

© 2020, American Psychological Association. This paper is not the copy of record and may not exactly replicate the final, authoritative version of the article. Please do not copy or cite without authors' permission. The final article will be available, upon publication, via its DOI: [10.1037/xge0001023](https://doi.org/10.1037/xge0001023)

Can You Trust What You Hear? Concurrent Misinformation Affects Recall Memory and Judgements of Guilt.

Greg J. Neil¹, Philip A. Higham², and Simon Fox³


¹Department of Psychology, Solent University


²Department of Psychology, University of Southampton

³Department of Criminology, Solent University

Word count: 14,060

Author Note

Greg J. Neil  <https://orcid.org/0000-0003-1360-5490>

Philip A. Higham  <https://orcid.org/0000-0001-6087-7224>

We have no known conflicts of interest to disclose.

This research was supported by a RIKE bid from Solent University to the authors.

Correspondence concerning this article should be addressed to Greg J Neil, Psychology, Solent University, East Park Terrace, Southampton, SO14 0YN, UK. Tel: 02382 016751; Email: greg.neil@solent.ac.uk

Portions of this research were presented at the Solent Research and Innovation Conference, 2017, Southampton, UK and the 3rd International Meeting of the Psychonomic Society, 2018, Amsterdam, NL.

Abstract

In most misinformation studies, participants are exposed to a to-be-remembered event, and then subsequently given misinformation in textual form. This misinformation impacts on people's ability to accurately report the initial event. In this paper, we present two experiments that explored a different approach to presenting misinformation. In the context of a murder suspect, the to-be-remembered event was audio of a police interview, whilst the misinformation was co-presented as subtitles with some words being different to, and more incriminating than, those that were actually said. We refer to this as *concurrent misinformation*. In Experiment 1, concurrent misinformation was inappropriately reported in a cued-recall test, and inflated participants' ratings of how incriminating the audio was. Experiment 2 attempted to employ warnings to mitigate the influence of concurrent misinformation. Warnings after the to-be-remembered event had no effect, whilst warnings before the event reduced the effect of concurrent misinformation for a sub-set of participants. Participants that noticed the discrepancy between the audio and the sub-titles were also less likely to judge the audio as incriminating. These results were considered in relation to existing theories underlying the misinformation effect, as well as the implication for the use of audio and text in applied contexts.

Keywords: Misinformation, concurrent misinformation, audio, warnings, discrepancy detection, memory.

Can You Trust What You Hear? Concurrent Misinformation Affects Recall Memory and Judgements of Guilt.

Despite being the one who called 911, Kathy Carpenter was accused of Nancy Pfister's murder (Finley, 2015). Nancy was found dead, locked in her bedroom closet and covered in a sheet. One piece of the prosecution's evidence was that on the 911 call, Kathy had said "I saw blood on her forehead", which was the version of events reflected in the transcript submitted by the prosecution. This was presented as incriminating evidence, because Kathy claimed she had not looked under the sheet, so could only have known there was blood on Nancy's forehead if Kathy herself had committed the crime. However, when the defense re-examined the audio, they demonstrated that Kathy had actually said, "I saw blood on her headboard", which was not incriminating as Kathy would have seen the headboard in getting to the bedroom cupboard where the body was found. The call had been inaccurately transcribed, with both sentences being plausible given the facts of the case. Such mistranscriptions can potentially bias jurors as to the actual content of audio recordings, distorting any decisions a jury might make based on that information. The current paper investigates the possible effect of such mistranscriptions, through the lens of the misinformation paradigm.

The Traditional Misinformation Effect

The misinformation effect has proven to be robust and repeatable (Loftus, 2005). In traditional misinformation experiments, participants are presented with an initial event, subsequently given some inaccurate information (the *misinformation*), and finally given a memory test for the initial event. The misinformation effect occurs when participants given misinformation perform worse on the memory test than do participants not given misinformation, either by decreased reports of actual event details, increased reports of misinformation, or both. For example, Loftus (1974) showed participants a video of a car

accident. Misinformation was induced by asking participants a question involving a barn that was not in the video. On a later memory test, misinformed participants reported seeing the non-existent barn to a greater extent than did control participants, given no misinformation. Loftus and Palmer (1974) further demonstrated that a similar effect can be produced by simply phrasing questions in different ways. Having viewed a video of a car crash, participants were asked how fast the cars were going when they *contacted* each other. If instead of *contacted* a verb was used that implied greater speed prior to the crash, such as *collided* or *hit*, participants' estimates of how fast the cars were moving increased.

Whilst Loftus and colleagues (1974) induced misinformation through misleading questions, robust misinformation effects can be found when misinformation is introduced by a variety of means. Examples include using (1) narratives (LaPaglia & Chan, 2013), (2) audio tapes (Thomas et al., 2010), (3) social pressure or co-witnesses (Goodwin et al., 2013), (4) doctored video evidence (Nash & Wade, 2009) and even (5) gestures (Gurney et al., 2016; Kirk et al., 2015). Although the misinformation effect occurs when introduced through any of those methods, its magnitude can vary. Misinformation tends to be more effective if (1) there is a long delay between the misinformation and the memory test (Higham, 1998), (2) sleep occurs before the misinformation is presented (Calvillo et al., 2016), (3) when people view misinformation as corrective feedback (Rindal et al., 2016), (4) the misinformation is presented through questions rather than narratives (Zaragoza & Lane, 1994), (5) memory for the original event is poor (Holliday et al., 1999), and (6) the misinformation is about peripheral rather than central details (Dalton & Daneman, 2006; Daneman et al., 2013).

The Concurrent Misinformation Effect

Most of the misinformation studies mentioned above presented the misinformation *after* the initial event.¹ This is generally considered to simulate the effects of follow up questions by the police (Clifford & Scott, 1978), of co-witnesses discussing what they thought they saw (Skagerberg & Wright, 2008), or of any other event that might influence memory of an event after the fact. However, there are other ways that misinformation might be introduced. Consider a trial in which the vital evidence is a recorded conversation, such as in the case of Kathy Carpenter mentioned earlier. The audio quality of such evidence may be quite poor, perhaps because of the quality of the microphones used, or background noise. With juries, poor audio quality is usually compensated for by providing transcripts or subtitles (if watching a video recording) to the jurors. If transcripts or subtitles do not match what is actually being said, then by providing transcripts/subtitles, jurors have been exposed to misinformation at the same time as (or sometimes before) their exposure to the original event (i.e., the original audio). As this misinformation would be introduced at the stage of perceiving the actual event, rather than later, we will refer to this as the *concurrent misinformation effect*. This type of misleading influence has received very little research attention, much less than the attention dedicated to the traditional misinformation effect.

The concurrent misinformation effect is important in a forensic setting because it could influence how juries interpret and remember evidence. Where audio is presented as evidence, juries must first evaluate the audio. Later, they must not only remember the evidence, they must also evaluate the reliability of the evidence, and judge whether the audio should be taken to be incriminating or exonerating. Ultimately, these evaluations must inform inferential judgements about the defendant, the most important being a guilty or not guilty judgment. Concurrent misinformation in the form of text potentially distorts the memory of

¹ There is a literature on misinformation in which the misinformation is itself the initial, and only, information presented (e.g., Lewandowsky et al., 2012), but this is not discussed here as we are primarily concerned with misinformation that alters memory or perception of another experienced event.

the audio evidence, leading to later inferential judgments, like guilt, potentially being biased or distorted. Clearly, this outcome is undesirable given the stakes of criminal trials. Despite this, whether concurrent misinformation would actually distort memory, and whether that could lead to distorted inferential decisions, is relatively under-researched.

The concurrent misinformation effect is only worth researching if transcription errors are likely to occur in practice. In this vein, Coulthard et al. (2016) set out several examples of cases in which police interviews had been inaccurately transcribed. For instance, a murder suspect who had boarded a train was transcribed as saying he “shot a man to kill” when in fact he said “show[ed] a man ticket”. Coulthard et al. suggested these errors usually occur because people transcribe what they expect to hear rather than the words spoken.

Additionally, Haworth (2018) pointed out that the way audio evidence is represented in court is far removed from the original audio, so errors are inevitable. Factual distortions can also occur in transcripts, misrepresenting the facts of a case. For example, Roberts and Lamb (1999) analyzed 68 investigative interviews with children and found that on 140 occasions, the investigators misheard the children, altering words that the children had said. Furthermore, when given the opportunity, the children only corrected these distortions one third of the time.

Examining this issue experimentally, Lange et al. (2010) demonstrated that transcription errors are easy to generate. In their first experiment, participants were given distorted audio to transcribe. Although the content of the audio was benign, some participants were told that the transcript was from a criminal trial, others were told that it was from job interviews, whilst the remaining participants were given no context. In the criminal context, people mistranscribed benign words (e.g., *muddy*) as incriminating words (e.g., *bloody*), thus creating incriminating transcripts. In the other contexts, participants created only benign transcripts. Lange et al. concluded that the trial context biased people to mishear words as

incriminating words, resulting in incriminating transcripts. In a second experiment, Lange et al. asked people to transcribe some audio, but first showed them an error-filled transcription produced by someone else. The new transcriber was biased to create a new transcription which included the same words as in the previously seen, error-filled, transcription, resulting in the new transcript containing the same errors as the previously seen transcript. Thus Lange et al.'s research shows that transcription errors can be induced through manipulating the context and being exposed to previous transcription errors, making it likely that transcription errors do occur in practice. There are many other possible avenues through which transcription errors might occur, such as in cases where interviewers mishear their interviewee, or mishear witnesses who are impaired.²

Lange et al.'s (2010) research elegantly demonstrated that transcribers can indeed be induced to make errors. However, they did not address whether those textual transcription errors bias judges' and/or jurors' memories and decisions. Indeed, whether such bias occurs is not clear from the rest of the literature either because most research investigating concurrent presentation of audio and text does not involve misinformation. Instead, the research has tended to explore how veridical memory is affected when both the audio and textual information are the same, rather than conflicting, a research domain we now briefly review.

Concurrent Presentation of Audio and Text

Some studies have found that the concurrent presentation of matching audio and text has a facilitating effect on learning. For example, Brasel and Gips (2014) found that concurrent subtitles and audio, compared to audio alone, improved memory for the content of television adverts. Dowell and Shmueli (2008) also found that concurrent text and audio aided memory, but only in comparison to audio only and not text alone. This research is

² We thank an anonymous reviewer for additional ideas of where our research is relevant.

consistent with *dual-coding theory* (Pavio, 1991), which posits that memory is better if information is encoded via two routes (e.g., visual and auditory) rather than just one (e.g., auditory alone; see Kanellopoulou et al., 2019 for a recent application of dual-coding theory to the use of subtitles).

Contrary to studies showing a positive effect on memory of concurrent text and audio, other studies have found that concurrent audio and text have a detrimental effect on memory. For example, Moreno and Mayer (2002a) found that for a highly complex learning task about botany, concurrent audio and text produced poorer memory than an audio-only condition. Kalyuga et al. (2004) similarly found that an audio-only learning condition produced better memory for mechanical engineering texts than concurrently presented text and audio.

There have been some attempts to explain the above noted inconsistencies in the effects of concurrently presented audio and text. Leahy and Sweller (2016) suggested that the effects of concurrent text and audio change depending on the presence of complex visuals or text length. Alternatively, Jamet and Le Bohec (2007) suggested that the degree of working memory overload was important. Different materials produce varying amounts of overload, with less versus more overload being associated with facilitative versus detrimental effects of concurrent text and audio, respectively.

Despite the above noted explanations for the equivocal effects of consistent concurrent text and audio on memory, it is not clear from this literature what would happen when concurrent text and audio present conflicting information. In terms of memory, textual misinformation seems likely to impair memory for the veridical audio because of the high potential for working memory overload, and the potential for the misinformation to otherwise replace or distort memory for the audio. On the other hand, if the conflicting details are clear and obvious, they may serve to focus attention specifically on those parts of the audio where the inconsistencies occur, making it highly memorable. Separate from the effects on memory,

it is also not clear from this literature whether inconsistent text and audio would influence later inferential judgements, such as about the guilt of a suspect. Thus, the first goal of the experiments presented below was to bring clarity to this issue and investigate how concurrent misinformation affects both memory and inferential judgments.

Preventing the Concurrent Misinformation Effect

The second goal of the experiments was to investigate whether the concurrent misinformation effect, if established, can be mitigated or prevented. Previous research has found that if participants detected discrepancies between the original event and the post-event misinformation in the traditional misinformation paradigm, memory distortions were substantially reduced. For example, Tousignant et al. (1986) demonstrated that participants who read a misinformation narrative slowly detected more discrepancies between the misinformation and the original event compared to those who read quickly. This discrepancy detection made participants more resistant to misinformation. Similarly, Higham, Blank, et al. (2017) required participants to indicate, while taking a 2AFC recognition memory test, which test questions were associated with a discrepancy between the event and the post-event detail. Memory performance was twice as accurate on items for which participants detected a discrepancy compared to items for which no discrepancy was detected. Other studies have also found that discrepancy detection can reduce the effect of misinformation (e.g., Blank, 1998; Oeberst & Blank, 2012; Putnam et al., 2017; Schooler & Loftus, 1986).

