*
Intercept	0.66	0.52 – 0.83	-3.498	< 0.001
Arm 2 – Static	1.15	0.84 – 1.58	0.857	0.391
Arm 3 – Video	1.23	0.90 – 1.68	1.295	0.195
Arm 4 – Interactive	1.04	0.76 – 1.41	0.231	0.818

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

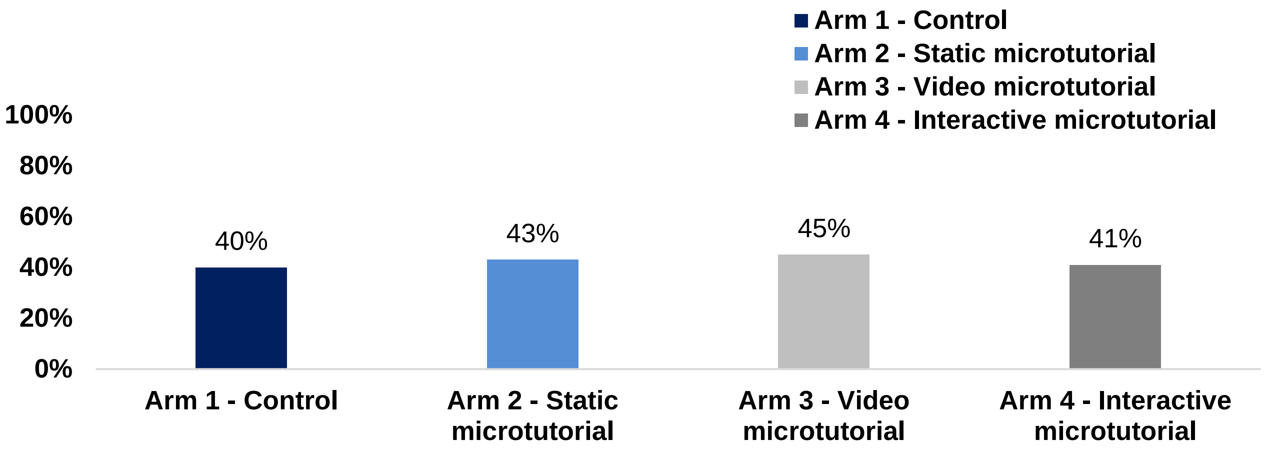


Figure 24. Proportion of completed reports that accurately classify the content of the potentially harmful content shown during the reporting capability task, by arm

7.5. Responses to Attitudinal Questions

In this section, we used ordinal regression models for making inferences. As in the section above, significance levels in figures are indicated using asterisks. Specifically, * p < 0.05; ** p < 0.01; *** p < 0.001. In each case where a significant difference is shown, the reported p-values are adjusted for multiple comparisons using the Bonferroni correction.

7.5.1. Confidence using features of VSP outside of the experiment

Participants who completed the Video (Arm 3) and Interactive (Arm 4) microtutorials reported being significantly more confident using features of a VSP outside of the experiment than participants in the control condition (Table 28). There was no significant difference between participants in the static microtutorial arm and those in the control condition.

Table 28. Model-based estimates of confidence in VSP use outside the experiment

	Odds Ratios	95% CI	*z*-value	*P*
Arm 2 – Static	1.18	0.97 – 1.42	1.695	0.090
Arm 3 – Video	1.50	1.24 – 1.81	4.208	< 0.001
Arm 4 – Interactive	1.36	1.13 – 1.65	3.224	0.001

Note: Arm 1 – Control is the reference level other arms are compared against; Excluding 59 “Don’t know” responses

The reported differences between groups were not sensitive to adjustment for multiple comparisons (Table 29). There were no differences in confidence in using features of a VSP outside of the experiment between participants in treatment groups.

Table 29. Estimates of confidence in VSP use outside of the experiment (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 – Static	1.18	0.92 – 1.51	1.695	0.540
Arm 1 – Control vs Arm 3 – Video	1.50	1.17 – 1.91	4.208	< 0.001
Arm 1 – Control vs Arm 4 – Interactive	1.36	1.07 – 1.75	3.224	0.008
Arm 2 – Static vs Arm 3 – Video	1.27	0.99 – 1.63	2.457	0.084
Arm 2 – Static vs Arm 4 – Interactive	1.16	0.90 – 1.49	1.502	0.799
Arm 3 – Video vs Arm 4 – Interactive	0.91	0.71 – 1.17	-0.945	1

Figure 25 shows the distribution of participant responses to the prompt ‘I feel confident using features of video sharing platforms’.

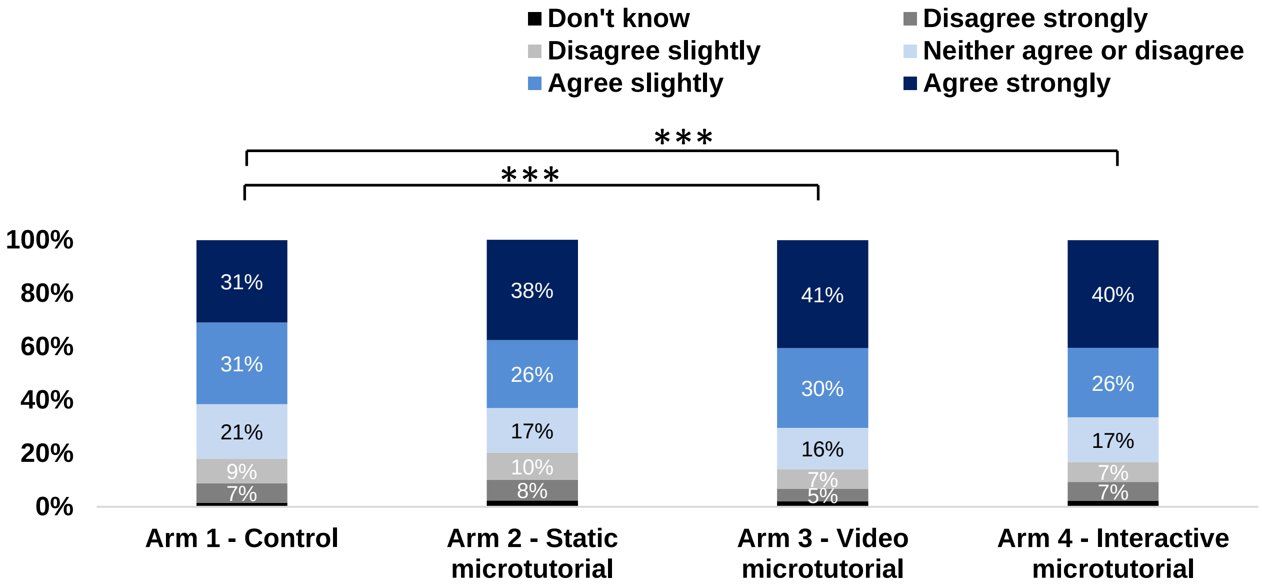


Figure 25. Participant responses to the prompt 'I feel confident using features of video sharing platforms' by arm

7.5.2. Ease of use

Participants in the video and interactive treatment arm were significantly more likely to indicate that they found the platform easy to use than participants in the control arm (Table 30). There were no observed differences in reported ease of use between participants who completed the static microtutorial and participants in the control arm.

Table 30. Model-based estimates of Ease of VSP use

	Odds Ratios	95% CI	*z*-value	*P*
Arm 2 – Static	1.18	0.97 – 1.43	1.666	0.096
Arm 3 – Video	1.71	1.40 – 2.08	5.266	< 0.001
Arm 4 – Interactive	1.61	1.32 – 1.96	4.660	< 0.001

Note: Arm 1 – Control is the reference level other arms are compared against; Excluding 46 "Don't know" responses

Adjusting for multiple comparisons, we also found that participants who completed the interactive and video microtutorials were significantly more likely to indicate that they found the platform features easy to use, relative to participants who completed the static microtutorial (Table 31). The reported differences between participants in control condition and participants who completed the video and interactive microtutorials were not sensitive to adjustment for multiple comparisons.

Table 31. Estimates of Ease of VSP use (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 – Static	1.18	0.91 – 1.52	1.666	0.574
Arm 1 – Control vs Arm 3 – Video	1.71	1.32 – 2.21	5.266	< 0.001
Arm 1 – Control vs Arm 4 – Interactive	1.61	1.24 – 2.09	4.660	< 0.001
Arm 2 – Static vs Arm 3 – Video	1.45	1.11 – 1.88	3.630	0.002
Arm 2 – Static vs Arm 4 – Interactive	1.36	1.05 – 1.77	3.028	0.015
Arm 3 – Video vs Arm 4 – Interactive	0.94	0.72 – 1.23	-0.586	1

Figure 26 shows the distribution of participant responses to the prompt 'I feel confident using features of video sharing platforms'.

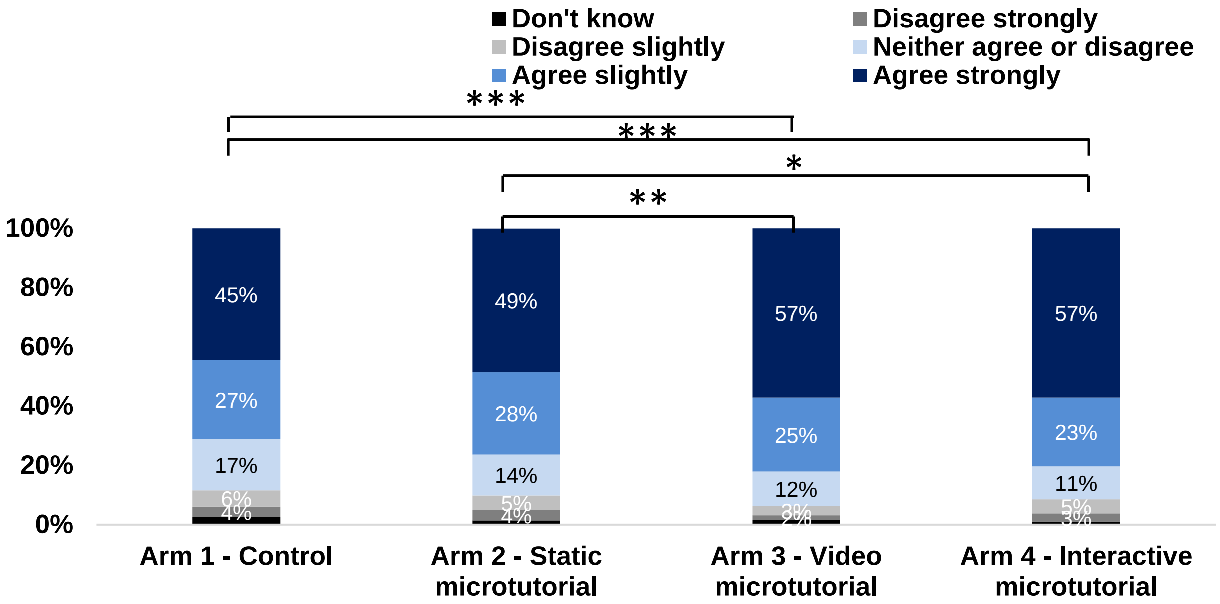


Figure 26. Participant responses to the prompt ‘I found the features of the platform (e.g., liking, disliking, reporting) easy to use’, by arm

7.5.3. Opportunities to learn

Participants in each treatment arm were significantly more likely to report that there were opportunities to learn how to use the video sharing platform than participants in the control condition (Table 32).