Warnings may also reduce the effect of concurrent misinformation. The traditional misinformation effect has been reduced by warnings given before (pre-warnings) or after (post-warnings) the misinformation. For the traditional misinformation effect, pre-warnings are effective (Dodd & Bradshaw, 1980; Greene et al., 1982; Schul, 1993), although their effectiveness can drop off over time (Chambers & Zaragoza, 2001). The evidence on post-warnings is more varied, with most studies finding some effect of post-warnings (Chambers

& Zaragoza, 2001; Christiaansen & Ochalek, 1983; Dodson et al., 2015), but other studies finding no effect (Frost et al., 2002; Greene et al., 1982). Blank and Launay (2014) conducted a meta-analysis of post-warnings' effectiveness, and concluded that in general, post-warnings do seem to be successful at reducing the misinformation effect. However, the effectiveness of the post-warnings varied, with effective warnings involving some element of explanation about why misinformation was present. Other evidence suggests that specificity can also be important, with weakly-worded or less-specific warnings being ineffective (Blank et al., 2013), whilst specific warnings (that inform participants exactly which items on the memory test pertain to misinformation) being very effective, conferring a substantial performance advantage compared to more general warning (Higham, Blank, et al., 2017).

Given the above literature, a reasonable prediction for concurrent misinformation is that both warnings and discrepancy detection will mitigate any effects the misinformation may have on memory or inferential judgements. However, there is evidence to suggest that the effects of concurrent misinformation may be harder to mitigate compared to the traditional misinformation. For example, processing co-presented text and audio is a cognitively demanding task (Brunyé et al., 2006; Jamet & Le Bohec, 2007), which may leave few cognitive resources available to implement instructions in warnings, or to notice discrepancies. Also, text, such as subtitles, is difficult to ignore (Brasel & Gips, 2014; d'Ydewalle et al., 1991), making the misinformation likely to be attended to rather than ignored, even with warnings or discrepancy detection. Finally, with the traditional misinformation effect, increasing contextual overlap between the original event and post-event also increases the misinformation effect (Johnson et al., 1988; Johnson et al., 1993; Pezdek & Greene, 1993). Contextual overlap is high when it is difficult to distinguish the misinformation and original event in memory, for example, if the misinformation is presented in the same form as the original event. Lindsay (1990) manipulated contextual overlap

between the to-be-remembered event and the misinformation by changing the time between encoding of the to-be-remembered event and the misinformation from 48 hours to immediately afterwards. The smaller the interval, the larger the contextual overlap, with larger misinformation effects being associated with greater contextual overlap. The concurrent misinformation paradigm goes one step further than Lindsay's experiment because the event and misinformation are presented at the same time, likely resulting in a robust concurrent misinformation effect. Given the above noted difficulties introduced by concurrent presentation of information and misinformation, a test of traditional misinformation prevention techniques on concurrent misinformation seems warranted, forming the second goal of our experiments.

Experiment Summary

In two experiments we investigated the effects of concurrent misinformation by combining the traditional misinformation task with the misleading transcript approach used by Lange et al. (2010). In the basic task, participants watched a video of a mock police interview, in which a suspect gave a description of events which was relatively non-incriminating, at least with respect to the criminal charges under consideration (murder, sexual assault). However, the audio for these videos was distorted such that it was difficult to understand. Each video was presented either with or without subtitles across three experimental groups. In the no-subtitles group, the video (with audio) was presented on its own with no subtitles. In the accurate-subtitles group, subtitles were provided which accurately represented the audio. In the incriminating-subtitles group, subtitles were also provided, but for five key sentences, the subtitles were changed to disclose something incriminating. The incriminating subtitle sentences were deliberately designed to resemble the audio. For example, instead of *he was muddy*, which was stated in the audio, the subtitles read *he was bloody*. Later, participants were given a cued-recall task in which they were

asked to recall these five keywords, to rate how incriminating the interview was (all experiments), and to decide whether or not the suspect was guilty (Experiment 2).

Based on the standard results in misinformation research, our predictions were that the incriminating subtitles would produce a concurrent misinformation effect. As a result, compared to participants in the accurate- and no-subtitles groups, participants in the incriminating-subtitles group would remember more incriminating words, give higher judgments of how incriminating the audio was, and judge the suspect as guilty more often. In the accurate- and incriminating-subtitle groups, a mild post-warning was administered, and the effectiveness of discrepancy detection was investigated. Experiment 2 further investigated whether warnings could be used to mitigate the concurrent misinformation effect, as well as testing whether the concurrent misinformation effect occurred with different materials and audio distortions.

Experiment 1

In Experiment 1 we tested the basic prediction that the presence of misleading subtitles in video-recorded mock police interviews would affect both memory and inferential judgments regarding that interview. The main manipulation was how incriminating the subtitles were, as described earlier. As Experiment 1 was our first investigation into this phenomenon, we deliberately maximized the likelihood that a concurrent misinformation effect would occur. First, we used subtitles to present the text rather than a transcript (Lange et al., 2010) because we reasoned that participants would be less likely to strategically ignore the former (e.g., by refusing to read the transcript). Second, participants were given no specific pre-warning that the subtitles were misleading. Thus, participants were likely to implicitly trust the veracity of the subtitles (Blank, 1998). As noted above, our main prediction was that with incriminating subtitles, participants would report higher ratings of incrimination regarding the suspect, and falsely recall more incriminating words as having

been stated in the audio, compared to accurate and absent subtitles. Finally, participants were asked at the end of the experiment to indicate if they noticed discrepancies between the subtitles and the audio and to indicate which questions on the memory test these discrepancies pertained to (Higham, Blank, et al., 2017).

A secondary matter in Experiment 1 concerned the extent to which participants might be likely to use other, secondary, features of the video in deciding whether the video evidence should be trusted. With that in mind, the visual quality of the videos was manipulated to be either good or poor. This created a situation where, if participants realized that the subtitles spoiled their ability to judge the clarity of the audio, then they may use other aspects of the overall video quality to act as a proxy for judging the audio quality. Consequently, they may use video quality to help judge how much they should weight the audio evidence. If the video quality is high, then participants may assume the audio quality is also high, and their knowledge of what has been said can be safely trusted and weighted strongly. On the other hand, if the video is poor, participants may assume the audio is poor and so their knowledge of what was said should be discounted. This may occur even when the knowledge they have regarding what was said derived from the subtitles rather than the audio. Thus, if the subtitles are incriminating, our prediction is that the incrimination ratings will be higher with good video quality than with poor.

Method

Participants

Overall, 88 participants were tested who were compensated with course credit for their time. Four participants were excluded due to missing data, leaving a total of 84. These were equally split between subtitle groups, making 28 participants in each group. Age and gender were not recorded. Misinformation effect sizes tend to be large. For instance, Higham, Blank, et al. (2017) found an effect size $\eta_p^2 = 0.52$ in their general-warning group of

Experiment 1. For our design, power calculations in G*Power suggested that our sample size allowed us to detect an effect size $\eta_p^2 = 0.11$ with power = 0.81, more than adequate to detect an effect of the size reported in Higham, Blank, et al.

Design

The overall design was a 3 x 2 mixed design, with group (no-subtitles, accurate-subtitles, incriminating-subtitles) manipulated between-participants and video quality (good, poor) manipulated within-participants. The incriminating-subtitles group was analogous to the misinformation condition used in traditional misinformation designs, whereas the no-subtitles group served as the control. The dependent variables were incrimination ratings and cued-recall performance.

Materials and Procedure

Ethics approval for the experiment was obtained from the University of Southampton and Solent University Ethics committees. To allow a within-participant manipulation of video quality, two different videos were used. To make these videos, we created two scripts (see Appendix A) corresponding to two, different police interviews, one with a suspect for a murder, and one with a suspect of sexual assault. Following Lange et al. (2010), the scripts contained five critical statements which were non-incriminating given the crime, but which could be misheard as incriminating on the basis of a single word in the statement (e.g., *I went to grab him/I went to stab him*). These scripts were then used in the filming of the two mock police interviews. Volunteer actors were recruited to play the parts of the police interviewer and the two suspects, and a seminar room was arranged to look like a police interview room. To preserve a natural feel to the interviews, the actors were permitted to ad-lib most of their lines in the police interview barring the five critical statements which were pre-scripted. The videos were then recorded on a video camera fixed to the top corner of the interview room. For each video, both accurate and (inaccurate) incriminating subtitles were created. The

subtitles were rated prior to conducting the experiment for how incriminating they were on a five-point scale by a group of participants obtained through MTurk (42 rated the murder script; 35 rated the sexual assault script). These ratings confirmed that for both scripts, the incriminating scripts (murder $M = 3.41$, 95% CI [3.16, 3.66]; assault $M = 4.10$, 95% CI [3.80, 4.40]) were more incriminating than the accurate, non-incriminating, scripts (murder $M = 2.72$, 95% CI [2.47, 2.97]; assault $M = 2.81$, 95% CI [2.46, 3.17]), $t(40) = 2.50$, $p = 0.02$, $d = 0.74$; $t(33) = 3.54$, $p = 0.001$, $d = 1.17$, respectively. The difference for the murder script had a medium effect size, whilst the assault difference had a large effect size.

Six versions of each video were created, all of which were black-and-white. First, the videos for the poor visual quality condition were created by degrading the visuals with static. This gave the videos a grainy, slightly blurred appearance. Good quality videos were left unedited. Secondly, the audio was degraded for all videos by applying an audio static filter. The overall impression is of rain falling, leaving the words difficult but not impossible to understand. Finally, the videos for the subtitle groups were created by applying either accurate or incriminating subtitles with the caption function, with font size 20, on a black background, with the subtitles centered in the bottom of the video. Thus, the six versions of each video varied on visual quality (poor vs good) and subtitle type (no-subtitles vs accurate vs incriminating). Assignment of video type (murder vs sexual assault) to visual quality and subtitle groups was fully counterbalanced across all participants. The videos were embedded in PowerPoint presentations.

Participants were tested individually. After consenting, they were asked to sit down at a computer, put headphones on, and follow the instructions at the beginning of the PowerPoint presentation. In the first section, the two videos were presented one after another, each preceded by a preamble which set the context as an interview with a suspect of a crime

(see Appendix A). Order of crime type and the visual quality of the videos was counterbalanced across participants.

Immediately after having watched both videos, participants were presented with a post-event questionnaire. The first section of the questionnaire stated that each suspect had been arrested and was now on trial. For all groups that included subtitles (incriminating, accurate), it was also stated that, because of a legal technicality, the subtitles had to be removed from the videos when they were presented to a jury. Consequently, participants were instructed that all judgments should be made based on the *audio* only. First, they were asked to rate how incriminating the audio of each video was, on a 5-point scale (1 = not at all incriminating, 5 = extremely incriminating). Following the incrimination judgment, participants were given a seven-item cued-recall test, with five of the questions pertaining to the critical phrases and two being filler questions. The order of questions was the same for all participants, with the filler questions being the 1st and 5th questions in the set. Finally (and *after* all other answers had been given), participants were asked to indicate if they noticed any discrepancies between the subtitles and the audio, by going back and putting tick marks by any questions about the discrepant information (Higham, Blank, et al., 2017). This requirement, as well as the statement to ignore subtitles, was omitted for the no-subtitle group. Following the discrepancy detection task, participants were thanked and debriefed.

Results

In the analyses that follow, there was only one statistically significant effect of video quality, which was tangential to the rest of the findings (see later). Consequently, the tables for Experiment 1 were collapsed across video quality for the reader's convenience, except for Table 3 in which the analysis was constrained by video quality. Video quality has been retained in the analyses to account for any relevant variance attributable to it. The effect size

measure of η_p^2 is reported, with sizes of .01, 0.06 and 0.14 indicating small, medium and large effects sizes as recommended by Cohen (1969).

Incrimination Ratings

To investigate whether there was an effect of concurrent misinformation on inferential judgments, ratings of incrimination were entered into a 3 (group) x 2 (video quality) mixed-model ANOVA – see Table 1 for summary statistics. There was a main effect of group with a medium effect size, $F(1,81) = 7.59, p < .001, \eta_p^2 = 0.16$, no effect of video quality, $F(1,81) = 3.72, p = 0.06, \eta_p^2 = 0.04$, and no interaction, $F(1,81) = 1.60, p = 0.21, \eta_p^2 = 0.04$. The group main effect was investigated with Bonferroni corrected pairwise comparisons. The concurrent misinformation in the incriminating-subtitles group inflated ratings of incrimination compared to the no-subtitle group ($p = .001$). In other words, there was a concurrent misinformation effect on the incrimination ratings. Incrimination ratings did not differ between the no-subtitles group versus the accurate-subtitles group ($p = 0.39$), nor between the accurate-subtitles group versus the incriminating-subtitles group ($p = 0.06$).

Table 1

Mean Incrimination Rating by Video Quality and Group from Experiment 1 (95% Confidence Intervals in Brackets)

Group	Mean Incrimination Rating
No-subtitles	2.57 _a [2.33, 2.80]
Accurate-subtitles	2.87 _{ab} [2.60, 3.14]
Incriminating-subtitles	3.34 _b [3.05, 3.63]

Note. Means with different subscript letters within columns are significantly different from each other.

Words Recalled

To investigate whether there was a concurrent misinformation effect on recall, the proportion of words recalled was analyzed. Recall responses were coded into seven categories: (1) critical correct words (e.g., *grab*); (2) critical incriminating, incorrect words (e.g., *stab*); (3) non-critical, incriminating, incorrect words (e.g., *attack*); (4) non-critical, non-incriminating, incorrect words (e.g., *touch*); (5) critical, correct words, but the response was to the wrong question (e.g., *grab* as the answer to question 2 when it should be the answer to question 1); (6) critical, incriminating, incorrect words, but the response was to the wrong question (e.g., *stab* as the incorrect answer to question 2 when it was the incorrect answer to question 1); and (7) “don’t know” or omitted responses. Collapsing across group and video quality, categories 3-6 accounted for 13%, 3%, 2% and 3% of the responses respectively (total = 21%). Throughout both experiments, the analysis was restricted to categories 1, 2 and 7.³ Category 1 (critical correct words recalled) and category 2 (critical incriminating, incorrect, words recalled) will henceforth be referred to as correct and false recall⁴, respectively - see Table 2 for summary statistics for Experiment 1.

The first analysis indicated that correct recall was low but non-zero in all groups, as all lower bounds of the confidence intervals (seen in Table 2) were higher than zero. Next, correct recall was entered into a 3 (subtitle) x 2 (video quality) mixed ANOVA, which revealed only an effect of group, with a large effect size, $F(1, 81) = 113.40, p < .001, \eta_p^2 = 0.73$. There were no other effects, largest $F < 1$. Bonferroni corrected pairwise comparisons demonstrated that the main effect of group was due to more correct answers being reported in the accurate-subtitles group compared to both incriminating and no-subtitles groups, both ps

³ Other analyses are possible, such as including critical incriminating words given but to the wrong question or including any incriminating word. The results of these analyses are consistent with the analyses presented here.

⁴ Note that it could be argued that recalling a word from the subtitles is correct in so far as it did actually appear in the video. However, participants were explicitly asked to recall only words from the audio, and so recalling a word from the subtitles is false recall in the sense that participants were not supposed to be engaging in such recall.

< .001. There was no difference between the number of correct answers reported in the incriminating and no-subtitle groups, $p = 1.00$.

False recall was also entered into an analogous 3 x 2 ANOVA. Again, the only effect was of group, with a large effect size, $F(1, 81) = 89.71, p < .001, \eta_p^2 = 0.69$, with no other effects, largest $F < 1$. Bonferroni corrected pairwise comparisons indicated that participants recalled more false words in the incriminating-subtitles group than in the accurate and no-subtitles groups, both $p < .001$. There was no difference in false recall between accurate and no-subtitles groups, $p = 1.00$. Thus, the concurrent misinformation effect was reflected in inflated false memory for incriminating words for the incriminating-subtitles group.

Finally, the proportion of “don’t know” responses was analyzed. Analyzing these responses is important because they potentially provide some insight into the types of strategies participants adopt to regulate memory accuracy when faced with misinformation. An analogous 3 x 2 ANOVA on the proportion of “don’t know” responses indicated that there was only a main effect of group, with a large effect size, $F(1, 81) = 34.98, p < .001, \eta^2 = 0.46$, with no other effects, largest $F < 1$. Bonferroni corrected pairwise comparisons indicated that participants used “don’t know” less in both the accurate and incriminating subtitle groups, compared to the no-subtitle group, $p < .001$. There was no difference in the rate of “don’t know” responding between the accurate and incriminating subtitle groups, $p = 1.00$. Thus, participants made far more recall attempts in the groups with subtitles than in the one without subtitles, and they used “don’t know” approximately equally often in the accurate and incriminating subtitle groups.

Table 2

Proportion of Response Types and Discrepancies Detected by Group for Experiment 1 (95% Confidence Interval in Brackets)

Group	Response Type			
	Discrepancies	Correct	False Recall	Don't Know
	Detected	Recall		
No-subtitles	-	0.11 _a [0.07, 0.15]	0.04 _a [0.02, 0.06]	0.56 _a [0.46, 0.66]
Accurate-subtitles	0.20 _a [0.09, 0.30]	0.59 _b [0.02, 1.16]	0.06 _a [0.04, 0.08]	0.19 _b [0.13, 0.25]
Incriminating-subtitles	0.27 _a [0.17, 0.37]	0.10 _a [0.04, 0.16]	0.54 _b [0.44, 0.64]	0.16 _b [0.10, 0.22]

Note. Means with different subscript letters within columns are significantly different from each other.

Discrepancy Detection

The proportion of discrepancies detected for the critical items was computed and entered into a 2 (group) x 2 (video quality) mixed ANOVA. See Table 2 for descriptive statistics.⁵ The accurate-subtitles group was included to act as a baseline to control for guessing; that is, we included it to determine the proportion of times participants guessed that there was a discrepancy when there was none. Participants claimed to notice more discrepancies in the fluent-video condition ($M = 0.26$, 95% CI [0.18, 0.34]) than the disfluent-video condition ($M = 0.20$, 95% CI [0.12, 0.28]), with a small effect size, $F(1,54) = 4.24$, $p = 0.04$, $\eta^2 = 0.07$. Surprisingly, however, there was no difference in the rate of discrepancy detection between accurate and incriminating subtitle groups, and no interaction, largest $F(1,54) = 1.06$, $p = 0.31$, $\eta^2 = 0.02$.

To investigate whether noticing discrepancies led to protection from the concurrent misinformation effect, as has been shown with the traditional misinformation effect (e.g., Higham, Blank, et al., 2017), we examined recall performance as a function of discrepancy

⁵ The no-subtitles group was not included in this analysis because the participants were not asked the discrepancy detection question.

detection. Specifically, an analysis was conducted to compare the false recall rate on critical questions in the incriminating subtitle group for which participants indicated that they noticed a discrepancy (*discrepancy detected* or DD) versus did not notice a discrepancy (*no discrepancy detected* or NDD). The false recall rates for DD and NDD categories were computed separately for each experimental condition. The rate for DD items was calculated by dividing the number of words falsely recalled for DD items by the total number of DD items in each condition. The NDD rate was calculated in the same way, using the relevant NDD response rate in place of the DD rate. See Table 3 for summary statistics. Participants who selected 0 or 5 discrepancies had no data in either the DD or NDD categories, and thus were excluded from the analysis. In order to keep the number of participants excluded as low as possible, this analysis was conducted separately for fluent and non-fluent videos. For the fluent condition, 14 participants were still excluded, 13 because they selected 0 discrepancies and one because they selected 5 discrepancies. In the non-fluent condition, 19 participants were excluded, with 16 selecting no discrepancies and 3 selecting 5 discrepancies. This left 14 participants in the fluent and 9 participants in the non-fluent conditions. In this analysis, if discrepancy detection did indeed help participants to avoid the concurrent misinformation effect, then the false recall rate in the DD condition should be lower than that in the NDD condition. Two separate paired-sample t-tests (one for the fluent-video condition and the other for the disfluent-video condition) indicated that there were no differences in false recall between the DD and NDD items in either condition, largest $t < 1.00$. Note that due to the small number of participants remaining in this analysis, this result should be treated with caution.

Discrepancy detection could also influence people's ratings. Ratings could not be compared across DD and NDD items, as participants only provided a single rating of incrimination for the whole video, rather than individual ratings per question. Thus, instead, a

correlation was calculated across participants in the incriminating subtitle group between the number of discrepancies detected and ratings of incrimination. In both the poor and the good video quality conditions, the number of discrepancies detected was negatively correlated with ratings of incrimination, $r(27) = -0.595, p < 0.001$ and $r(27) = -0.445, p = 0.02$, respectively.

Table 3

Proportion Incorrect Words Recalled in the Incriminating-Subtitles Group by Video Quality and Discrepancy Detection Condition from Experiment 1 (95% Confidence Intervals in Brackets)

Discrepancy Condition	Video Quality	
	Poor	Good
Overall	0.56 [0.44, 0.68]	0.53 [0.41, 0.65]
Discrepancy Detected	0.48 _a [0.15, 0.81]	0.43 _a [0.19, 0.66]
No Discrepancy Detected	0.58 _a [0.34, 0.81]	0.54 _a [0.28, 0.79]

Note. Means with different subscript letters within columns are significantly different from each other.

Discussion

The results suggest that by co-presenting misleading textual information alongside audio, a concurrent misinformation effect was obtained. This concurrent misinformation distorted both participants' memory for audio in the interviews and their inferential judgments (i.e., incrimination ratings) about those interviews. In terms of memory, participants falsely recalled more critical incriminating words when given incriminating subtitles compared to when no subtitles were present. All of the effect sizes for the group differences were large, other than the one for incrimination ratings, which had a medium effect size.

The fact that false recall was low in the no-subtitles group is worthy of note. It indicates that participants did not self-generate or guess the incriminating words using the surrounding context of the words they could understand or the overall potentially incriminating context of the police interview. Thus, the incriminating words that were produced on the memory test were not naturally salient or a just a good fit to the context. Instead, the data suggest that the concurrent misinformation effect that we observed was a memory-based error.

For inferential judgments, the incrimination ratings were highest in the incriminating subtitle group. Thus, whilst most studies of misinformation concentrate on memory measures, Experiment 1 demonstrated that misinformation can also have a meaningful impact on decision making. This would be worrying in an applied setting as it implies that jurors who have been misled by written incriminating misinformation (e.g., an erroneous transcript) may have difficulty discounting that misinformation, giving the written misinformation more weight than it deserves.

There was also some evidence that discrepancy detection was difficult for participants. There were no differences in the number of discrepancies detected across subtitle groups, and in the incriminating-subtitles group, the same false recall rate was observed in the DD and NDD conditions. In other words, participants did not detect discrepancies when they should have, and when they did detect discrepancies, it did not prevent them falsely recalling incriminating words. In fact, in the incriminating subtitle group, about half of the participants in the fluent condition and more than half in the non-fluent condition detected no discrepancies at all, further indicating how difficult discrepancy detection was in this task. On the other hand, there was some evidence that detecting discrepancies was associated with a reduction in the effect of concurrent misinformation on inferential judgments; correlational analyses indicated that as participants detected more

discrepancies, the lower were their incrimination ratings. Thus, whilst discrepancy detection was difficult, discrepancy detection did appear to be associated with lower incrimination ratings overall.

Finally, contrary to our hypothesis, video fluency had no effect on incrimination ratings. In fact, the only effect of visual fluency was found in the number of discrepancies detected: participants detected more discrepancies with fluent videos than with disfluent ones, even though there were no differences in the actual number of discrepancies between these two conditions.

Experiment 2

In the traditional misinformation paradigm, warnings can play a part in protecting participants from misinformation. Thus, Experiment 2 explored whether the same might be true of the concurrent misinformation effect by introducing both a pre-warning and a stronger post-warning than the one used in Experiment 1.

As well as the addition of warnings, several other elements of the design were changed in Experiment 2. First, to ensure that the results from Experiment 1 were not peculiar to those particular materials or rating scales, a new video was recorded, using a different method to distort the sound and a different 5-point scale (see below for details). Second, we made the distortion in the audio slightly less severe to try to make it easier for participants to detect discrepancies. If participants have a better chance of hearing the authentic audio, then they may find discrepancy detection easier, and successful discrepancy detection may be more effective at controlling the concurrent misinformation effect. Third, whilst Experiment 1 only examined how incriminating the participants found the video, a jury may be making other judgments in a court situation, the central judgment being a guilty/not-guilty verdict. Additionally, jurors would need to judge whether audio should be considered reliable evidence, but this judgement is known to be contaminated by having knowledge of

what is said in the audio (Higham, Neil, et al., 2017). Consequently, additional ratings were included in Experiment 2: a guilty/not-guilty verdict, ratings of how clear the audio sounds, and a rating of whether a conviction would be justified.