Table 32. Model-based estimates of reported opportunities to learn during the trial

	Odds Ratios	95% CI	*z*-value	*P*
Arm 2 – Static	2.26	1.86 – 2.75	8.213	< 0.001
Arm 3 – Video	3.25	2.66 – 3.96	11.623	< 0.001
Arm 4 – Interactive	3.34	2.74 – 4.08	11.885	< 0.001

Note: Arm 1 – Control is the reference level other arms are compared against; excluding 87 “Don’t know” responses

Adjusting for multiple comparisons, we also found that participants who completed the interactive and video microtutorials were significantly more likely to indicate that there were opportunities to learn how to use the video sharing platform, relative to participants who completed the static microtutorial (Table 33). The reported differences between participants in the treatment conditions and those in the control condition were not sensitive to adjustment for multiple comparisons.

Table 33. Estimates of reported opportunities to learn during the trial (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 – Static	2.26	1.75 – 2.92	8.213	< 0.001
Arm 1 – Control vs Arm 3 – Video	3.25	2.50 – 4.21	11.623	< 0.001
Arm 1 – Control vs Arm 4 – Interactive	3.34	2.57 – 4.34	11.885	< 0.001
Arm 2 – Static vs Arm 3 – Video	1.43	1.11 – 1.85	3.673	0.001
Arm 2 – Static vs Arm 4 – Interactive	1.48	1.15 – 1.90	3.957	< 0.001
Arm 3 – Video vs Arm 4 – Interactive	1.03	0.80 – 1.33	0.282	1

Figure 27 shows the distribution of participant responses to the prompt ‘There were opportunities to learn how to use the video sharing platform’.

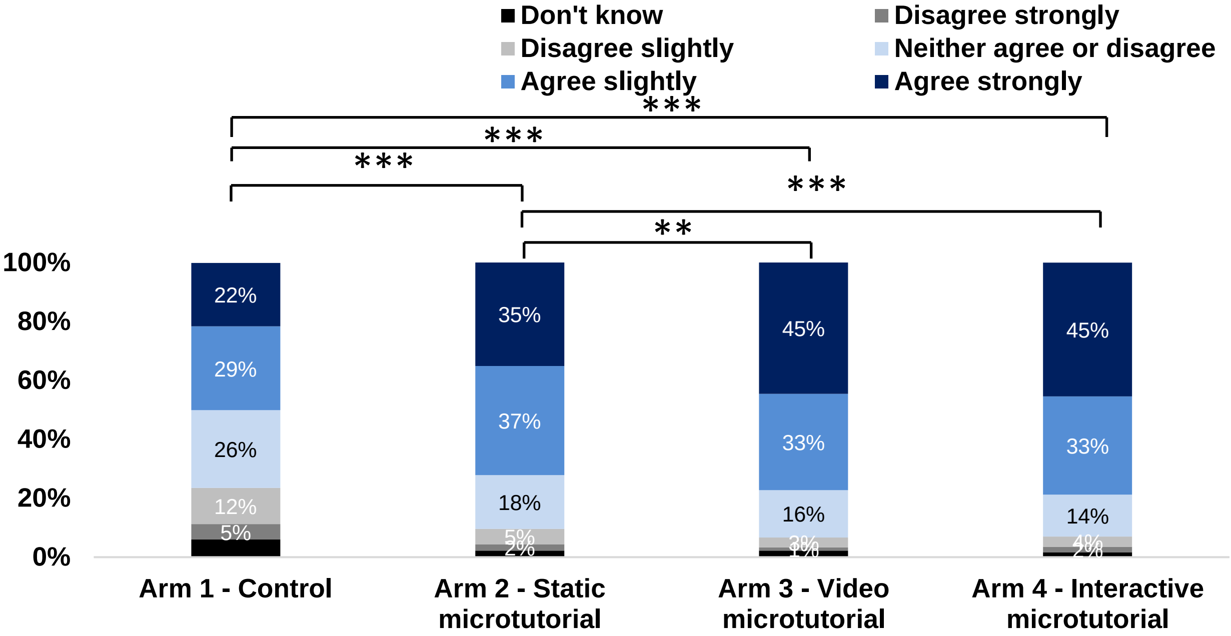


Figure 27. Participant responses to the prompt ‘There were opportunities to learn how to use the video sharing platform’, by arm

7.5.4. Opportunities to report

Similarly, we found that participants in each treatment arm were significantly more likely to indicate that there were opportunities to report content than participants in the control arm (Table 34).

Table 34. Model-based estimates of reported opportunities to report during the trial

	Odds Ratios	95% CI	*z*-value	*P*
Arm 2 – Static	1.48	1.21 – 1.80	3.880	< 0.001
Arm 3 – Video	1.82	1.49 – 2.22	5.916	< 0.001
Arm 4 – Interactive	2.20	1.80 – 2.69	7.649	< 0.001

Note: Arm 1 – Control is the reference level other arms are compared against; excluding 65 “Don’t know” responses

Adjusting for multiple comparisons, we also found that participants who completed the interactive microtutorial were significantly more likely to indicate there were opportunities to report content than participants who completed the static microtutorial (Table 35).

Table 35. Estimates of reported opportunities to report during the trial (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 – Static	1.48	1.14 – 1.92	3.880	< 0.001
Arm 1 – Control vs Arm 3 – Video	1.82	1.40 – 2.36	5.916	< 0.001
Arm 1 – Control vs Arm 4 – Interactive	2.20	1.69 – 2.87	7.649	< 0.001
Arm 2 – Static vs Arm 3 – Video	1.23	0.95 – 1.60	2.030	0.254
Arm 2 – Static vs Arm 4 – Interactive	1.49	1.14 – 1.94	3.821	< 0.001
Arm 3 – Video vs Arm 4 – Interactive	1.21	0.93 – 1.58	1.823	0.410

Figure 28 shows the distribution of participant responses to the prompt ‘The design of the video sharing platform provided opportunities for me to report videos’.

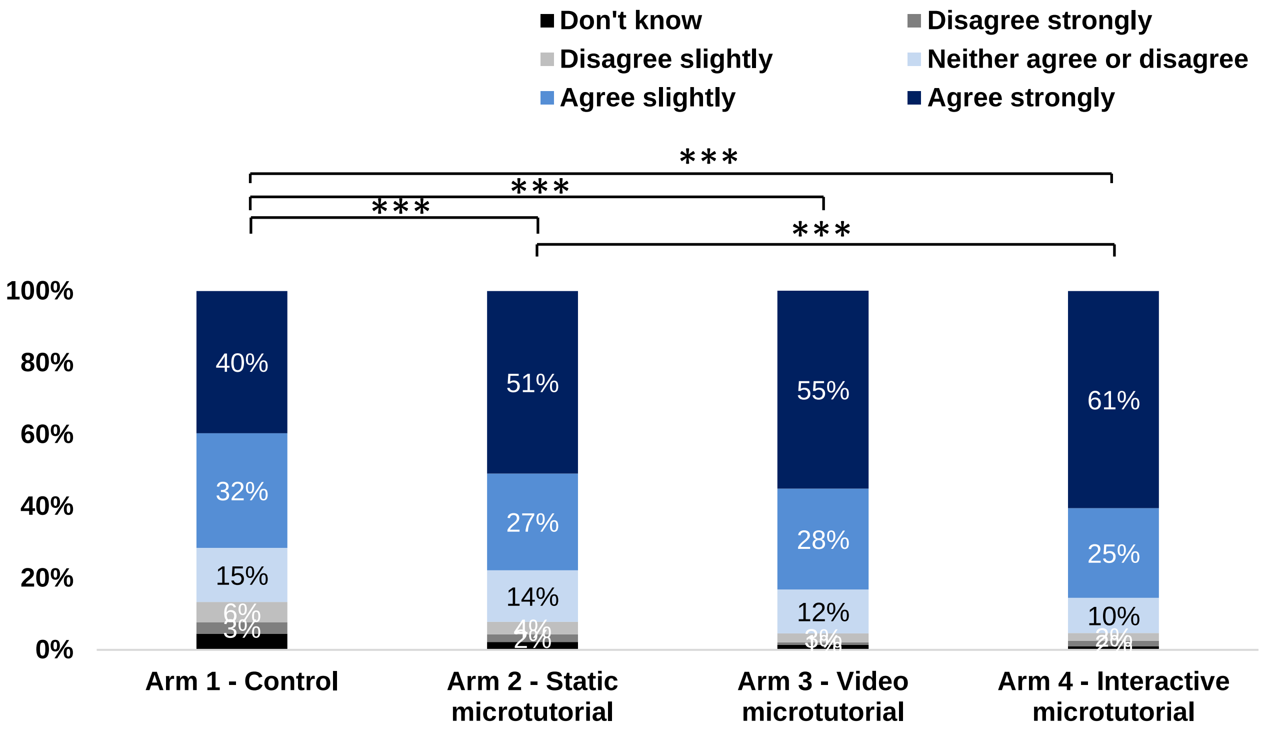


Figure 28. Participant responses to the prompt ‘The design of the video sharing platform provided opportunities for me to report videos’, by arm

7.5.5. Questions asked to participants shown microtutorials

The next series of questions were only asked to participants who were shown static, video or interactive microtutorials. These questions asked about features of the microtutorial. In the questions, we referred to the microtutorial as an “Introduction” (this was considered simpler than explaining to participants what the microtutorial was).

7.5.5.1. Confidence to report

Participants who watched the video microtutorial reported significantly higher confidence to report videos than participants who completed the static microtutorial (Table 36).

Table 36. Model-based estimates of reported confidence to report during the trial

	Odds Ratios	95% CI	*z*-value	*P*
Arm 3 – Video	1.41	1.16 – 1.70	3.538	< 0.001
Arm 4 – Interactive	0.83	0.68 – 1.00	-1.939	0.052

Note: Arm 2 – Static is the reference level other arms are compared against; excluding 32 “Don’t know” responses

After adjusting for multiple comparisons, we also found that participants who completed the interactive microtutorial were significantly less likely to indicate that the introduction gave them confidence to report relative to participants who watched the video microtutorial (Table 37).

Table 37. Estimates of reported confidence to report during the trial (p values and Cis corrected for multiple comparisons using the Bonferroni correction)

Comparison	Odds Ratios	95% CI	*z*-value	*P*
Arm 2 – Static vs Arm 3 – Video	1.41	1.12 – 1.76	3.538	0.001
Arm 2 – Static vs Arm 4 – Interactive	0.83	0.66 – 1.04	-1.939	0.157
Arm 3 – Video vs Arm 4 – Interactive	1.70	1.35 – 2.13	5.475	< 0.001

Figure 29 shows the distribution of participant responses to the prompt ‘The introduction gave me confidence to report videos’.