Finally, the design of Experiment 2 was simplified. First, the manipulation of video quality was dropped because it exerted few effects in Experiment 1. Second, the accurate-subtitles group was dropped so that we could focus primarily on the effect of warning in moderating the concurrent misinformation effect.

Method

Participants

There were 120 participants (79 males and 41 females; age: $M = 35.7$, 95% CI [32.80, 38.60]), split equally between four groups. Participants were collected online through MTurk and were financially compensated for their time with \$6. This sample size allowed us to detect an effect size η_p^2 of 0.11 with power of 0.90, more than ample to detect misinformation effects (see earlier).

Design

The design consisted of a single between-participants factor of group with four levels. There was a control group with no subtitles, and three different incriminating subtitle groups. The weak-warning group had incriminating subtitles with only the weak warning used in the previous experiment, so it was identical to the incriminating subtitle group in Experiment 1. The pre- and post-warning groups had incriminating subtitles along with strong warnings either before or after the video, respectively. The pre-warning group also received the weak warning after the video as in the weak-warning group.

Materials and Procedure

The materials were broadly similar to Experiment 1, with the following changes. First, only one video was used, which was a different version of the “murder” scenario used

in Experiment 1, with the phrases slightly altered to provide a more naturalistic narrative (see Appendix B for phrases). To increase the likelihood that there was consistent delivery of the critical phrases and the rest of the video for a given actor, the entire video was scripted, and thus there was no ad-libbing. Thus, the actors did not know that the critical phrases differed from the rest of the script. Additionally, the video was recorded in a real police interview training room. The distortion applied to the audio was also different from Experiment 1 –a low-pass filter was applied, which cut off sound above 300Hz.⁶ A low-pass filter makes audio seem muffled, as if you were listening through a door or had accidentally covered the microphone on your recording device. The entire experiment was delivered online using Sphinx online surveys.

The procedure was nearly identical to Experiment 1 with the following changes. Everyone viewed the new murder interview video, which was hosted on YouTube. There was an additional page of instructions which guided participants to set the volume levels of their computer and YouTube such that everyone listened to the video at the same volume. All participants were also instructed to complete the experiment using headphones, on a desktop computer, and in a quiet environment. Participants in the no-subtitle and weak-warning subtitle groups received instructions as in Experiment 1. Those in the pre-warning group were given the following warning before the video:

“The video you are about to watch has subtitles. However, the transcribers that produce such subtitles sometimes make mistakes because they mishear words.

In the video you are about to watch, the transcriber has made some of these mistakes. Consequently, the subtitles will not always be accurate, so be sure to pay close attention to what you can hear.”

⁶ Whilst this may seem lower than low-pass filters used in other experiments, participants listened to the videos entirely using headphones. This made the audio somewhat clearer compared to speakers. Pilot testing suggested that at this level, around 40-50% of the words in the audio could be understood, enough to understand the gist of the interview.

Those in the post-warning group were given the same warning after the video (but before the questionnaire) with the wording changed appropriately to past tense.

Having watched the video, and read warnings if appropriate, the participants were given a questionnaire that was similar to the one used in the first experiment, with several changes. First, the rating scale was altered, to enable the scale to apply to a variety of judgments rather than just the incrimination judgment. The rating scale was still a 5-point scale, but the scale ran from 1 = *entirely disagree* through to 5 = *entirely agree*, with all the questions phrased as statements. In addition to rating whether they agreed that the audio was incriminating, they also rated whether they agreed that the audio was clear and whether a conviction would be justified based on the audio. They also provided their own guilty/not-guilty verdict. The recall questionnaire took the same form as in Experiment 1, using cued-recall questions, except that 10 questions were used instead of 7, to equalize the number of filler and critical questions at 5 each. Finally, at the end of the experiment there were several checks to ensure that the participant had followed instructions correctly. These checks included questions regarding whether they used headphones, whether they completed the experiment on a desktop computer, whether they watched the video more than once, and whether they understood the instructions.

Results

Rating Data

Mean ratings for audio clarity, whether a conviction would be justified, and whether the audio was incriminating can be found in Table 4. The different types of ratings were analyzed separately with one-way, between-participants ANOVAs. For judgments of audio clarity, there was no effect of group, $F(3,116) = 2.20, p = 0.09, \eta_p^2 = 0.04$. For incrimination ratings, there was an effect of group with a large effect size, $F(3,116) = 8.15, p < .001, \eta_p^2 =$

0.17. Bonferroni corrected pairwise comparisons revealed that all groups with incriminating subtitles were given higher ratings than the no-subtitles group (weak-warning: $p = 0.04$; pre-warning: $p < .001$; post-warning: $p < .001$). Thus, as in Experiment 1, there was a concurrent misinformation effect on incrimination judgments. However, warnings were ineffective, as there were no differences between the weak-warning, pre-warning and post-warning groups (lowest $p = 0.44$).

For ratings regarding whether the conviction was justified, there was an effect of group with a medium effect size, $F(3,116) = 3.44$, $p = .019$, $\eta_p^2 = 0.08$. Bonferroni corrected pairwise comparisons indicated that this was due to the post-warning group having higher ratings than the no-subtitles group ($p = 0.03$), with there being no difference between the other groups (lowest $p = 0.09$). Thus, for the conviction-justified rating, there was only a concurrent misinformation effect in the post-warning group. Descriptively, the other warning groups gave higher ratings than the no-subtitles group, but not high enough to reach statistical significance.

Table 4

Mean Ratings of Audio Clarity, Incrimination and Conviction Justified (95% Confidence Interval in Brackets) and Number of Guilty Verdicts (Percentage of Participants in Brackets) by Group from Experiment 2

Group	Rating Type			
	Audio Clarity	Incrimination	Conviction justified	Guilty
No-Subtitles	1.20 _a [1.02, 1.38]	2.17 _a [1.80, 2.54]	1.63 _a [1.18, 2.08]	3 _a (10.3%)
Weak-warning	1.63 _a [1.26, 2.00]	3.03 _b [2.60, 3.46]	2.40 _{ab} [1.95, 2.85]	13 _b (43.3%)
Pre-warning	1.53 _a [1.29, 1.76]	3.40 _b [2.95, 3.85]	2.43 _{ab} [1.98, 2.88]	10 _b (33.3%)
Post-warning	1.67 _a [1.38, 1.96]	3.60 _b [3.13, 4.07]	2.57 _b [2.12, 3.02]	14 _b (46.7%)

Overall	1.51 [1.37, 1.65]	3.05 [2.81, 3.28]	2.26 [2.02, 2.49]
---------	-------------------	-------------------	-------------------

Note. Means with different subscript letters within columns are significantly different from each other.

Guilty Verdicts

The number and percentage of guilty verdicts by group can be found in Table 4. Guilty verdicts were analyzed with a 2 (yes/no guilty) x 4 (group) chi-squared test, with one participant being excluded from the analysis due to missing data. The chi-squared test revealed significant differences in the rate of guilty verdicts between groups, $\chi^2(3, N = 120) = 10.60, p = 0.01, V = 0.30$. This result was followed up with a series of 2 (yes/no guilty) x 2 (group) chi-squared tests. The first compared the weak-warning group to the no-subtitles group, which indicated that the weak-warning group had a higher rate of guilty verdicts than did the no-subtitles group, $\chi^2(1, N = 60) = 8.12, p < 0.01, V = 0.37$. Another set of 2x2 chi-squared tests compared the other warning groups to the weak-warning group.⁷ The chi-squared tests indicated that neither the pre-warning, nor post-warning, group differed from the weak-warning group, $\chi^2(1, N = 60) = 0.63, p = 0.43$ and $\chi^2(1, N = 60) = 0.07, p = 0.79$, respectively. In other words, there was a concurrent misinformation effect in the weak warning compared to the no-subtitle group, in which the weak-warning group had a higher rate of guilty verdicts than the no-subtitles group. The other warning group showed the same level of guilty verdicts as the weak-warning group. Thus, the weak-warning group showed a concurrent misinformation effect compared with the no-subtitles group, but warnings were

⁷ The weak-warning group was used as the basis for comparison here rather than the no-subtitle group because the weak-warning group represented the baseline misinformed group with the mildest form of warning. Thus, if pre- or post-warnings mitigated the misinformation effect, then those groups should have lower rates of guilty verdicts than the weak-warning group. On the other hand, no difference between the misinformed groups would indicate that the pre- and post-warnings were ineffective.

not effective in mitigating this effect, as the concurrent misinformation effect remained in the pre- and post-warning groups.

Word Recall

As in Experiment 1, the rates of correct recall, false recall and “don’t know” responses were analyzed. See Table 5 for response type distributions. Collapsing across warning group, the response rates for categories 3-6 (omitted from the table) were 17%, 1%, 1% and 8% respectively (total = 27%). Once again, referring to the confidence intervals in Table 6, correct recall was greater than zero in all groups. A one-way, between-participants ANOVA on the proportion of correct recall indicated group did not exert a significant effect, $F < 1$. A similar ANOVA on the false recall revealed an effect of group with a large effect size, $F(3, 116) = 24.21, p < .001, \eta_p^2 = 0.38$. Bonferroni corrected pairwise comparisons indicated that, compared to the no-subtitle baseline, participants falsely recalled more incriminating words in all warning groups (all $ps < .001$), which once again demonstrates a concurrent misinformation effect on memory. Compared to the weak-warning group, neither a pre- ($p = 0.09$) nor post-warning ($p = 1.00$) reduced false recall. Finally, the ANOVA on “don’t know” responses revealed an effect of group with a large effect size, $F(3, 116) = 48.85, p < .001, \eta_p^2 = 0.56$. This was due to all warning groups having lower “don’t know” responding rates than the no-subtitles group ($p < .001$), whilst there was no difference in “don’t know” rates between any of the warning groups (all $p = 1.00$).

Overall, these analyses show that warnings did not prevent the concurrent misinformation effect. Participants also displayed similar levels of “don’t know” responding regardless of warning type, but those rates were much less than in the no-subtitles group. Although warnings did not significantly mitigate the concurrent misinformation effect, pre-warnings descriptively reduced false recall descriptively more than weak-warning or the post-warning (see Table 5). This possibility will be further investigated in the following analysis.

Table 5

Proportion of Recall Types By Group for Experiment 2 (95% Confidence Interval in Brackets).

Group	Response Type		
	Correct Recall	False Recall	Don't know
No-subtitles	0.16 _a [0.08, 0.24]	0.01 _a [-0.01, 0.03]	0.63 _a [0.51, 0.75]
Weak-Warning	0.24 _a [0.12, 0.36]	0.41 _b [0.31, 0.51]	0.10 _b [0.04, 0.16]
Pre-warning	0.27 _a [0.17, 0.37]	0.27 _b [0.19, 0.35]	0.11 _b [0.05, 0.17]
Post-warning	0.23 _a [0.15, 0.31]	0.41 _b [0.33, 0.49]	0.10 _b [0.03, 0.15]

Note. Means with different subscript letters within columns are significantly different from each other.

Discrepancy Detection

As in Experiment 1, all analyses in this section excluded the no-subtitles group who were not asked the discrepancy detection question. Table 6 shows the mean proportion of the five critical items for which discrepancies were detected in the three warning groups, as well as false recall conditioned on DD and NDD. A one-way ANOVA indicated that there were no differences in the proportion of discrepancies detected between the groups, $F(2,87) = 1.31, p = .27, \eta_p^2 = 0.03$. Next, we examined the effect of discrepancy detection on false recall across the three groups given subtitles with a 3 (group: weak-warning, pre-warning, post-warning) X 2 (discrepancy detection: DD, NDD) mixed ANOVA. Sixteen participants were excluded from this analysis because they noticed no discrepancies and one was excluded because they noticed all five discrepancies. This left data from 70 participants to analyze. The ANOVA showed that there were main effects of group, $F(2, 70) = 6.32, p = 0.003, \eta_p^2 = 0.15$, and discrepancy detection, $F(1, 70) = 18.46, p < .001, \eta_p^2 = 0.21$, both with large effect sizes, but no interaction, $F < 1$. The discrepancy detection main effect indicated that people had lower

false recall for DD items than NDD items. The group main effect was investigated with Bonferroni corrected pairwise comparisons. This analysis demonstrated that, whilst false recall was equal between the weak- and post-warning groups ($p = 1.00$), it was lower in the pre-warning group than in either the weak- or post-warning groups ($p = 0.01$ for both tests).

Table 6

Proportion of Critical Items for Which Discrepancies were Detected (Far-Left Column) and the False Recall Rate for DD and NDD Items (Middle Columns) by Warning Group from Experiment 2 (95% Confidence Intervals in Brackets)

Group	Discrepancy condition			
	Proportion of Discrepancies Detected	False Recall DD	False Recall NDD	False Recall Overall
No-subtitles	-	-	-	0.01 (0.01)
Weak-warning	0.38 _a [0.26, 0.50]	0.30 [0.16, 0.44]	0.50 [0.34, 0.66]	0.40 _a [0.30, 0.50]
Pre-warning	0.39 _a [0.29, 0.49]	0.07 [-0.05, 0.19]	0.31 [0.19, 0.43]	0.19 _b [0.09, 0.29]
Post-warning	0.49 _a [0.39, 0.59]	0.24 [0.12, 0.36]	0.55 [0.41, 0.69]	0.39 _a [0.29, 0.49]
Overall	0.42 [0.36, 0.48]	0.20 _I [0.12, 0.28]	0.45 _{II} [0.37, 0.53]	

Note. DD = discrepancy detected; NDD = no discrepancy detected. The discrepancy detection question was not asked in the no-subtitles group so there were only data corresponding to the overall false recall rate for that group. Means with different subscript letters within columns are significantly different from each other. Means with different subscript roman numerals within rows are significantly different from each other.

Note that there was a difference between the warning groups in this analysis, which may seem to contradict the absence of any differences between the warning groups in the previous analysis on false recall. However, only participants who noticed some (but not all)

discrepancies were included in this second analysis.⁸ In contrast, the previous analysis on false recall included all participants, even those who noticed no discrepancies. Thus, when the analysis is focused on participants who detected at least one discrepancy, false recall was significantly less in the pre-warning group than in either the weak- or post-warning groups.

The lower false recall in the pre-warning group compared to the other groups could be due to participants focusing more on the audio in the pre-warning group, enabling them to correctly encode the actual answers, as well as paying less attention to the subtitles. Thus, the pre-warned participants might have avoided the influence of the subtitles, reducing false recall, whilst encoding more audio information, increasing correct recall. If that was the case, then there should be higher correct recall in the pre-warning group compared to the other groups. To test this explanation, correct recall was re-analyzed, excluding the same participants as for the previous analysis (Table 7). The ANOVA showed that there was an effect of discrepancy detection with a large effect size, $F(1, 70) = 59.33, p < .001, \eta_p^2 = 0.46$, but no other effects, largest $F(1, 70) = 1.69, p = 0.23, \eta_p^2 = 0.04$. Thus, whilst correct recall was superior for items on which discrepancies were detected across all groups, there was no correct recall advantage in the pre-warning group compared to the other warning groups. These analyses suggest that participants did not respond to the pre-warning by focusing more on the audio than in the other groups.⁹

⁸ Although any participant who noticed all discrepancies was also excluded from the second analysis involving discrepancy detection as a variable, the vast majority of exclusions were participants who noticed no discrepancies (16/17 = 94%).

⁹ It is possible that participants may have increased their focus on the audio and ignored the subtitles, but they failed to encode any additional critical words because the audio was so poor. Under these circumstances, correct recall would not be enhanced despite the enhanced attentional focus. However, given that we took a number of measures in this experiment to improve the clarity of the audio (and the data suggest that these measures were successful), complete failure to detect any additional words seems unlikely.

Table 7

Proportion of Critical Items Correctly Recalled for DD and NDD Questions by Warning Group in Experiment 2 (95% Confidence Intervals in Brackets)

Group	Discrepancy Condition		
	Correct recall DD	Correct Recall NDD	Correct Recall Overall
Weak-warning	0.33 _a [0.31, 0.35]	0.05 _b [-0.05, 0.15]	0.19 [0.09, 0.29]
Pre-warning	0.58 _a [0.42, 0.74]	0.07 _b [-0.10, 0.25]	0.33 [0.23, 0.43]
Post-warning	0.50 _a [0.34, 0.66]	0.08 _b [0.02, 0.14]	0.29 [0.19, 0.39]
Overall	0.47 _a [0.37, 0.57]	0.07 _b [0.03, 0.11]	

Note. DD = discrepancy detected; NDD = no discrepancy detected. Means with different subscript letters within rows are significantly different from each other.

Finally, the extent to which discrepancy detection influenced judgments was again investigated. For incriminating ratings and conviction-justification ratings, correlations across participants were calculated between mean ratings and the number of discrepancies detected. As the number of discrepancies detected did not vary by warning group, the correlations were calculated after collapsing across the warning groups, thereby increasing the power of the analyses. Both incrimination rating and conviction justified ratings were weakly negatively correlated with proportion of discrepancies noticed, $r(89) = -0.25$, $p = 0.02$ and $r(89) = -0.24$, $p = 0.02$ respectively. The effect of discrepancy detection on guilty verdicts was investigated by conducting a t-test using guilty verdicts as an independent variable, and the proportion of discrepancies detected as the dependent variable. Those who gave innocent verdicts detected more discrepancies ($M = 0.49$, [95% CI = 0.41, 0.57]) than those who gave guilty verdicts ($M = 0.33$, [95% CI = 0.25, 0.41]), $t(88) = 2.54$, $p = 0.01$, $d = 0.54$, with a medium effect size.

Discussion

As in Experiment 1, concurrent misinformation effects were found in both memory and inferential judgment measures, even with different materials and a different response scale, with effect sizes ranging from medium to large. The recall and discrepancy detection data suggest that participants' audio identification rates were improved compared to Experiment 1; collapsing across incriminating and no-subtitle groups, correct recall increased from 0.10 in Experiment 1 to 0.22 here. The rate of discrepancy detection was also higher in Experiment 2 than Experiment 1; collapsing across fluency, participants detected 0.27 discrepancies in the incriminating-subtitles group of Experiment 1 compared to 0.38 in the weak-warning group here.¹⁰ Compared to Experiment 1, discrepancy detection was more successful at mitigating the effect of concurrent misinformation in Experiment 2. For example, unlike Experiment 1, there was lower false recall for DD items than for NDD items. Discrepancy detection also appeared to mitigate the effect of concurrent misinformation on judgments, with higher discrepancy detection being associated with lower incriminating ratings and fewer guilty judgements.

Warning participants about the presence of misinformation had limited effects on performance. On the one hand, warnings did not increase discrepancy detection; nor did they prevent the concurrent misinformation effect on guilty judgments or incrimination ratings. The fact all three groups differed in their incrimination ratings from the no-subtitles baseline with a large effect size, coupled with the fact that there was no difference between the three warning groups, suggests that the concurrent misinformation effect on those ratings was robust. However, for conviction-justified ratings, the situation was a little different. For those ratings, only the post-warning group showed a clear concurrent misinformation effect, with statistically higher ratings than the no-subtitles group. Whilst this may imply that the pre-

¹⁰ These two groups were compared because, other than two videos being used in Experiment 1 and one video in Experiment 2, the groups were otherwise the same.

warning diminished the concurrent misinformation effect, the weak-warning group did not show a concurrent misinformation effect either. Thus, rather than being an effect of warning type per se, the fact that incriminating subtitles only affected conviction-justified ratings in the group given a post-warning may indicate that this particular rating type was simply less sensitive to misinformation. In this vein, the main effect of group on conviction-justified ratings had a lower effect size (0.08) than that associated with the group main effect on incrimination rating (0.17). Finally, in the analysis of the full dataset, neither the pre- nor the post-warning reduced misinformation-induced false recall compared to the weak-warning group.

On the other hand, if participants who failed to notice any discrepancies were removed from the analysis, pre-warnings did reduce false recall compared to the other warning groups. It may be that the participants who failed to detect any discrepancies ignored, forgot, did not believe, or chose not to act on the pre-warning, rendering it ineffective. Thus, the efficacy of pre-warnings may depend in part on participants' willingness to act upon them, or the extent to which the pre-warning is believable.

It is important to consider how some pre-warned participants were able to mitigate the effect of concurrent misinformation. Although we cannot state with confidence what they *did* do, the data tell us something about what they did *not* do. First, pre-warnings did not appear to make participants more cautious in their responding. Had that occurred, it is likely that the rate of "don't know" responding would have been high. Research has found that people can effectively use the option to report or withhold information to strategically regulate their memory accuracy (e.g., Higham, 2007; Koriat & Goldsmith, 1996). By responding conservatively, the accuracy of the responses that *were* provided (i.e., output-bound

accuracy) would likely be high. However, as shown in Table 5, the “don’t know” rate was no higher in the pre-warning group than in the groups who received a weak- or post-warning.¹¹

Second, pre-warnings did not appear to persuade participants to focus their attention on the audio to a greater extent than if no pre-warning was given. Participants could easily have closed their eyes, blurred their vision, or adopted another strategy to better encode the audio and reduce the distraction from the misleading information in the subtitles. Had they done so, they would likely have had better memory for the correct words. However, the pre-warning group had the same level of correct recall as other warning groups. This result suggests that the subtitles (or the effort to ignore them) interfered with pre-warned participants ability to focus on the audio. We will further consider the implications of these findings in the General Discussion.

General Discussion

The results of two experiments indicated that incriminating but misleading subtitles presented concurrently with distorted audio in a video recording of a mock police interview created both false memories and biased inferential decision-making. In Experiment 1, ratings of incrimination were inflated, and false recall was higher, in the incriminating-subtitles group compared to the no-subtitles group. Successful discrepancy detection did not mitigate the concurrent misinformation effect on memory, although it did mitigate decision-making to some extent. However, many participants detected no discrepancies in that experiment, and baseline accuracy was low. Experiment 2 improved the baseline accuracy, and the results indicated that the concurrent misinformation effect generalized to different materials and other ratings, including judgments of guilt. Neither pre- nor post-warnings mitigated the concurrent misinformation’s effect on either memory or inferential decisions, although pre-

¹¹ In contrast, the pre-warned “don’t know” rate was much lower than in the no-subtitles group, most likely because the latter participants could not produce a plausible candidate answer to many of the test questions either via retrieval or guessing.

warnings did reduce false recall when the analysis was restricted to participants who noticed at least one, but not all, discrepancies. Unlike Experiment 1, increased discrepancy detection was associated with lower false recall and less inferential bias. Specifically, false recall was lower when discrepancies were detected than when they were not, and higher discrepancy rates was associated with lower ratings of incrimination and fewer guilty judgments. These results will be discussed in relation to possible underlying mechanisms, and the applied implications.

The Concurrent Misinformation Effect and Underlying Mechanisms

Whilst most misinformation studies focus on the effects of post-event information, our results demonstrate that information present at the time of encoding can also have a powerful effect. Whilst concurrent and traditional misinformation effects likely share some common mechanisms, our data suggest that there may also be differences. Discrepancy detection was less effective in reducing the concurrent misinformation effect compared to the traditional misinformation effect. For example, discrepancy detection had no effect at all on performance in Experiment 1. Whilst discrepancy detection was more effective in Experiment 2, it only reduced the concurrent misinformation effect for participants who detected discrepancies, and even then, a significant residual misinformation effect remained. In contrast, Higham, Blank, et al. (2017) found that discrepancy detection was associated with almost complete elimination of the traditional misinformation effect. Additionally, with the traditional misinformation effect, pre-warnings are usually extremely effective at reducing the effect (e.g., Schul, 1993), and so are strongly worded post-warnings (Blank & Launay, 2014). However, in Experiment 2, pre-warnings had a limited effect, and post-warnings had no effect other than to increase the number of discrepancies detected.

Given that the concurrent and traditional misinformation effects responded differently to countermeasures, the concurrent misinformation effect may relate differently to the

underlying mechanisms usually employed to explain the traditional misinformation effect. As with traditional misinformation mitigation, both warnings and discrepancy detection could target participants' *strategic* decision making. There are many strategic decisions participants might make in a misinformation experiment. They may decide to trust, or ignore, the misinformation because their original memory is poor, or strong (McCloskey & Zaragoza, 1985). Participants could also make a pre-encoding choice to ignore the subtitles, and thus, ignore the misinformation (Dodd & Bradshaw, 1980). However, subtitles are quite difficult to ignore or filter out (Brasel & Gips, 2014; d'Ydewalle et al., 1991), and thus participants may have found it difficult to avoid encoding the subtitles, regardless of decisions to ignore them, and regardless of warnings that they should be ignored¹². Further compounding this issue, encoding visuals, subtitles and audio simultaneously can overload working memory (Brunyé et al., 2006; Jamet & Le Bohec, 2007; Moreno & Mayer, 2002a). Working memory overload would leave few cognitive resources to make assessments about the encoded memory strength of the audio, or to make strategic decisions during encoding about whether to encode only audio, only subtitles, or both together. Thus, the lack of time to make decisions about encoding, the cognitive overload involved in co-processing concurrent text and audio, and the compelling nature of subtitles likely led to poor or unachievable strategic decisions in our experiments, even in the face of warnings. Consequently, concurrent misinformation may be more difficult to ignore, and more difficult to mitigate, than post-event misinformation.