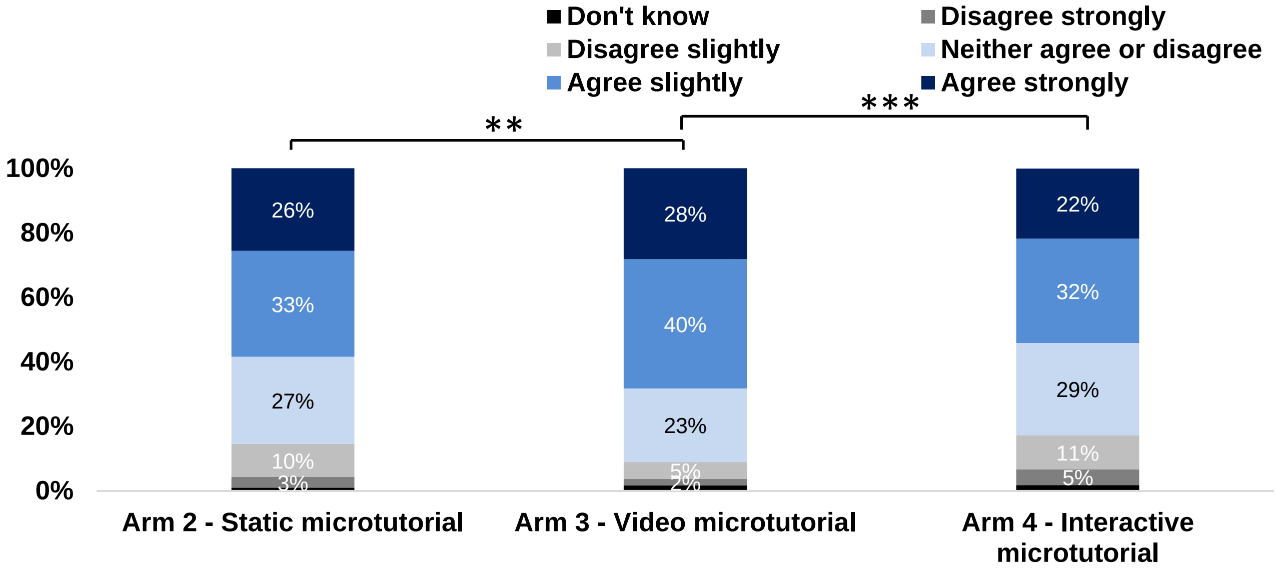


Figure 29. Participant responses to the prompt 'The introduction gave me confidence to report videos', by arm

7.5.5.2. Need for introduction

There were no significant differences in participants' perceived need for an introduction to video sharing platforms (Table 38). In each treatment arm, nearly half of participants indicated that they felt they did not need an introduction (Arm 2: 44%, Arm 3: 46%, Arm 4: 47%).

Table 38. Model-based estimates of reported need for introduction

	Odds Ratios	95% CI	*z*-value	*P*
Arm 3 – Video	1.14	0.94 – 1.37	1.330	0.184
Arm 4 – Interactive	1.17	0.97 – 1.40	1.630	0.103

Note: Arm 2 – Static is the reference level other arms are compared against; excluding 21 "Don't know" responses

Figure 30 shows the distribution of participant responses to the prompt 'I did not need the introduction'.

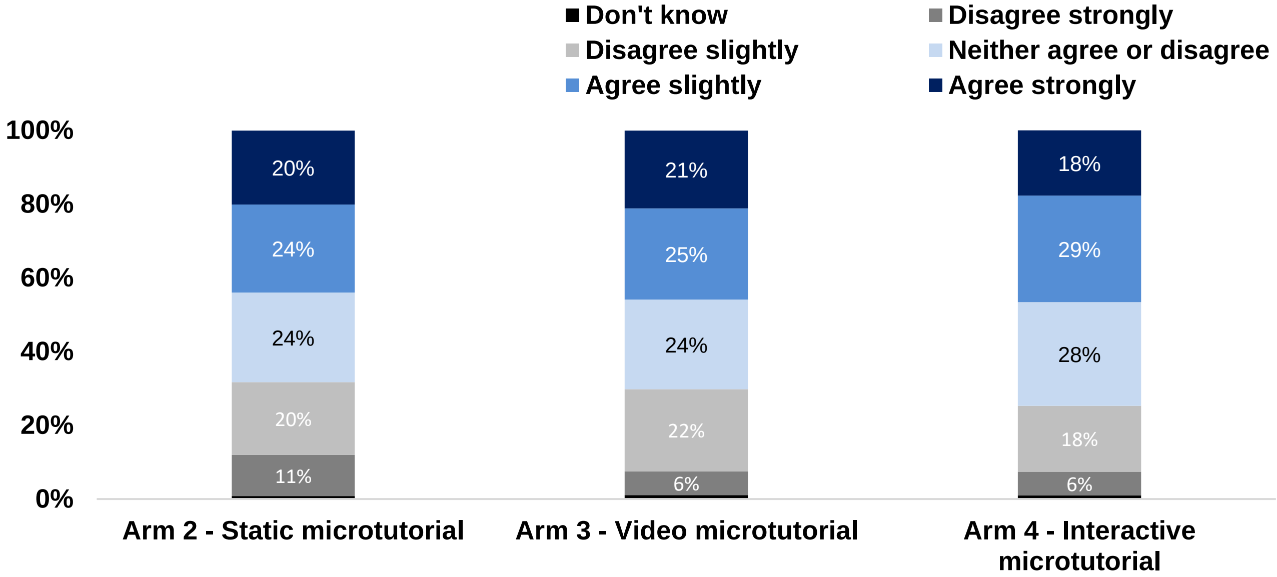


Figure 30. Participant responses to the prompt 'I did not need the introduction', by arm

7.5.5.3. Annoyingness of introduction

Participants who completed the interactive microtutorial were significantly more likely to report that they found the introduction annoying, relative to participants who completed the static microtutorial (Table 39). There was no observed difference in likelihood of reporting the introduction as annoying between participants who completed the video and static microtutorials.

Table 39. Model-based estimates of reported annoyingness of introduction

	Odds Ratios	95% CI	*z*-value	*P*
Arm 3 – Video	1.08	0.90 – 1.31	0.851	0.395
Arm 4 – Interactive	1.54	1.28 – 1.85	4.565	< 0.001

Note: Arm 2 – Static is the reference level other arms are compared against; excluding 16 "Don't know" responses

Adjusting for multiple comparisons, we also found that participants who completed the interactive microtutorial were significantly more likely to report that they found the introduction annoying, relative to participants who completed the video microtutorial (Table 40). The previously reported differences between groups were not sensitive to adjustment for multiple comparisons.

Table 40. Estimates of reported annoyingness of introduction (p values and CIs corrected for multiple comparisons using the Bonferroni correction)

Comparison	Odds Ratios	95% CI	*z*-value	*P*
Arm 2 – Static vs Arm 3 – Video	1.08	0.87 – 1.35	0.851	1
Arm 2 – Static vs Arm 4 – Interactive	1.54	1.23 – 1.92	4.565	< 0.001
Arm 3 – Video vs Arm 4 – Interactive	0.70	0.56 – 0.88	-3.669	< 0.001

Figure 31 shows the distribution of participant responses to the prompt 'I found the introduction annoying'.

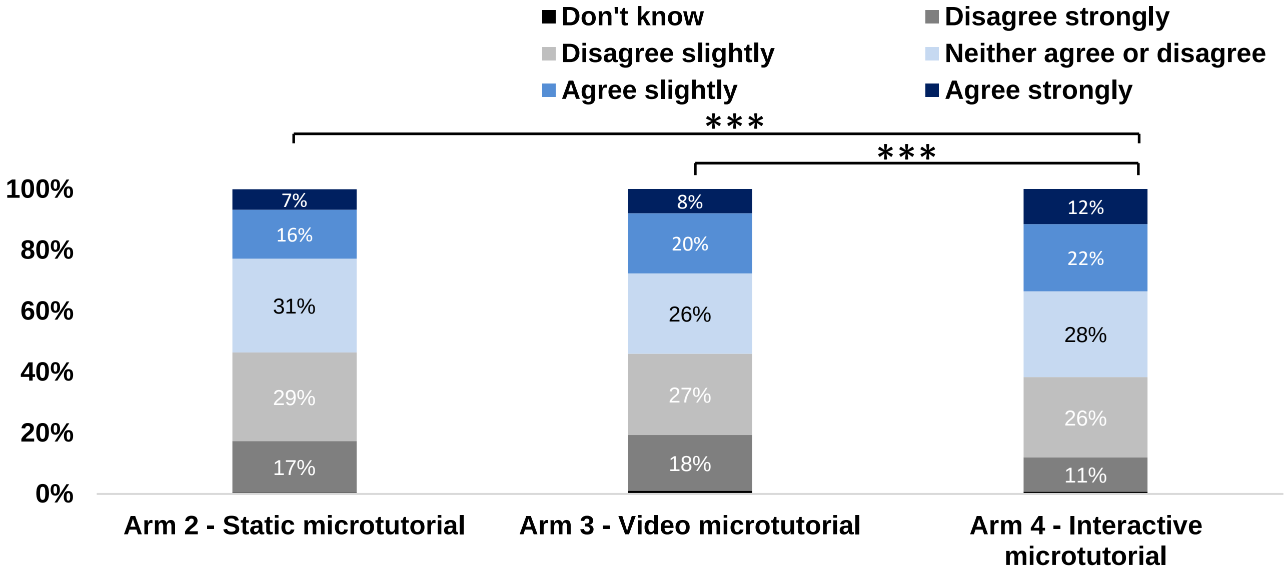


Figure 31. Participant responses to the prompt 'I found the introduction annoying', by arm

7.5.5.4. Simplicity of introduction

There were no significant differences in participants' beliefs that the introduction was 'too simple' across treatment arms. Indeed, across all treatment arms only approximately one quarter of respondents indicated that they found the introduction too simple (Table 41).

Table 41. Model-based estimates of reported simplicity of introduction

	Odds Ratios	95% CI	*z*-value	*P*
Arm 3 – Video	0.97	0.80 – 1.17	-0.334	0.738
Arm 4 – Interactive	1.04	0.86 – 1.26	0.439	0.661

Note: Arm 2 – Static is the reference level other arms are compared against; excluding 20 "Don't know" responses

Figure 32 shows the distribution of participant responses to the prompt 'I found the introduction too simple'.

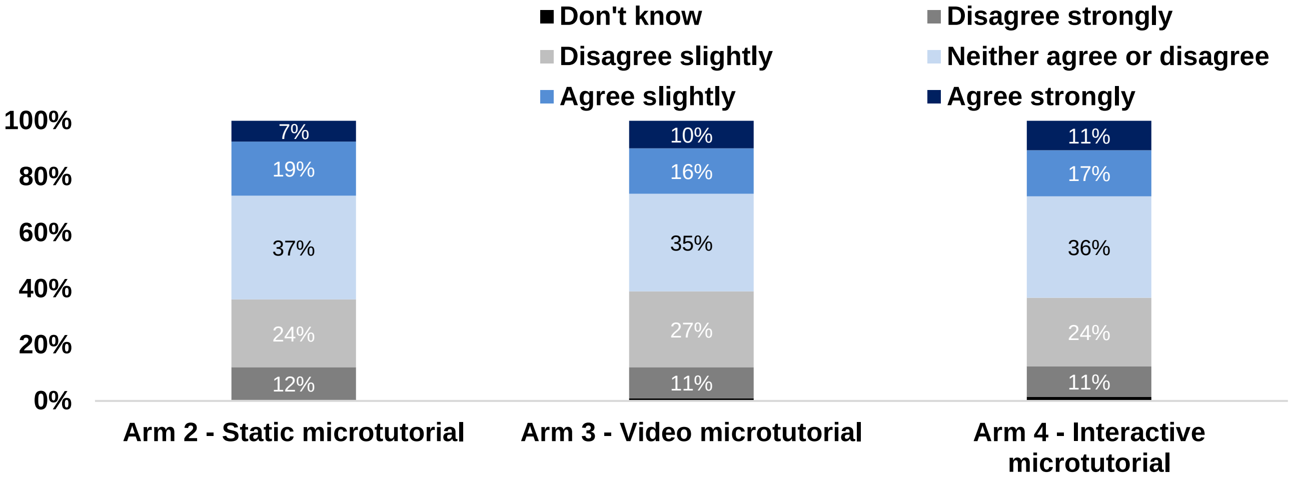


Figure 32. Participant responses to the prompt ‘I found the introduction too simple’, by arm

7.5.5.5. Length of introduction

Participants who completed the interactive or video microtutorials were significantly more likely to report that the introduction was too long than participants who completed the static microtutorial (Table 42).