The second mechanism often linked to the traditional misinformation effect refers to the fate of the original memory when misinformation is introduced. Usually, it is assumed that the original memory (the audio in our experiments) suffers one of two fates. One possibility is that access to the original memory is *blocked* by the misinformation. Blocking

¹² Even in the traditional misinformation paradigm, participants do not always ignore misinformation when they have the time to compare it with the original event. When given the chance to directly compare the original information and the misinformation, participants can still make errors based on the misinformation Polak et al. (2016)

might occur because the misinformation was more recently encountered and/or has a stronger memory trace associated with it than the original event, making the original memory harder (or impossible) to retrieve (Belli, 1989; Eakin et al., 2003; Gordon & Shapiro, 2012).

Alternatively, the original memory is thought to be *destructively updated* – the original memory is overwritten or permanently altered by the misinformation (Loftus, 1979; Skagerberg & Wright, 2008).

Whilst the destructive updating hypothesis had been mostly discounted, a recent form of the destructive updating hypothesis is found in research on *disrupted memory reconsolidation* (Chan & LaPaglia, 2013). This occurs when testing a participant about a witnessed event causes them to retrieve the original memory. If misinformation is introduced at this point, it distorts the original memory before it is reconsolidated in memory, producing *retrieval enhanced suggestibility* compared to cases where no immediate testing occurs (Chan & LaPaglia, 2011; Chan et al., 2009). Whilst disrupted memory reconsolidation can potentially explain some data derived from the traditional misinformation paradigm, it simply cannot occur in the concurrent misinformation paradigm because misinformation is introduced simultaneously with the original event memory. Indeed, our data suggest that neither blocking or destructive updating occurred in our experiments, as correct recall of critical items in the misinformed groups was no different than in the no-subtitles groups, where no misinformation was present (see Tables 2 and 5). In fact, in Experiment 2, correct recall of critical items was descriptively higher in the incriminating subtitle group than in the no-subtitle group.

Given the discussion above, a *source confusion* (Higham, 1998; Lindsay & Johnson, 1989; Schacter et al., 1984) explanation would appear to be most consistent with our data. The idea of source confusion is that participants misattribute memories from the misinformation source to the original event source, inducing participants to report

misinformation as if it were in the original event (Higham, 1998; Lindsay & Johnson, 1989; Schacter et al., 1984). Applied to our experiments, participants may have encoded both the audio and the subtitles, but had difficulty recalling the source of these memories, attributing words read in the subtitles as heard in the audio. Participants recalled both correct and incorrect words in the incriminating-subtitles groups, suggesting that both audio and subtitles were indeed encoded, and that words read in the subtitles were reported as if they had been heard in the audio. Thus, participants would seem to have had difficulty identifying the source of their memories, or else ignored instructions to only report words from the audio.

Assuming a source confusion explanation for our data, both early discrepancy detection whilst viewing the video and pre-warnings could have helped mitigate misinformation by encouraging participants to improve their source encoding (Ecker et al., 2010; Higham, Blank, et al., 2017; Oeberst & Blank, 2012). However, this may have been difficult, as source confusion is likely to be more severe with concurrent information than with traditional misinformation, due to the difficulties inherent in how the memories are initially formed (Raj & Bell, 2010). For example, source encoding could be poor due to the earlier-noted cognitive overload induced by concurrently presented materials (Moreno & Mayer, 2002b). This overload can lead to difficulties binding memory to source (Mammarella & Fairfield, 2008), making later source memory decisions inaccurate. Source encoding could also be impaired because, as noted in the introduction, the contextual overlap between the audio and textual sources is likely to be high, and high contextual overlap impairs source encoding (Johnson et al., 1988; Johnson et al., 1993; Pezdek & Greene, 1993). Thus, both discrepancy detection and pre-warnings may have been less effective for concurrent misinformation than with traditional misinformation, as traditional misinformation usually has less contextual overlap, and no cognitive overload from co-presentation of sources, because the event information and the misinformation are temporally separated. In future

research, it would be interesting to investigate whether these difficulties could be overcome, perhaps using measures to help participants filter out irrelevant sources, such as focusing on speech characteristics at encoding (Kovacs & Newcombe, 2006).

As well as source monitoring at encoding, source can also be monitored at retrieval. Specifically, in Johnson et al.'s (1993) *Source Monitoring Framework*, source monitoring involves an assessment of the qualitative features of retrieved memories and an inference about what the source must have been based on that information (e.g., if the perceptual features of a memory were of someone talking, the source was a voice; if perceptual features of a memory were written words, the source was a book). Discrepancy detection and both pre- and post-warnings might encourage participants to improve their monitoring of source at retrieval (Blank & Launay, 2014; Higham, Blank, et al., 2017; Oeberst & Blank, 2012). However, such efforts may be hampered in the concurrent misinformation paradigm compared to traditional misinformation paradigm because, as noted above, there are many factors that might prevent effective source encoding to begin with. Thus, the sources may be too similar or poorly encoded to disentangle at retrieval, even when discrepancy detection or warnings encourage additional effort to be applied. Mitigating source confusion at retrieval may also be difficult as is not clear whether our participants encoded audio-visual sources in a task-appropriate way. For instance, they may have encoded audio and subtitles as separate sources, or they may have encoded them as simply video. If the source was encoded as video, measures that focused participants on interrogating source more closely at retrieval could not be effective.

Even though neither discrepancy detection nor pre-warnings were as effective with the concurrent misinformation effect as for the traditional misinformation effect, they both mitigated the concurrent misinformation effect to some degree. Considering discrepancy detection first, its association with a reduction of the concurrent misinformation effect may

have been causal or more indirect. In the former case, the sheer act of detecting a discrepancy may have laid down a memory trace that would help to discriminate source at test. For example, participants may have encoded in memory the difference between what was stated in the subtitles versus what was said in the audio (e.g., “the suspect **said** *muddy*, but the subtitles **stated** *bloody*”). If so, the effect of discrepancy detection on reducing the concurrent misinformation effect was causal because it produced source discrimination during encoding which would have been helpful on the test. However, it is also possible that discrepancy detection had no causal effect on performance per se, but instead was simply an indicator of better audibility. More audible words may have been less susceptible to misinformation, as their clarity might make them more obviously at odds with the subtitles, and thus marked as discrepant more often compared to less audible words. In this scenario, discrepancy detection was simply a proxy for audibility with the latter having the causative effect on performance. Whether discrepancy detection operated independently from audibility could be tested by including a direct measure of audibility for the critical phrases, and directly manipulating the audibility of different phrases.

Considering pre-warning next, recall that if the analysis was restricted to participants who detected some, but not all, discrepancies, pre-warnings mitigated the concurrent misinformation effect to some extent in Experiment 2. Compared to no pre-warning, the natural assumption is that pre-warnings helped participants in one of three ways: (1) it helped participants detect more discrepancies during encoding (2) it caused participants to focus their attention on the audio and to ignore the subtitles, or (3) it caused participants to adopt a more conservative reporting strategy by responding “don’t know” more often on the recall test. However, as noted earlier, none of these explanations fully accounts for the data.

One clue about what pre-warned participants might have been doing comes from a closer analysis of the low “don’t know” rate for pre-warned participants in Experiment 2.

The response rates of non-incriminating, incorrect, words were 0.15, 0.21 and 0.12 for the weak-warning, pre-warning and post-warning groups respectively. Therefore, at least descriptively, pre-warned participants chose to guess with a non-incriminating (but incorrect) word rather than respond “don’t know”. Pre-warned participants might have chosen to adopt this strategy because they were trying to avoid making serious errors; by guessing with a non-incriminating word when they were unsure of an answer on the recall test, they may not have answered the question correctly, but at least they did not falsely incriminate the suspect. These results suggest that there is a need to consider more sophisticated accuracy regulation strategies that participants might adopt in response to pre-warnings than just withholding responses. It seems that pre-warned participants may be reluctant to respond “don’t know” too often and would rather produce a benign error than no response at all.

Both pre- and post-warnings might be more effective if they were strengthened. Blank and Launay (2014) classified post-warnings at four levels of specificity, with more specificity constituting stronger warnings. The same classification system could also be applied to pre-warnings. At lowest level of specificity, a warning only mentions the possible presence of misinformation. At the next level, a warning assures that misinformation is present, but does not specify its nature or the amount of it. At the third level, the warning not only assures that misinformation is present, but also states that *no* information from the misinformation source (e.g., post-event narrative in the traditional misinformation paradigm) provides an accurate answer on the memory test. Finally, at the most specific level, a warning specifies exactly what the misinformation is and where it occurs (e.g., Wright, 1993).

The fourth level is usually only employed in studies investigating the fate of the original event memory trace as false reports or endorsements of misinformation are unlikely to occur with such a strong, specific warning. However, such false reports or endorsements of misinformation can still occur at the third level, which uses *opposition logic* (Jacoby et al.,

1989). The name derives from the fact that this type of warning sets the ability to recollect the source of misinformation in opposition to the tendency to report misinformation due to its familiarity. Lindsay (1990) applied opposition logic to a traditional misinformation study and found that as long as the event and post-event sources were hard to discriminate, misinformation was observed.

In the concurrent misinformation case, a warning using opposition logic would specify that no detail presented in the subtitles would be a valid response on the memory test.¹³ If performance in the opposition condition was compared to another *inclusion* condition for which details occurring in *either* source (audio or subtitles) were acceptable responses, then it would be possible to estimate controlled and automatic memory influences using *process dissociation procedure* (Jacoby, 1991). With this procedure, the difference in the endorsement rate of misleading details between the opposition and inclusion conditions provides a measure of controlled influences on memory (C). On the other hand, the endorsement rate in the opposition condition divided by (1-C) provides an estimate of automatic influences on memory. Some researchers have successfully applied process dissociation theory to investigate traditional misinformation effects in children (see Holliday et al., 2002 for a review), so it may be a fruitful avenue for future research on the concurrent misinformation effect.

Applied Context

The most obvious applied implication of the experiments presented above is that, in support of Lange et al.'s (2010) results, providing accompanying text can bias people's interpretation of audio. Whilst Lange et al. focused on biasing transcribers, our results suggest that transcribers' misinterpretations can go on to affect the memories and inferential

¹³ Of course, the memory test would have to be constructed so that this was true. For example, filler questions about details occurring in both the audio and subtitles would have to be omitted.

decisions of those who are provided with the transcriptions, such as judges and juries. Together the results of both studies suggest that misinformation in subtitles and transcriptions can have wide-ranging effects within the criminal justice system.

Although it seems intuitive to assume that, if concurrent misinformation affected memory, it should also affect inferential judgments, this need not have been the case. Such judgments occurred after the memory judgement and may have involved additional deliberation and effort. Effortful processing has been shown to reduce the misinformation effect (Lindsay & Johnson, 1989); thus, the effortful processing involved in inferential judgments could have mitigated the concurrent misinformation effect. However, in our experiments, the additional processing involved in making the inferential decisions, such as guilt, did not eliminate the concurrent misinformation effect. Alternatively, people could have reflected on the fact that the audio was difficult to understand due to its poor quality, and thus should be discounted as evidence regardless of accompanying text. However, clearly the audio was not discounted due to its poor audibility because even in the pre-warning group, participants rated the audio as incriminating, and judged the defendant as guilty at greater-than-baseline levels. The fact that concurrent misinformation affected inferences such as guilt judgments is worrying to the extent that it may have the same effect on jury decision-making in actual courtrooms.

Given the potential of concurrent misinformation to affect guilt judgements, the use of text with audio, even with warnings, should be approached with care. As noted in the introduction, mistranscriptions can occur with distorted audio (Lange et al., 2010), and audio used in court can be distorted for many reasons. For instance, audio could be difficult to understand due to background noise, or due to differences between the accent of the speaker and the listener. If the subject's mouth is not visible, the audio can be unclear, as even those with normal hearing use lip reading when interpreting speech (Remez et al., 2005; Sumbly &

Pollack, 1954). Poor audio is also not an automatic reason why courts might reject audio evidence; in the US, if the prosecution and defense cannot agree on a transcript due to poor audio, they are each permitted to make a different transcript and present both to the jury (Miller, 2016). Whilst judges may warn juries about the possibility of inaccurate transcripts, care should be taken here too. When there is no misinformation present, warnings can actually lead to lower memory performance than when no warning is used (Szpitalak & Polczyk, 2010).