Table 42. Model-based estimates of length of introduction

	Odds Ratios	95% CI	*z*-value	*P*
Arm 3 – Video	1.47	1.22 – 1.77	4.086	< 0.001
Arm 4 – Interactive	2.48	2.06 – 3.00	9.455	< 0.001

Note: Arm 2 – Static is the reference level other arms are compared against; excluding 20 “Don’t know” responses

After adjusting for multiple comparisons, we also found that participants who completed the interactive microtutorial were significantly more likely to report that the introduction was too long than participants who completed the video microtutorial (Table 43). The previously reported differences were not sensitive to adjustment for multiple comparisons.

Table 43. Estimates of perceived length of introduction (p values and CIs corrected for multiple comparisons using the Bonferroni correction)

Comparison	Odds Ratios	95% CI	*z*-value	*P*
Arm 2 – Static vs Arm 3 – Video	1.47	1.18 – 1.83	4.086	< 0.001
Arm 2 – Static vs Arm 4 – Interactive	2.48	1.98 – 3.11	9.455	< 0.001
Arm 3 – Video vs Arm 4 – Interactive	0.59	0.47 – 0.74	-5.443	< 0.001

Figure 33 shows the distribution of participant responses to the prompt ‘the introduction was too long’.

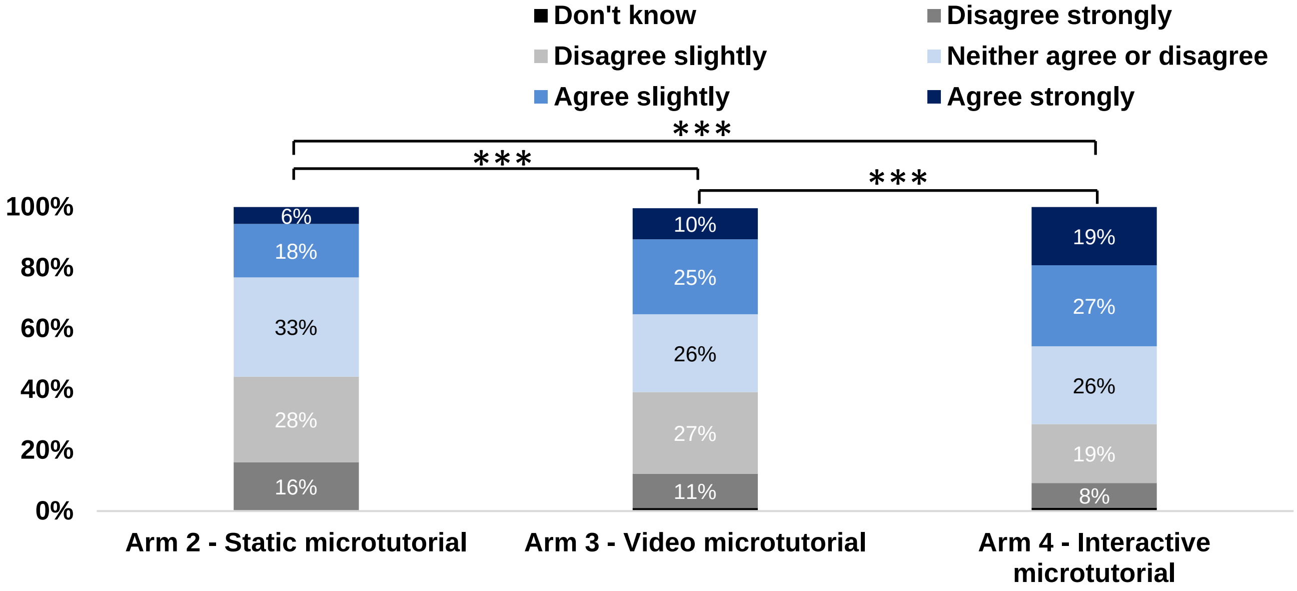


Figure 33. Participant responses to the prompt ‘the introduction was too long’, by arm

7.5.5.6. Engagingness of introduction

There were no significant differences in the reported engagingness of the introduction across treatment arms (Table 44). In each arm, less than one quarter of participants indicated that they did not find the introduction engaging.

Table 44. Model-based estimates of engagingness of introduction

	Odds Ratios	95% CI	*z*-value	*P*
Arm 3 – Video	1.19	0.99 – 1.43	1.841	0.066
Arm 4 – Interactive	1.12	0.93 – 1.35	1.169	0.242

Note: Arm 2 – Static is the reference level other arms are compared against; excluding 12 “Don’t know” responses

Figure 34 shows the distribution of participant responses to the prompt ‘I found the introduction engaging’.

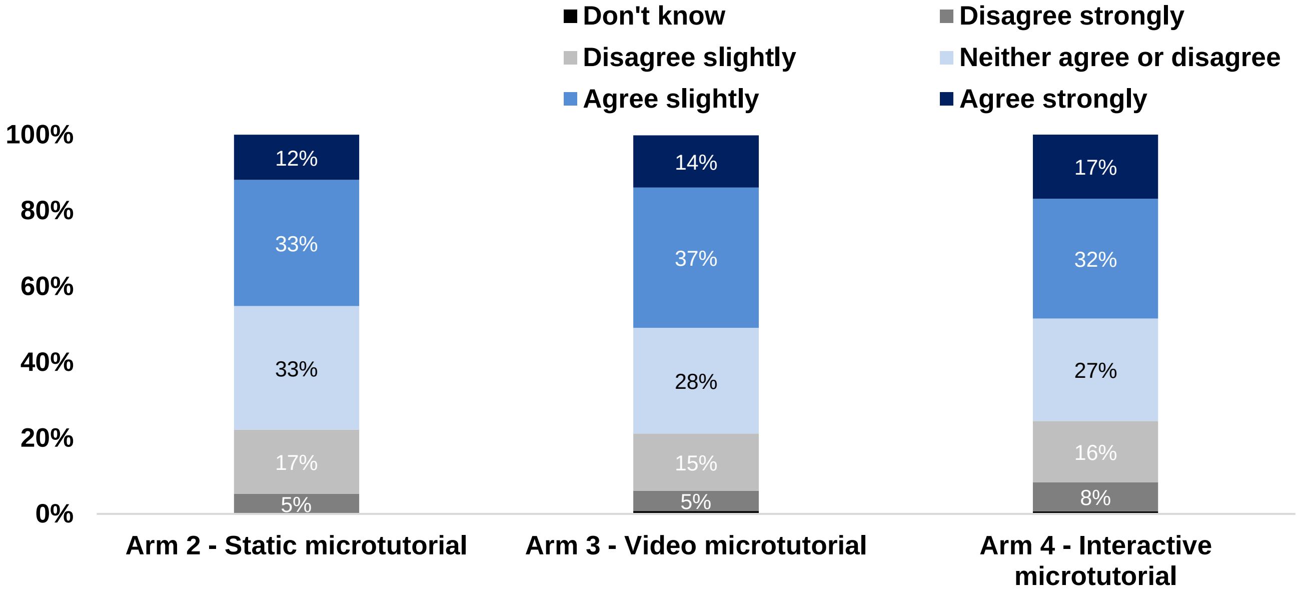


Figure 34. Participant responses to the prompt ‘I found the introduction engaging’, by arm

7.5.5.7. Learned something new during the introduction

We found no significant differences between arms in the likelihood of participants agreeing that they learned something new during the introduction. The model estimates show a significant difference in the likelihood of participants agreeing that they learnt something new during the introduction amongst those who completed the interactive and static microtutorials (Table 45).

Table 45. Model-based estimates of learning something new during the introduction

	Odds Ratios	95% CI	*z*-value	*P*
Arm 3 - Video	1.20	1.00 – 1.45	1.943	0.052
Arm 4 – Interactive	1.20	1.00 – 1.45	1.966	0.049

Note: Arm 2 – Static is the reference level other arms are compared against; excluding 23 “Don’t know” responses

However, after adjusting for multiple comparisons, the differences are no longer significant (Table 46).

Table 46. Estimates of reported learning of something new (p values and CIs corrected for multiple comparisons using the Bonferroni correction)

Comparison	Odds Ratios	95% CI	*z*-value	*P*
Arm 2 – Static vs Arm 3 – Video	1.20	0.96 – 1.51	1.943	0.156
Arm 2 – Static vs Arm 4 – Interactive	1.20	0.97 – 1.50	1.966	0.148
Arm 3 – Video vs Arm 4 – Interactive	1.00	0.80 – 1.25	0.004	1

Figure 35 shows the distribution of participant responses to the prompt ‘I learned something new by going through the introduction’.

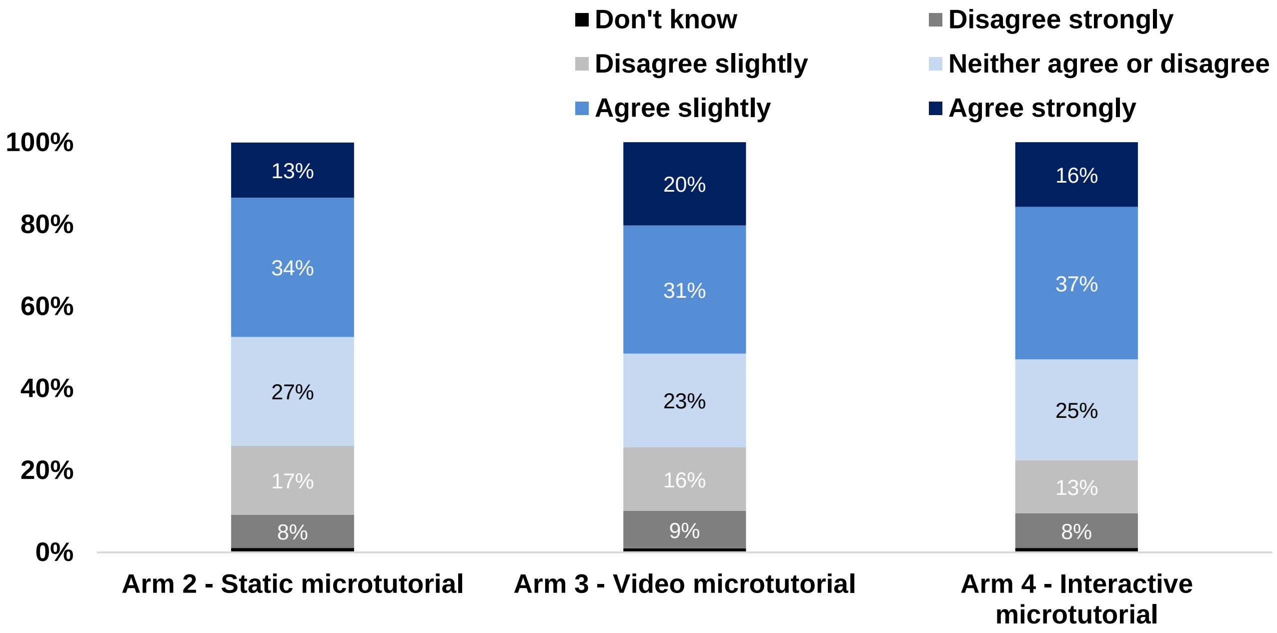


Figure 35. Participant responses to the prompt ‘I learned something new by going through the introduction’, by arm

7.5.5.8. Design quality of introduction

Participants who watched the video microtutorial were significantly less likely to report that the introduction was poorly designed compared to participants who completed the static microtutorial (Table 47). There were no observed differences between participants who completed the interactive and static microtutorial.

Table 47. Model-based estimates of design ratings of introduction

	Odds Ratios	95% CI	*z*-value	*P*
Arm 3 – Video	0.74	0.61 – 0.89	-3.137	0.002
Arm 4 – Interactive	1.12	0.93 – 1.34	1.149	0.250

Note: Arm 2 – Static is the reference level other arms are compared against; excluding 20 “Don’t know” responses

After adjusting for multiple comparisons, we also found that participants who completed the interactive microtutorial were significantly more likely to indicate that the introduction was poorly compared to participants who completed the video microtutorial (Table 48). The previously reported difference between participants who completed the static and video microtutorial was not sensitive to adjustment for multiple comparisons.

Table 48. Estimates of design ratings of introduction (p values and CIs corrected for multiple comparisons using the Bonferroni correction)

Comparison	Odds Ratios	95% CI	*z*–value	*P*
Arm 2 – Static vs Arm 3 – Video	0.74	0.59 – 0.93	-3.137	0.005
Arm 2 – Static vs Arm 4 – Interactive	1.12	0.89 – 1.39	1.149	0.751
Arm 3 – Video vs Arm 4 – Interactive	0.66	0.53 – 0.83	-4.318	< 0.001

Figure 36 shows the distribution of participant responses to the prompt ‘I thought the introduction was poorly designed’.

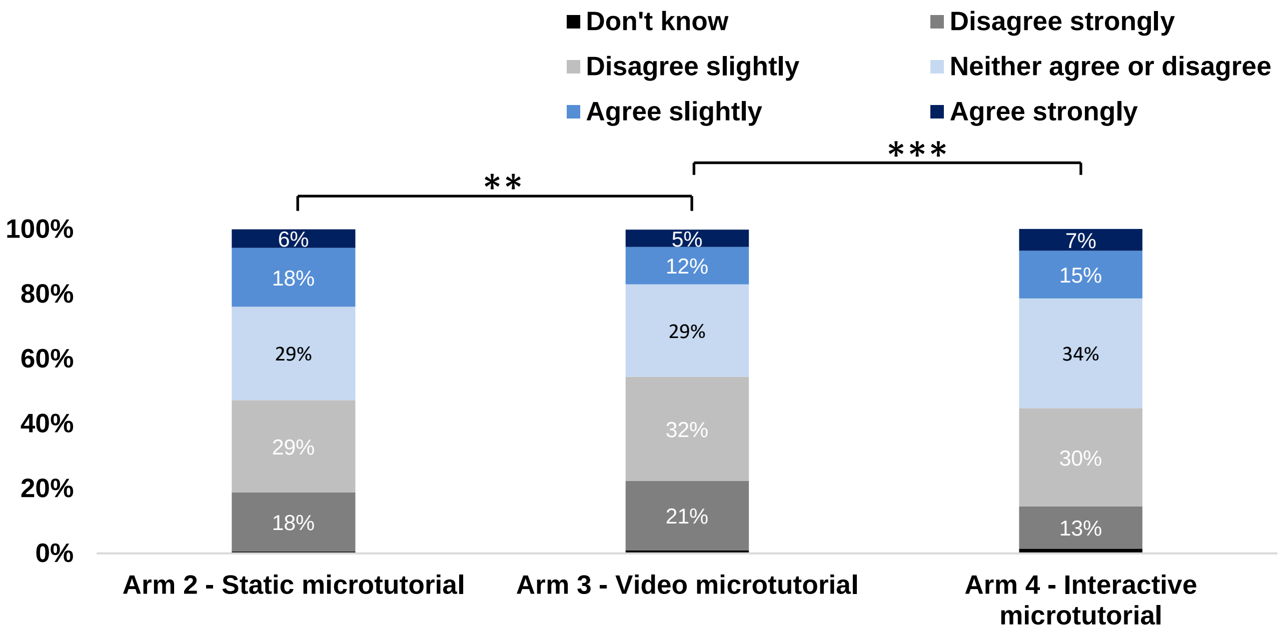


Figure 36. Participant responses to the prompt ‘I thought the introduction was poorly designed’, by arm

7.6. Descriptive statistics

We also conducted exploratory analyses relating to how the primary outcome measure (1.1 in Table 2) or other secondary outcomes may vary across multiple sub–level groupings like age groups or self–reported gender to provide additional insights into participant behaviour. However, this research was not designed to explore potential demographic differences across participants. As a result, it is not known whether this study was sufficiently powered to detect any potential effects. As such, any conclusions drawn from this reporting should not be interpreted as representative of the population of panel members who are VSP users.

7.6.1. Reporting of potentially harmful videos by age and gender

Table 49 shows the observed probability of reporting potentially harmful videos by age. Participants in the age groups ’25–39’ and ’40–54’ had a slightly higher mean observed probability of reporting potentially harmful content than participants in other age groups, but these differences are marginal.

Table 49. Mean observed probability of reporting potentially harmful videos, by age group

Age group	Mean	SD
18–24	0.05	0.22

25–39	0.07	0.25
40–54	0.07	0.26
55–69	0.05	0.22

Table 50 shows the observed probability of reporting of potentially harmful videos by gender. Female participants had a lower mean observed probability of reporting potentially harmful videos (0.06) than male participants (0.07), but the difference is marginal.

Table 50. Mean observed probability of reporting potentially harmful videos, by gender

Gender	Mean	SD
Male	0.07	0.25
Female	0.06	0.24
Other	0.04	0.19
Prefer not to say	0.08	0.29

7.6.2. View time of potentially harmful videos

Table 51 shows the mean and median times (in seconds) participants spent watching potentially harmful videos, by age. Participants in the age group of 18-24 slightly less time viewing potentially harmful content than participants in any other group. There was very little difference in terms of mean viewing time between participants in the other three age brackets.

Table 51. Mean and median viewing time (in seconds) of potentially harmful videos, by age group

Age group	Mean	SD	Median
18–24	24.55	20.55	22.32
25–39	27.31	20.48	30.36
40–54	27.76	19.94	28.8
55–69	27.28	19.36	25.34

Table 52 shows the mean and median times (in seconds) participants spent watching potentially harmful videos, by gender. Male participants spent marginally longer watching potentially harmful content than female participants. Given the small number of participants identifying as other (18), or those who preferred not to disclose their gender (4), it is not possible to make any inferences into observed differences between these groups.

Table 52. Mean and median viewing time (in seconds) of potentially harmful videos, by gender

Gender	Mean	SD	Median
Male	27.75	20.10	30.24
Female	26.62	20.12	24.74
Other	17.95	19.00	9.04
Prefer not to say	12.12	16.08	6.01

8. Early discussion

All microtutorials were found to be effective at increasing the probability of reporting potentially harmful videos. The most effective microtutorial was the interactive microtutorial. The reported effects were robust to all sensitivity analyses and were replicated using a partial reporting outcome (pressing of a flag button).

Similarly, all microtutorials were found to be effective at increasing the probability of participants successfully completing a report of a potentially harmful video, when prompted to do so (i.e., during the reporting capability task). Once again, the most effective microtutorial was the interactive microtutorial.

Critically, there is some evidence to suggest that the interactive and video microtutorials were effective at teaching participants how to report videos as shown in the Reporting Capability Task. This is because not knowing how to report was the most frequently chosen reason for not reporting by participants who were not shown any microtutorial or who were shown the static microtutorial. In contrast, this option was selected by a much a smaller proportion of participants who were shown the video or interactive microtutorials and who chose not to report during the reporting capability task.

Despite the increase in the likelihood of participants completing a report on potentially harmful videos, there was no evidence that microtutorials had any effect on the reporting of neutral videos. This is a meaningful finding, because it suggests that microtutorials did not increase the probability of filing spurious reports.

Microtutorials increased overall interaction with all the video content. They statistically increased the probability of interacting with the potentially harmful videos, this was especially the case for video and interactive microtutorials. Dislikes constituted a substantial proportion of these interactions, and participants exposed to microtutorials were more likely to dislike potentially harmful videos than those not exposed to microtutorials.

There were no differences in the probability of skipping potentially harmful videos between participants exposed to microtutorials and those who were not.

There were no differences in the amount of time participants spent watching potentially harmful videos. Thus, the increase in rates of reporting potentially harmful videos did not seem to affect the length of time spent watching potentially harmful videos. One reason for this could be that microtutorials did not increase the likelihood of skipping potentially harmful videos.

Although all microtutorials were found to provide opportunities to learn how to use the VSP simulation, the ones that participants reported learning most from were the video and interactive microtutorials. Differences in annoyingness and perceived length between microtutorials did not seem to have a detrimental impact on the effectiveness of the interactive microtutorial.

Our research findings may show higher effect sizes than would be observed in the real world for two reasons. First, participants were forced to complete the microtutorials and did not have the option of skipping them. In addition, users of VSPs always have a choice to close their browsers, skipping any content they do not want to engage with. Second, we used a simulation of a VSP rather than working with an actual VSP to test the effectiveness of our interventions. Consequently, future research could examine whether the effects reported in this experiment are sensitive to making the microtutorials optional and whether the developed interventions work in real-life VSP environment.

9. Appendix A: Post-trial questionnaire

Trial Arm (GROUP)	No. of questions	Format of questions
Arm 1 -- Control	4	Likert scale
Arm 2 -- Static microtutorial	12	Likert scale
Arm 3 -- Video microtutorial	12	Likert scale
Arm 4 -- Interactive microtutorial	12	Likert scale

Part 1: Unprompted recall

CONFIDENT

ASK ALL

SINGLE CODE

Thinking about your experiences of using video sharing platforms outside of this survey, to what extent do you agree with the following statement:

I feel confident using features of video sharing platforms

1	Disagree strongly
2	Disagree slightly
3	Neither agree nor disagree
4	Agree slightly
5	Agree strongly
6	Don’t know

Thinking about your experiences just now, using the video sharing platform in this survey, to what extent to you agree with the following statements:

EASY

ASK ALL

SINGLE CODE

I found the features of the platform (e.g., liking, disliking, reporting) easy to use

1	Disagree strongly
2	Disagree slightly
3	Neither agree nor disagree
4	Agree slightly
5	Agree strongly
6	Don’t know

OPPORTUNITY

ASK ALL

SINGLE CODE

There were opportunities to learn how to use the video sharing platform

1	Disagree strongly
2	Disagree slightly
3	Neither agree nor disagree
4	Agree slightly
5	Agree strongly
6	Don’t know

DESIGN

ASK ALL

SINGLE CODE

The design of the video sharing platform provided opportunities for me to report videos

1	Disagree strongly
2	Disagree slightly
3	Neither agree nor disagree
4	Agree slightly
5	Agree strongly
6	Don’t know

SCRIPTER NOTES: RANDOMISE THE ORDER OF EASY, OPPOURTUNITY, AND DESIGN.

Part 2: Prompted recall

SCRIPTER NOTES: REMAINING QUESTIONS ARE ONLY ASKED TO ARM 2, 3, AND 4.

Thinking about the introduction you were given to the video sharing platform in this survey, before you started interacting with it, to what extent to you agree with the following statements:

SCRIPTER NOTES**: [INSERT SCREENSHOT FROM THEIR EXPERIMENTAL ARM]**

REPORT_CONFIDENCE

ASK IF GROUP = 2, 3, 4

SINGLE CODE

The introduction gave me confidence to report videos

1	Disagree strongly
2	Disagree slightly
3	Neither agree nor disagree
4	Agree slightly
5	Agree strongly
6	Don’t know

NEED

ASK IF GROUP = 2, 3, 4

SINGLE CODE

I did not need the introduction

1	Disagree strongly
2	Disagree slightly
3	Neither agree nor disagree
4	Agree slightly
5	Agree strongly
6	Don’t know

ANNOYING

ASK IF GROUP = 2, 3, 4

SINGLE CODE

I found the introduction annoying

1	Disagree strongly
2	Disagree slightly

3 Neither agree nor disagree

4 Agree slightly

5 Agree strongly

6 Don’t know

SIMPLE

ASK IF GROUP = 2, 3, 4

SINGLE CODE

I found the introduction too simple

1 Disagree strongly

2 Disagree slightly

3 Neither agree nor disagree

4 Agree slightly

5 Agree strongly

6 Don’t know

LONG

ASK IF GROUP = 2, 3, 4

SINGLE CODE

The introduction was too long

1 Disagree strongly

2 Disagree slightly

3 Neither agree nor disagree

4 Agree slightly

5 Agree strongly

6 Don’t know

ENGAGING

ASK IF GROUP = 2, 3, 4

SINGLE CODE

I found the introduction engaging

1 Disagree strongly

2 Disagree slightly

3 Neither agree nor disagree

4 Agree slightly

5 Agree strongly

6 Don’t know

LEARN

ASK IF GROUP = 2, 3, 4

SINGLE CODE

I learned something new by going through the introduction

1 Disagree strongly

2 Disagree slightly

3 Neither agree nor disagree

4 Agree slightly

5 Agree strongly

6 Don’t know

POOR_DESIGN

ASK IF GROUP = 2, 3, 4

SINGLE CODE

I thought the introduction was poorly designed

1 Disagree strongly

2 Disagree slightly

3 Neither agree nor disagree

4 Agree slightly

5 Agree strongly

6 Don’t know

SCRIPTER NOTES: RANDOMISE THE ORDER OF ALL QUESTIONS IN SECTION 2.