Where text is used, its form may be important. The form of text used in court can vary – videos in court may be appear with subtitles, but they are also commonly presented with written transcripts. Such transcripts are likely to be more prone to the concurrent misinformation effect than subtitles because transcripts require looking away from any visuals associated with the audio. Thus, the transcriptions can direct people’s attention away from the visuals, which is problematic as those visuals may help to resolve any ambiguities created by differing transcript and audio. Ultimately, given the possibility for accidental misinformation, it may be safer to not use transcripts in court cases at all (Miller, 2016), especially since having a transcript can create an illusion of audibility (Bernstein et al., 2012; Higham, Neil, et al., 2017). It is also difficult to know if a transcript is actually accurate, as if the audio is unclear, there may be no way to determine what is actually being said. However, the risk of misleading transcripts would need to be weighed against the benefit of accurate transcripts helping juries to understand poor audio.

In future research, the effects of delay and repetition may be particularly relevant for concurrent misinformation. Written misinformation has been shown to be more effective across long delays (Higham, 1998), or when misinformation is repeated (Zaragoza & Mitchell, 1996), both characteristics which typify how audio evidence may be used in a court context. Evidence is pored over and repeated, and decisions of guilt may not be made for

days or weeks after the presentation of some pieces of evidence. Given this, the effects of delay could be especially problematic for several reasons. Source memory decays faster than event memory (Higham, 1998), which would exacerbate the difficult source encoding circumstances of concurrent text and audio noted above. Also, warnings might be given at the time that audio was presented, but where decisions on the basis of that audio are made later, listeners may not remember the warning¹⁴.

Finally, whilst here we have focused on the concurrent misinformation effect in a forensic context, it is relevant to many non-forensic contexts. For instance, investigative journalists frequently record conversations with the subject of their investigations whilst undercover. Later, journalists may choose to present their recordings to their editors, or the public, with the audio subtitled. Errors in transcription, or malicious intent to mislead, could result in misleading subtitles, resulting in a concurrent misinformation effect manifesting in people misremembering the actual audio, and making unjustified inferential judgements about the interview topic. Many other situations similarly involve judgments made about transcribed, but distorted, audio recordings, such as non-primary investigators listening to interpretations of black box recordings from crashed airplanes. Education too could be prone to concurrent misinformation effects. Educators sometimes accidentally verbally contradict written information provided to their students prior to teaching sessions (e.g., lecture slide handouts), creating a situation where misinformation has been introduced, but where students may not immediately realize. Smith et al. (2017) even demonstrated that misinformation can be introduced to students when they copy lecture notes from each other. Written notes about a lecture were given to a participant, with the notes containing deliberate mistakes. The participants then watched the lecture that the notes were about. Misinformation from the notes was present in their later memory reports about the lecture. Thus, in all of these

¹⁴ We thank an anonymous reviewer for this observation.

situations, it is important to explore the extent to which concurrent misinformation could be mitigated. With recording capabilities being ubiquitous in the form of mobile phones and tablets, it is an ideal time to put the way that audio is used and misused under the microscope.

Context

The idea for this paper initially occurred to us through two strands of work. The first was through work in voice recognition (e.g., Stevenage et al., 2013), which is often compared to face recognition. The second was through our work on the audio hindsight bias (e.g., Higham, Neil, et al., 2017), in which participants misjudged how audible distorted words were. During this work, we came across Lange et al.'s (2010) work on how mistranscriptions can occur with distorted audio, and how these mistranscriptions are more likely in a forensic context. Thus, the current paper represents an intersection of our previous work. Interested readers may wish to read the literature concerning voice processing and memory, as a general theme in this literature is that you cannot always generalize research in other modalities to voices.

References

- Belli, R. F. (1989). Influences of misleading postevent information: Misinformation interference and acceptance. *Journal of Experimental Psychology: General*, *118*(1), 72-85. <https://doi.org/10.1037/0096-3445.118.1.72>
- Bernstein, D. M., Wilson, A. M., Pernat, N. L., & Meilleur, L. R. (2012). Auditory hindsight bias. *Psychonomic Bulletin & Review*, *19*(4), 588-593.
- Blank, H. (1998). Memory states and memory tasks: an integrative framework for eyewitness memory and suggestibility. *Memory*, *6*(5), 481-529. <https://doi.org/10.1080/741943086>
- Blank, H., & Launay, C. (2014). How to protect eyewitness memory against the misinformation effect: A meta-analysis of post-warning studies. *Journal of Applied Research in Memory and Cognition*, *3*(2), 77-88. <https://doi.org/10.1016/j.jarmac.2014.03.005>
- Blank, H., Ost, J., Davies, J., Jones, G., Lambert, K., & Salmon, K. (2013). Comparing the influence of directly vs. indirectly encountered post-event misinformation on eyewitness remembering. *Acta Psychologica*, *144*(3), 635-641. <https://doi.org/10.1016/j.actpsy.2013.10.006>
- Brasel, S. A., & Gips, J. (2014). Enhancing television advertising: same-language subtitles can improve brand recall, verbal memory, and behavioral intent [journal article]. *Journal of the Academy of Marketing Science*, *42*(3), 322-336. <https://doi.org/10.1007/s11747-013-0358-1>
- Brunyé, T. T., Taylor, H. A., Rapp, D. N., & Spiro, A. B. (2006). Learning procedures: the role of working memory in multimedia learning experiences. *Applied Cognitive Psychology*, *20*(7), 917-940. <https://doi.org/10.1002/acp.1236>

- Calvillo, D. P., Parong, J. A., Peralta, B., Ocampo, D., & Van Gundy, R. (2016). Sleep increases susceptibility to the misinformation effect. *Applied Cognitive Psychology, 30*(6), 1061-1067.
- Chambers, K. L., & Zaragoza, M. S. (2001). Intended and unintended effects of explicit warnings on eyewitness suggestibility: Evidence from source identification tests. *Memory & Cognition, 29*(8), 1120-1129. <https://doi.org/10.3758/BF03206381>
- Chan, J. C., & LaPaglia, J. A. (2011). The dark side of testing memory: Repeated retrieval can enhance eyewitness suggestibility. *Journal of Experimental Psychology: Applied, 17*(4), 418. <https://doi.org/10.1037/a0025147>
- Chan, J. C., & LaPaglia, J. A. (2013). Impairing existing declarative memory in humans by disrupting reconsolidation. *Proceedings of the National Academy of Sciences, 110*(23), 9309-9313. <https://doi.org/10.1073/pnas.1218472110>
- Chan, J. C., Thomas, A. K., & Bulevich, J. B. (2009). Recalling a witnessed event increases eyewitness suggestibility: The reversed testing effect. *Psychological Science, 20*(1), 66-73. <https://doi.org/10.1111/j.1467-9280.2008.02245.x>
- Christiaansen, R. E., & Ochalek, K. (1983). Editing misleading information from memory: Evidence for the coexistence of original and postevent information. *Memory & Cognition, 11*(5), 467-475. <https://doi.org/10.3758/BF03196983>
- Clifford, B. R., & Scott, J. (1978). Individual and situational factors in eyewitness testimony. *Journal of Applied Psychology, 63*(3), 352. <https://doi.org/10.1037/0021-9010.63.3.352>
- Cohen, J. (1969). *Statistical power analysis for the behavioral sciences*. Academic press.
- Coulthard, M., Johnson, A., & Wright, D. (2016). *An introduction to forensic linguistics: Language in evidence*. Routledge. <https://www.jstor.org/stable/41055334>

- d'Ydewalle, G., Praet, C., Verfaillie, K., & Rensbergen, J. V. (1991). Watching Subtitled Television: Automatic Reading Behavior. *Communication Research, 18*(5), 650-666.
<https://doi.org/10.1177/009365091018005005>
- Dalton, A. L., & Daneman, M. (2006). Social suggestibility to central and peripheral misinformation. *Memory, 14*(4), 486-501.
<https://doi.org/10.1080/09658210500495073>
- Daneman, M., Thannikkotu, C., & Chen, Z. (2013). Are there age-related differences in social suggestibility to central and peripheral misinformation? *Experimental Aging Research, 39*(3), 342-369. <https://doi.org/10.1080/0361073X.2013.779201>
- Dodd, D. H., & Bradshaw, J. M. (1980). Leading questions and memory: Pragmatic constraints. *Journal of Verbal Learning and Verbal Behavior, 19*(6), 695-704.
[https://doi.org/10.1016/S0022-5371\(80\)90379-5](https://doi.org/10.1016/S0022-5371(80)90379-5)
- Dodson, C. S., Powers, E., & Lytell, M. (2015). Aging, confidence, and misinformation: Recalling information with the cognitive interview. *Psychology and Aging, 30*(1), 46.
<https://doi.org/10.1037/a0038492>
- Dowell, J., & Shmueli, Y. (2008). Blending Speech Output and Visual Text in the Multimodal Interface. *Human Factors, 50*(5), 782-788.
<https://doi.org/10.1518/001872008x354165>
- Eakin, D. K., Schreiber, T. A., & Sergent-Marshall, S. (2003). Misinformation effects in eyewitness memory: the presence and absence of memory impairment as a function of warning and misinformation accessibility. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29*(5), 813. <https://doi.org/10.1037/0278-7393.29.5.813>

- Ecker, U. K., Lewandowsky, S., & Tang, D. T. (2010). Explicit warnings reduce but do not eliminate the continued influence of misinformation. *Memory & Cognition*, *38*(8), 1087-1100. <https://doi.org/10.3758/MC.38.8.1087>
- Finley, L. (2015). Murder in Aspen. *Fourty Eight Hours*. <https://www.cbsnews.com/news/48-hours-probes-murder-of-aspen-legend-nancy-pfister/>
- Frost, P., Ingraham, M., & Wilson, B. (2002). Why misinformation is more likely to be recognised over time: A source monitoring account. *Memory*, *10*(3), 179-185. <https://doi.org/10.1080/09658210143000317>
- Goodwin, K. A., Kukucka, J. P., & Hawks, I. M. (2013). Co-witness confidence, conformity, and eyewitness memory: An examination of normative and informational social influences. *Applied Cognitive Psychology*, *27*(1), 91-100. <https://doi.org/10.1002/acp.2877>
- Gordon, L. T., & Shapiro, A. M. (2012). Priming correct information reduces the misinformation effect. *Memory & Cognition*, *40*(5), 717-726. <https://doi.org/10.3758/s13421-012-0191-7>
- Greene, E., Flynn, M. S., & Loftus, E. F. (1982). Inducing resistance to misleading information. *Journal of Verbal Learning and Verbal Behavior*, *21*(2), 207-219. [https://doi.org/10.1016/S0022-5371\(82\)90571-0](https://doi.org/10.1016/S0022-5371(82)90571-0)
- Gurney, D. J., Ellis, L. R., & Vardon-Hynard, E. (2016). The saliency of gestural misinformation in the perception of a violent crime. *Psychology, Crime & Law*, *22*(7), 651-665. <https://doi.org/10.1080/1068316X.2016.1174860>
- Haworth, K. (2018). Tapes, transcripts and trials: The routine contamination of police interview evidence. *The International Journal of Evidence & Proof*, *22*(4), 428-450. <https://doi.org/10.1177/1365712718798656>

- Higham, P. A. (1998). Believing details known to have been suggested. *British Journal of Psychology*, *89*(2), 265-283. <https://doi.org/10.1111/j.2044-8295.1998.tb02684.x>
- Higham, P. A. (2007). No Special K! A signal-detection framework for the strategic regulation of memory accuracy. *Journal of Experimental Psychology: General*, *136*(1), 1-22. <https://doi.org/http://dx.doi.org/10.1037/0096-3445.136.1.1>
- Higham, P. A., Blank, H., & Luna, K. (2017). Effects of postwarning specificity on memory performance and confidence in the eyewitness misinformation paradigm. *Journal of Experimental Psychology: Applied*, *23*(4), 417-432. <https://doi.org/10.1037/xap0000140>
- Higham, P. A., Neil, G. J., & Bernstein, D. M. (2017). Auditory Hindsight Bias: Fluency Misattribution Versus Memory Reconstruction. *Journal of Experimental Psychology: Human Perception and Performance*, *43*(6), 1144. <https://doi.org/10.1037/xhp0000405>
- Holliday, R. E., Douglas, K. M., & Hayes, B. K. (1999). Children's eyewitness suggestibility: Memory trace strength revisited. *Cognitive Development*, *14*(3), 443-462. [https://doi.org/10.1016/S0885-2014\(99\)00014-3](https://doi.org/10.1016/S0885-2014(99)00014-3)
- Holliday, R. E., Reyna, V. F., & Hayes, B. K. (2002). Memory processes underlying misinformation effects in child witnesses. *Developmental Review*, *22*(1), 37-77. <https://doi.org/10.1006/drev.2001.0534>
- Jacoby, L. L. (1991). A process dissociation framework: Separating automatic from intentional uses of memory. *Journal of Memory and Language*, *30*(5), 513-541. [https://doi.org/10.1016/0749-596X\(91\)90025-F](https://doi.org/10.1016/0749-596X(91)90025-F)
- Jacoby, L. L., Woloshyn, V., & Kelley, C. (1989). Becoming famous without being recognized: Unconscious influences of memory produced by dividing attention. *Journal of Experimental Psychology: General*, *118*(2), 115.