10. Appendix B: Demographic breakdown of participants by device type

Appendix B Item 1: Split of participants by device operating system and age group

Device operating system	18–24 (%)	25–39 (%)	40–54 (%)	55–69 (%)
Android	23.5	41.9	44.3	35.0
Ios	62.5	31.3	19.0	8.4
Windows	8.3	20.3	28.0	40.1
macOS	4.3	4.7	4.4	9.9
iPadOS	0.3	0.2	0.8	1.1
ChromeOS	1.3	1.2	2.0	2.2
Linux	0.0	0.3	0.8	1.1
Unknown	0.0	0.0	0.0	0.0

Appendix B Item 2: Split of participants by device operating system and gender

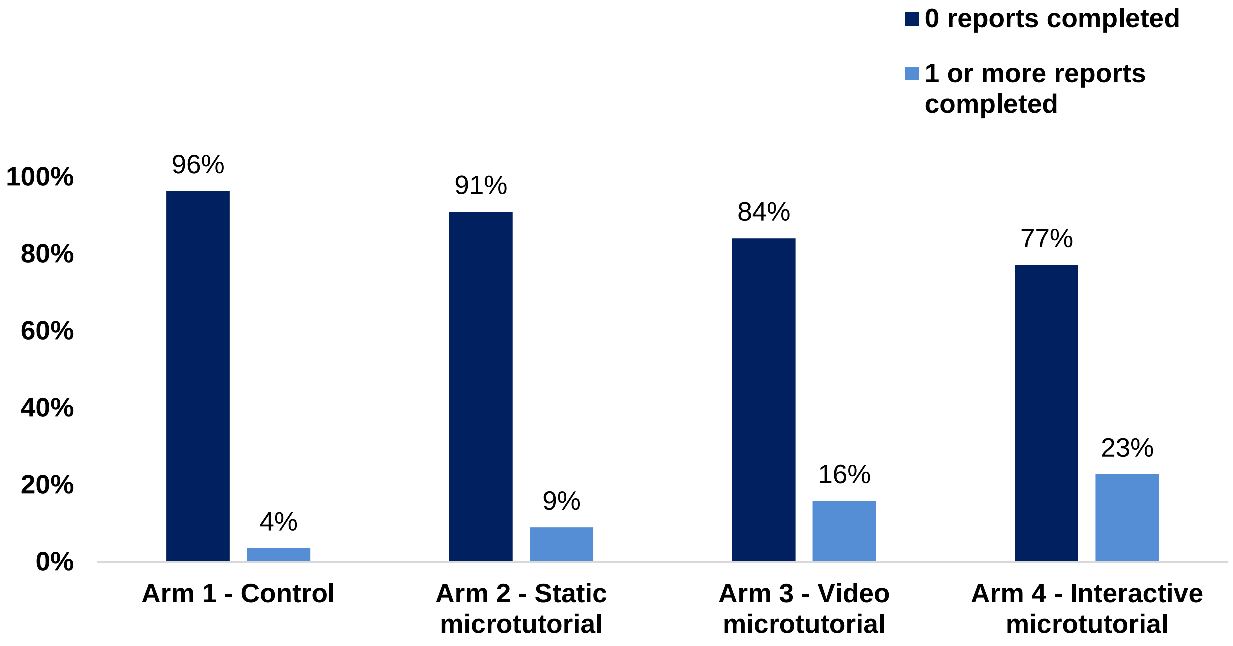
Device operating system	Male (%)	Female (%)	Other (%)	Prefer not to say (%)
Android	37.2	39.8	38.9	50.0
Ios	22.3	31.5	22.2	50.0
Windows	32.3	18.4	27.8	0.0
macOS	4.8	6.5	11.1	0.0
iPadOS	0.8	1.8	0.0	0.0
ChromeOS	1.7	1.7	0.0	0.0
Linux	0.9	0.3	0.0	0.0
Unknown	0.0	0.0	0.0	0.0

Appendix B Item 3: Split of participants by device operating system and SEG

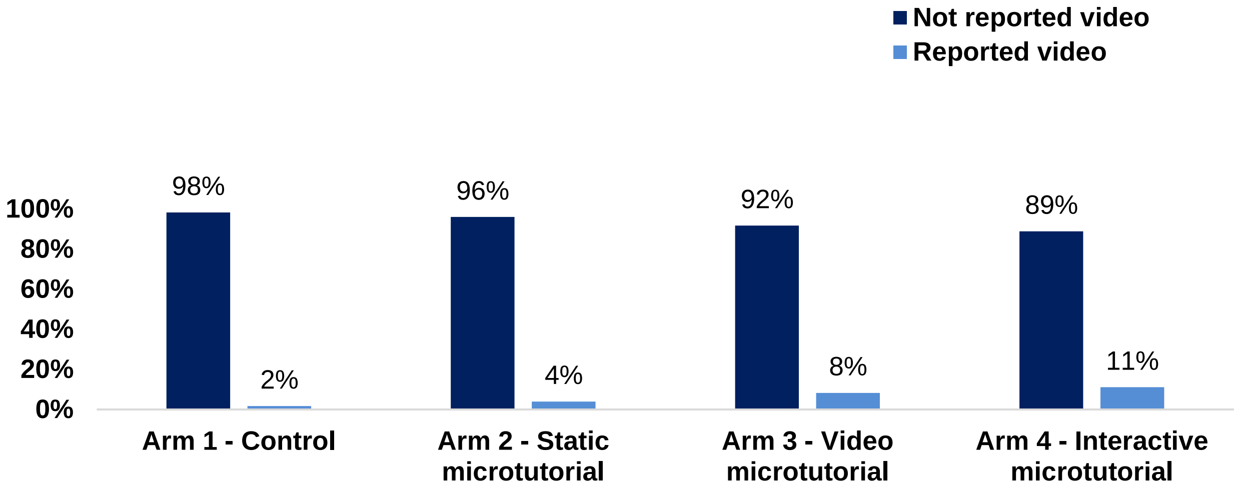
Device operating system	ABC1 (%)	C2DE (%)
Android	34.3	43.9
Ios	27.4	26.4
Windows	28.2	21.6
macOS	6.4	4.8
iPadOS	1.6	0.9
ChromeOS	1.7	1.6
Linux	0.4	0.8
Unknown	0.0	0.0

11. Appendix C: Additional descriptive statistics

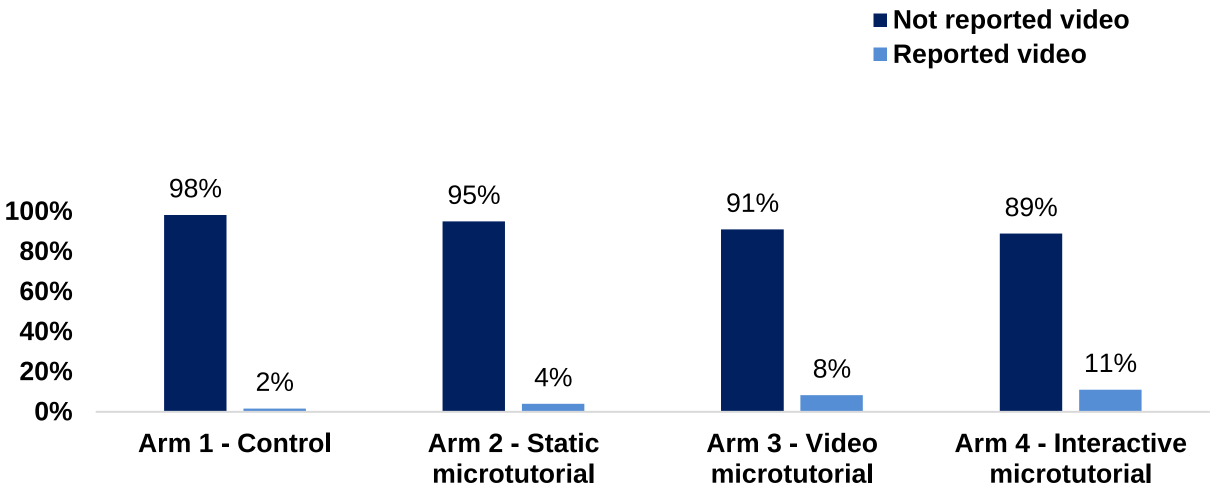
Appendix C Item 1: Percentage of participants who reported at least one potentially harmful video across treatment arms



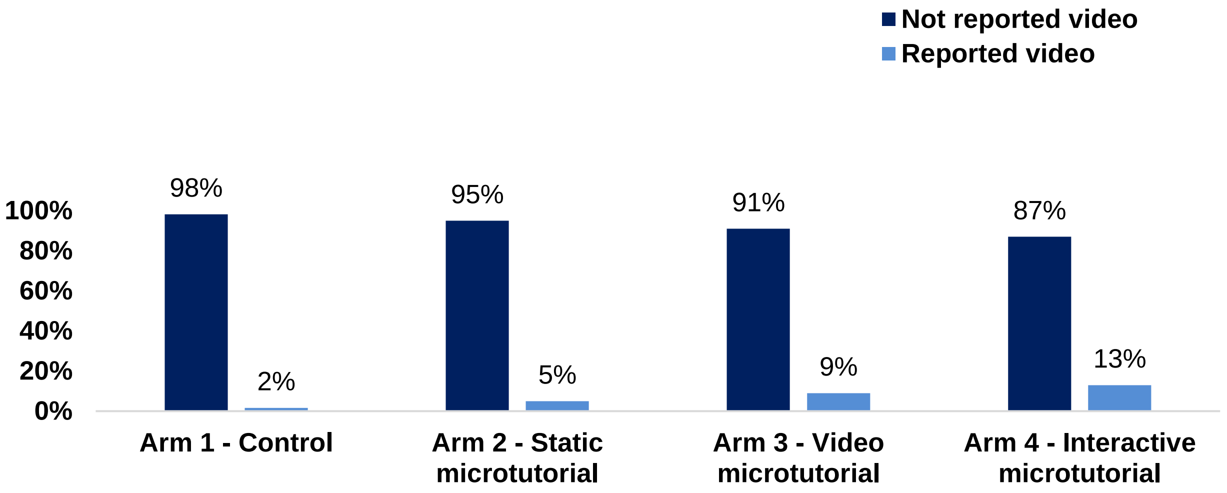
Appendix C Item 2: Percentage of observations in which a potentially harmful video was reported, per arm (Primary analysis dataset)