- Jamet, E., & Le Bohec, O. (2007). The effect of redundant text in multimedia instruction. *Contemporary Educational Psychology, 32*(4), 588-598.
<https://doi.org/https://doi.org/10.1016/j.cedpsych.2006.07.001>
- Johnson, M. K., Foley, M. A., & Leach, K. (1988). The consequences for memory of imagining in another person's voice. *Memory & Cognition, 16*(4), 337-342.
<https://doi.org/10.3758/BF03197044>
- Johnson, M. K., Hashtroudi, S., & Lindsay, D. S. (1993). Source monitoring. *Psychological Bulletin, 114*(1), 3.
- Kalyuga, S., Chandler, P., & Sweller, J. (2004). When Redundant On-Screen Text in Multimedia Technical Instruction Can Interfere With Learning. *Human Factors, 46*(3), 567-581. <https://doi.org/10.1518/hfes.46.3.567.50405>
- Kirk, E., Gurney, D., Edwards, R., & Dodimead, C. (2015). Handmade Memories: The Robustness of the Gestural Misinformation Effect in Children's Eyewitness Interviews [journal article]. *Journal of Nonverbal Behavior, 39*(3), 259-273.
<https://doi.org/10.1007/s10919-015-0210-z>
- Koriat, A., & Goldsmith, M. (1996). Monitoring and control processes in the strategic regulation of memory accuracy. *Psychological Review, 103*(3), 490-517.
<https://doi.org/10.1037/0033-295x.103.3.490>
- Kovacs, S. L., & Newcombe, N. S. (2006). Developments in source monitoring: The role of thinking of others. *Journal of Experimental Child Psychology, 93*(1), 25-44.
<https://doi.org/https://doi.org/10.1016/j.jecp.2005.06.006>
- Lange, N., Thomas, R., Dana, J., & Dawes, R. (2010). Contextual Biases in the Interpretation of Auditory Evidence. *Law and Human Behavior, 1-10*.
<https://doi.org/10.1007/s10979-010-9226-4>

- LaPaglia, J. A., & Chan, J. C. (2013). Testing increases suggestibility for narrative-based misinformation but reduces suggestibility for question-based misinformation. *Behavioral Sciences & The Law, 31*(5), 593-606. <https://doi.org/10.1002/bsl.2090>
- Leahy, W., & Sweller, J. (2016). Cognitive load theory and the effects of transient information on the modality effect. *Instructional Science, 44*(1), 107-123. <https://doi.org/10.1007/s11251-015-9362-9>
- Lewandowsky, S., Ecker, U. K., Seifert, C. M., Schwarz, N., & Cook, J. (2012). Misinformation and its correction: Continued influence and successful debiasing. *Psychological Science in the Public Interest, 13*(3), 106-131.
- Lindsay, D. S. (1990). Misleading suggestions can impair eyewitnesses' ability to remember event details. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16*(6), 1077. <https://doi.org/10.1037/0278-7393.16.6.1077>
- Lindsay, D. S., & Johnson, M. K. (1989). The eyewitness suggestibility effect and memory for source. *Memory & Cognition, 17*(3), 349-358. <https://doi.org/10.3758/BF03198473>
- Loftus, E. F. (1974). Reconstructing Memory: The Incredible Eyewitness. *Jurimetrics Journal, 15*(3), 188-193. <https://www.jstor.org/stable/29761487>
- Loftus, E. F. (1979). The malleability of human memory: Information introduced after we view an incident can transform memory. *American Scientist, 67*(3), 312-320. <https://www.jstor.org/stable/27849223>
- Loftus, E. F. (2005). Planting misinformation in the human mind: A 30-year investigation of the malleability of memory. *Learning & Memory, 12*(4), 361-366. <https://doi.org/10.1101/lm.94705>.

- Loftus, E. F., & Palmer, J. C. (1974). Reconstruction of automobile destruction: An example of the interaction between language and memory. *Journal of verbal learning and verbal behavior*, *13*(5), 585-589.
- Mammarella, N., & Fairfield, B. (2008). Source monitoring: The importance of feature binding at encoding. *European Journal of Cognitive Psychology*, *20*(1), 91-122.
<https://doi.org/10.1080/09541440601112522>
- McCloskey, M., & Zaragoza, M. (1985). Misleading postevent information and memory for events: Arguments and evidence against memory impairment hypotheses. *Journal of Experimental Psychology: General*, *114*(1), 1. <https://doi.org/10.1037/0096-3445.114.1.1>
- Miller, A. E. (2016). Jury suggestibility: the misinformation effect and why courts should care about inaccuracies in transcripts that accompany recorded evidence. *Law & Psychology Review*, *40*, 363.
- Moreno, R., & Mayer, R. (2002a). Learning science in virtual reality multimedia environments: Role of methods and media. *Journal of Educational Psychology*, *94*, 598-610. <https://doi.org/10.1037/0022-0663.94.3.598>
- Moreno, R., & Mayer, R. (2002b). Verbal redundancy in multimedia learning: When reading helps listening. *Journal of Educational Psychology*, *94*(1), 156.
- Nash, R. A., & Wade, K. A. (2009). Innocent but proven guilty: Eliciting internalized false confessions using doctored-video evidence. *Applied Cognitive Psychology*, *23*(5), 624-637. <https://doi.org/10.1002/acp.1500>
- Oeberst, A., & Blank, H. (2012). Undoing suggestive influence on memory: The reversibility of the eyewitness misinformation effect. *Cognition*, *125*(2), 141-159.
<https://doi.org/10.1016/j.cognition.2012.07.009>

- Pezdek, K., & Greene, J. (1993). Testing eyewitness memory: Developing a measure that is more resistant to suggestibility. *Law and Human Behavior, 17*(3), 361.
<https://doi.org/10.1007/BF01044514>
- Polak, M., Dukała, K., Szpitalak, M., & Polczyk, R. (2016). Toward a Non-memory Misinformation Effect: Accessing the Original Source Does Not Prevent Yielding to Misinformation [journal article]. *Current Psychology, 35*(1), 1-12.
<https://doi.org/10.1007/s12144-015-9352-8>
- Putnam, A. L., Sungkhasettee, V. W., & Roediger III, H. L. (2017). When misinformation improves memory: The effects of recollecting change. *Psychological Science, 28*(1), 36-46. <https://doi.org/10.1177/0956797616672268>
- Raj, V., & Bell, M. A. (2010). Cognitive processes supporting episodic memory formation in childhood: The role of source memory, binding, and executive functioning. *Developmental Review, 30*(4), 384-402.
- Remez, R., Vatikiotis-Bateson, E., Bailly, G., & Perrier, P. (2005). Three puzzles of multimodal speech perception. *Audiovisual Speech, 12-19*.
<https://doi.org/10.1017/CBO9780511843891.003>
- Rindal, E. J., DeFranco, R. M., Rich, P. R., & Zaragoza, M. S. (2016). Does reactivating a witnessed memory increase its susceptibility to impairment by subsequent misinformation? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 42*(10), 1544. <https://doi.org/10.1037/xlm0000265>
- Roberts, K. P., & Lamb, M. E. (1999). Children's responses when interviewers distort details during investigative interviews. *Legal and Criminological Psychology, 4*(1), 23-31.
- Schacter, D. L., Harbluk, J. L., & McLachlan, D. R. (1984). Retrieval without recollection: An experimental analysis of source amnesia. *Journal of Verbal Learning and Verbal Behavior, 23*(5), 593-611. [https://doi.org/10.1016/S0022-5371\(84\)90373-6](https://doi.org/10.1016/S0022-5371(84)90373-6)

- Schooler, J. W., & Loftus, E. F. (1986). Individual differences and experimentation: Complementary approaches to interrogative suggestibility. *Social Behaviour, 1*(2), 105-112. <https://doi.org/10.1006/jesp.1993.1003>
- Schul, Y. (1993). When warning succeeds: The effect of warning on success in ignoring invalid information. *Journal of Experimental Social Psychology, 29*(1), 42-62. <https://doi.org/10.1006/jesp.1993.1003>
- Skagerberg, E. M., & Wright, D. B. (2008). The co-witness misinformation effect: Memory blends or memory compliance? *Memory, 16*(4), 436-442. <https://doi.org/10.1080/09658210802019696>
- Smith, K. C., Multhaup, K. S., & Ihejirika, R. C. (2017). From Eyewitness to Academic Contexts: Examining the Effect of Misinformation in First and Second Languages. *Applied Cognitive Psychology, 31*(5), 546-557.
- Stevenage, S. V., Neil, G. J., & Hamlin, I. (2013). When the face fits: Recognition of celebrities from matching and mismatching faces and voices. *Memory, 22*(3), 284-294. <https://doi.org/10.1080/09658211.2013.781654>
- Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *The Journal of the Acoustical Society of America, 26*(2), 212-215. <https://doi.org/10.1121/1.1907309>
- Szpitalak, M., & Polczyk, R. (2010). Warning against warnings: Alerted subjects may perform worse. Misinformation, involvement and warning as determinants of witness testimony. *Polish Psychological Bulletin, 41*(3), 105-112.
- Thomas, A. K., Bulevich, J. B., & Chan, J. C. (2010). Testing promotes eyewitness accuracy with a warning: Implications for retrieval enhanced suggestibility. *Journal of Memory and Language, 63*(2), 149-157. <https://doi.org/10.1016/j.jml.2010.04.004>

Tousignant, J. P., Hall, D., & Loftus, E. F. (1986). Discrepancy detection and vulnerability to misleading postevent information. *Memory & Cognition*, *14*(4), 329-338.

<https://doi.org/10.3758/BF03202511>

Wright, D. B. (1993). Misinformation and warnings in eyewitness testimony: A new testing procedure to differentiate explanations. *Memory*, *1*(2), 153–166.

<https://doi.org/10.1080/09658219308258229>

Zaragoza, M. S., & Lane, S. M. (1994). Source misattributions and the suggestibility of eyewitness memory. *Journal of Experimental Psychology: Learning Memory & Cognition*, *20*(4), 934-945. <https://doi.org/10.1037/0278-7393.20.4.934>

Zaragoza, M. S., & Mitchell, K. J. (1996). Repeated exposure to suggestion and the creation of false memories. *Psychological Science*, *7*(5), 294-300.

<https://doi.org/10.1111/j.1467-9280.1996.tb00377.x>

Appendix A: Key phrases for Experiment 1

Murder video key phrases and preamble

Preamble - The accused (Fred Jackson) went to a New Year's Eve party and shortly before midnight, he got into an argument with the deceased (Sam). The altercation moved from the kitchen into the back garden, and witnesses say the two men continued to have a heated argument. Sam left the party shortly after and was later found dead at his home from a knife wound.

Critical phrases – with key words for both accurate and incriminating groups separated by /

- Do that again and I'll BILL/KILL you.
- I'm going to SHUT/CUT you up...
- I went to GRAB/STAB him.
- I was surprised at how MUDDY/BLOODY he was.
- ...a really, really terrible NIGHT/FIGHT.

Sexual assault video key phrases and preamble

John is accused of giving the Sarah a date-rape drug at a bar and then taking her to his home and sexually assaulting her. John saw Sarah at the bar on her own, walked up to her and sparked up a conversation. They later left for his home.

Critical phrases – with key words for both accurate and incriminating groups separated by /

- She was being kind of, quite flirty, umm, so I gave her a HUG/DRUG.
- ...forget him [boyfriend], I'm going to STEAL/FEEL her.
- ...she was very drunk, so I thought the best thing to do would be to, uh, get her FOOD/NUDE.
- I, uh, put her into...into the bed, and, uh, made sure that she, uh, that I STAYED/LAID with her.
- I sort of wondered if she had remembered, sort of, what IT'D/I'D done to her.

Appendix B: Key phrases for Experiment 2

Preamble - The accused (Fred Jackson) went to a New Year's Eve party and shortly before midnight, he got into an argument with the deceased (Sam). The altercation moved from the kitchen into the back garden, and witnesses say the two men continued to have a heated argument. Sam left the party shortly after and was later found dead at his home from a knife wound.

Critical phrases – with key words for what was actually said and the incriminating replacement separated by /

- To be honest, I really wanted to have a good NIGHT/FIGHT at the party.
- I lost my balance, broke my beer bottle, and SPLASHED/SLASHED him.
- I was trying to protect myself so, Um and I was like, you know, I'm gonna SHUT/CUT you up.
- I went to GRAB/STAB him.
- I was surprised to see how MUDDY/BLOODY he was.