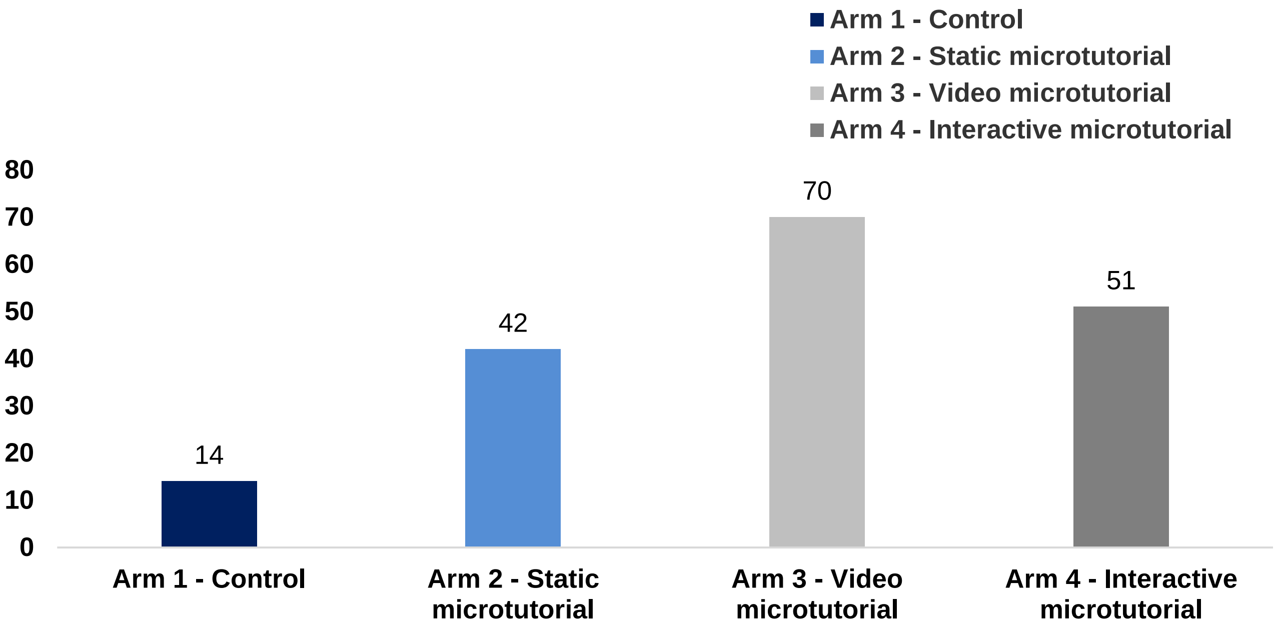
Appendix C Item 3: Percentage of observations in which a potentially harmful video was reported, per arm (Sensitivity analysis dataset including only participants who recorded a non-zero watch time for at least one video.)



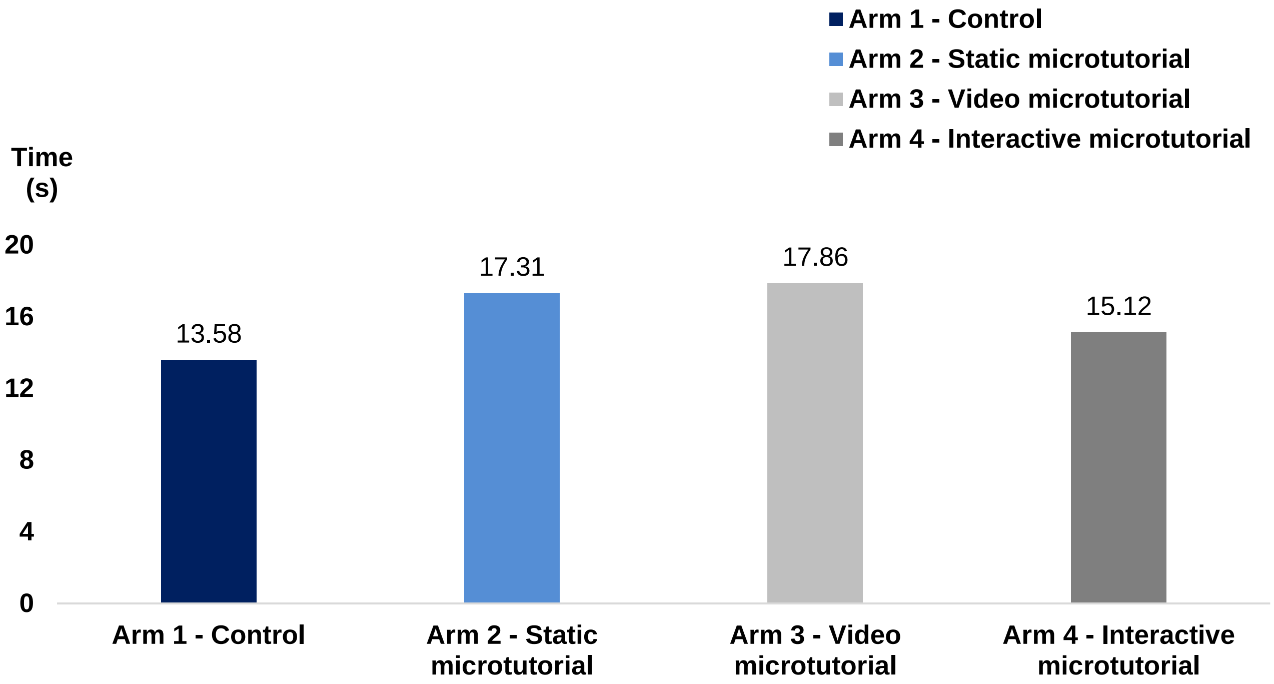
Appendix C Item 4: Percentage of observations in which a potentially harmful video was reported, per arm (Sensitivity analysis dataset including only participants who recorded a non-zero watch time for all videos.)



Appendix C Item 5: Raw count of 'incomplete reports'[27] across treatment arms

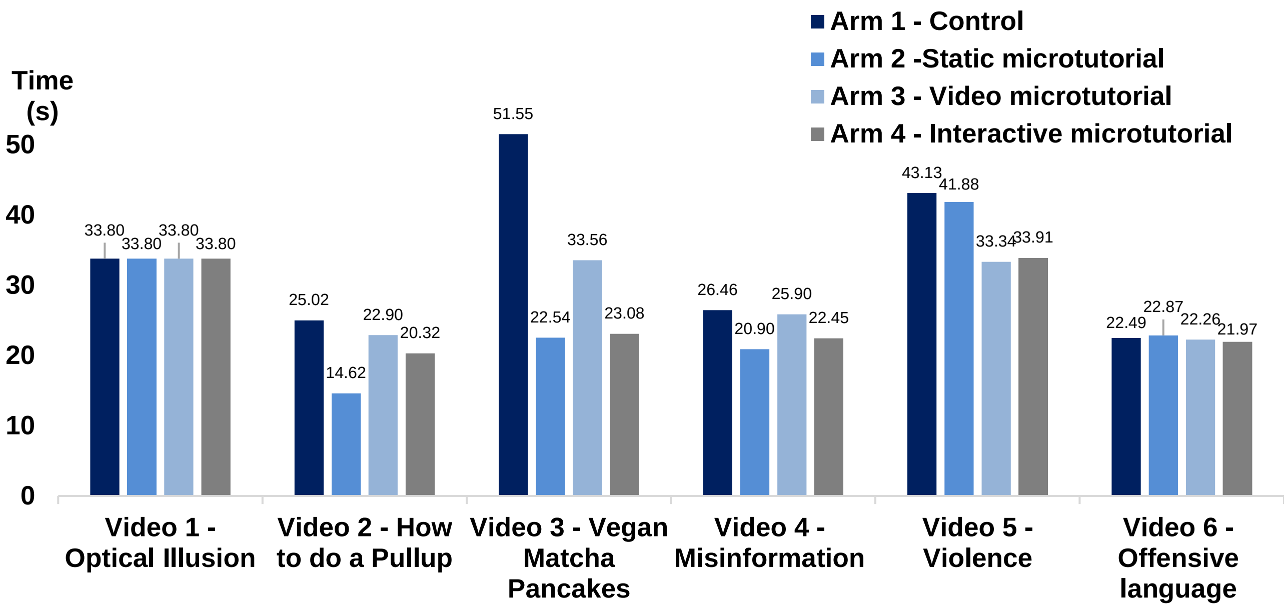


Appendix C Item 6: Mean time taken to complete a report across treatment arms

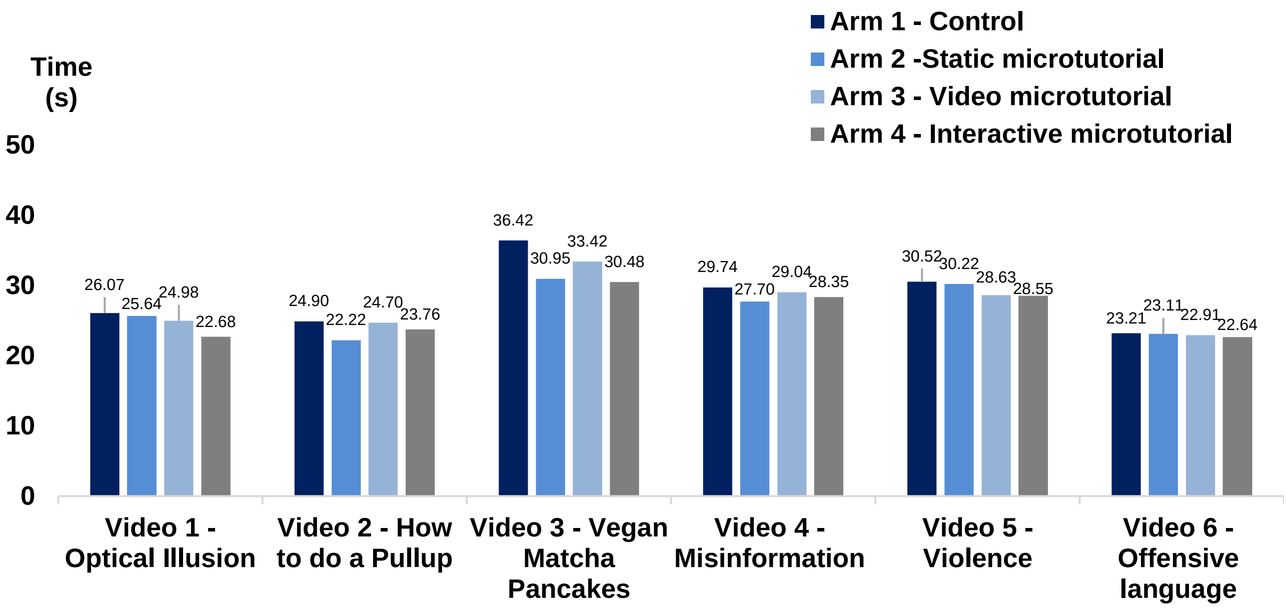


[27] An incomplete report is classified as any observation in which a user clicked the flag icon without submitting a report.

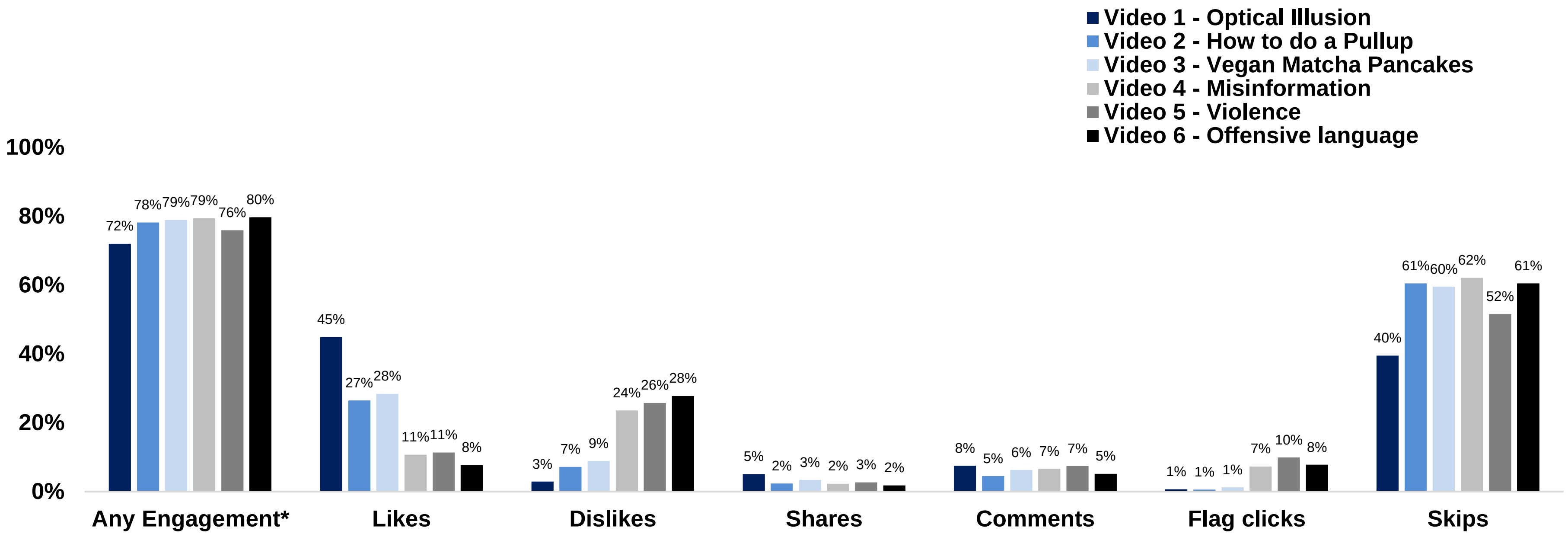
Appendix C Item 7: Average (Median) viewing time of each video split by arm



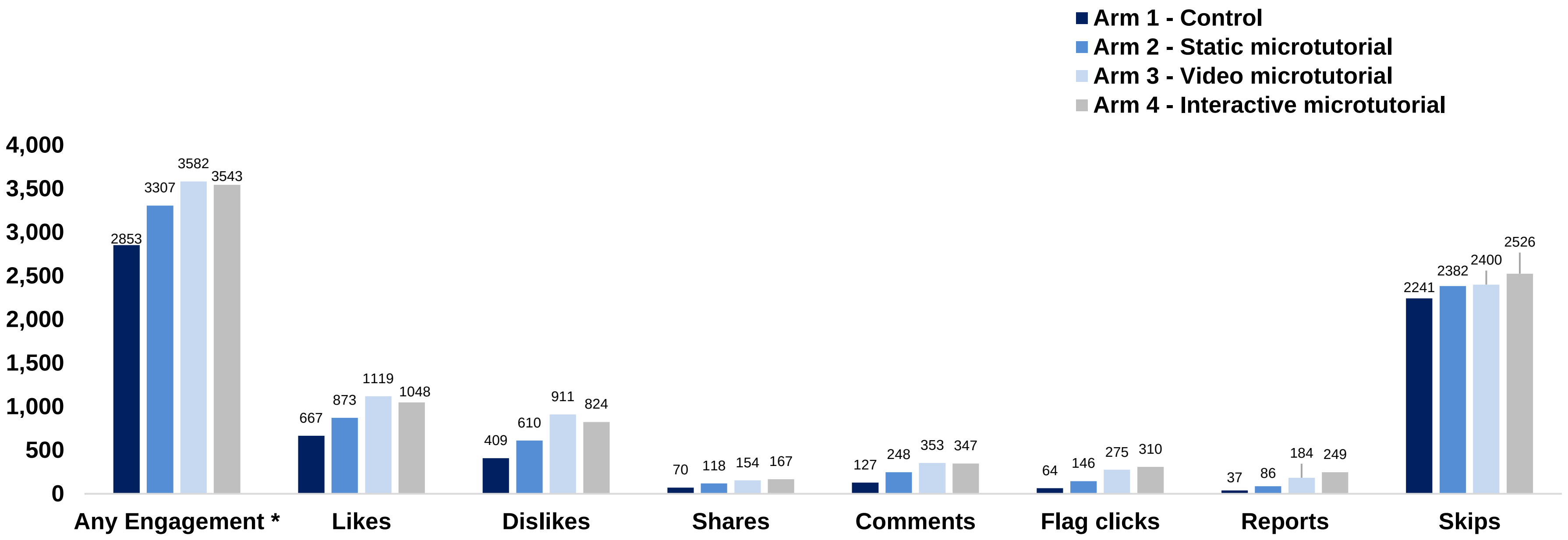
Appendix C Item 8: Average (Mean) viewing time of each video split by arm



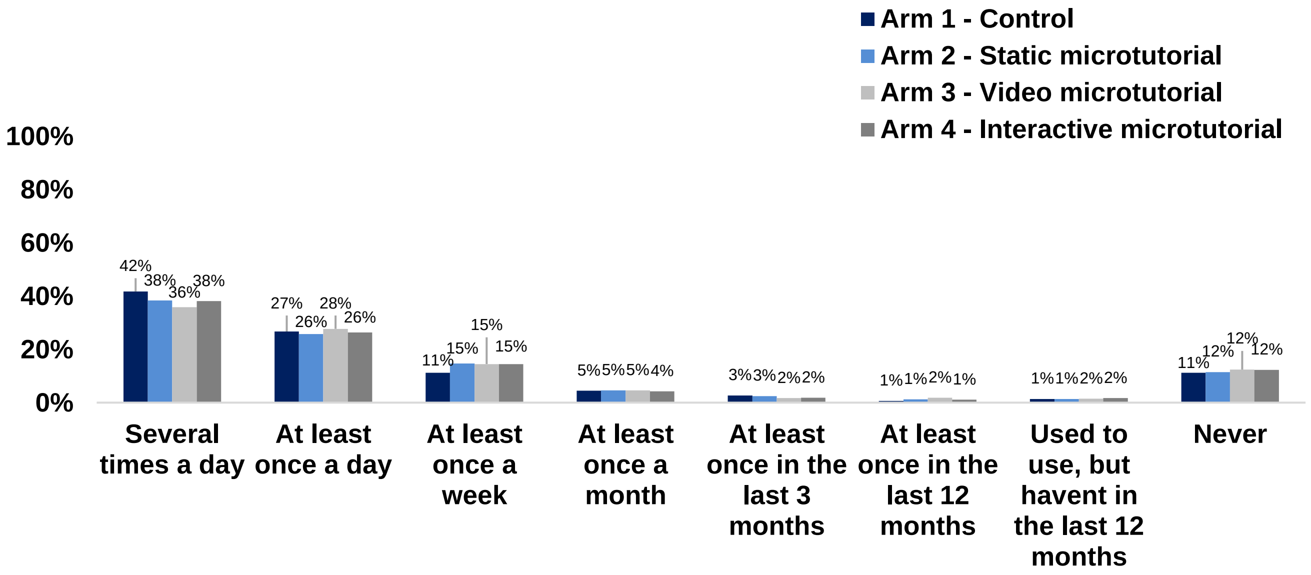
Appendix C Item 9: Participant engagement per video



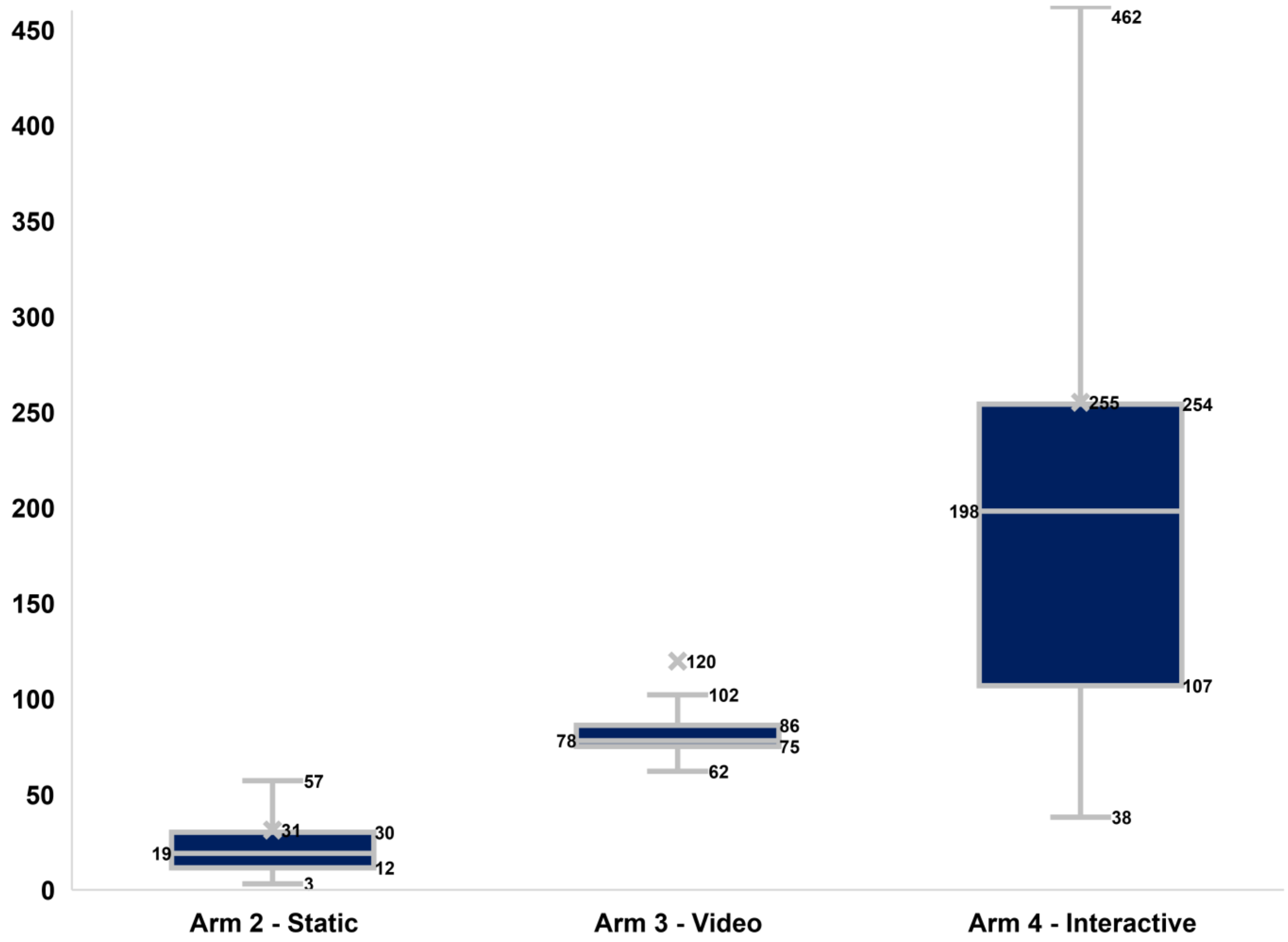
Appendix C Item 10: Participant engagement per arm (raw count)



Appendix C Item 11: VSP use across trial arms



Appendix C Item 12: Box plots of time spent completing microtutorials, per arm



Appendix C Item 13: Time spent completing microtutorials, per arm

Microtutorial	Min	Max	Median	Mean	SD
Arm 2 – Static	3	2233	19	31.18	96.43
Arm 3 - Video	10	6027	78	119.50	344.14
Arm 4 – interactive	38	23827	198	254.97	928.36