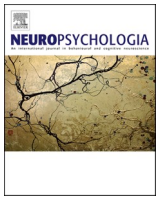
Neuropsychologia 225 (2026) 109428

Contents lists available at ScienceDirect

Neuropsychologia

journal homepage: www.elsevier.com/locate/neuropsychologia



# Faster initial retrieval of misinformation corrections predicts better long-term memory: An ERP study

Sean Guo [a], Danni Chen [a], Wanrou Hu [a,b], Xiaoqing Hu [a,c,*]

[a] Department of Psychology, The University of Hong Kong, Hong Kong SAR, China
[b] School of Life Science and Technology, University of Electronic Science and Technology of China, Chengdu, China
[c] HKU-Shenzhen Institute of Research and Innovation, Shenzhen, China

ARTICLE INFO

*Keywords:*
Misinformation
Encoding
Working memory
Belief updating

ABSTRACT

The ability to effectively encode corrections to misinformation is vital for making well-informed decisions. However, false information often influences people's beliefs and judgments even after it is corrected. Here, we employed the electroencephalogram (EEG) to examine neurocognitive processes when participants encoded events, their causes, and corrections/affirmations to said causes. Participants then judged the veracity of the causes immediately (day 1) and one day after (day 2). Behaviorally, faster correct judgements on day 1 predicted more accurate day 2 delayed veracity judgements, suggesting that measuring response times could identify weakly encoded corrections. Re-exposure to misinformation on day 2 did not modulate subsequent memory for corrections. Although we did not observe any subsequent memory effects, we found reduced P300 amplitudes for corrections than affirmations. When only considering subsequent accurate veracity judgements, we observed a greater frontal slow wave amplitude when encoding corrections than affirmations. These findings provide preliminary neural evidence that processing corrections may be more difficult than affirmations.

As the distribution and proliferation of misinformation in society becomes more prevalent, understanding how to effectively correct misinformed beliefs and the relationship between memory and belief is of paramount importance (Ecker et al., 2022). Causal misinformation can be especially damaging because cause-effect relationships scaffold our interpretation of the world and influence how we ascribe blame (Cushman, 2008; Lombrozo, 2010). Even if outdated causal information is updated, it can still exert long-lasting impacts on decision-making, a phenomenon known as the continued influence effect (CIE, Chan et al., 2017; Ecker et al., 2022; Wilkes and Leatherbarrow, 1988). Failure to encode or integrate corrections with misinformation may be a driving force of the CIE (Ecker et al., 2022; Kendeou et al., 2019), and even when misinformation is successfully corrected at first, misinformation belief may increase after a delay (Rich and Zaragoza, 2020; Swire-Thompson et al., 2023). Here, we use electroencephalography (EEG) to examine neural processes underlying encoding of corrections to misinformation and relate these processes to subsequent memory for corrections. After a delay, we examine whether repetition of corrected misinformation leads to increased belief (i.e., the illusory truth effect).

To facilitate understanding, we provide definitions of key words that are referenced throughout the study here. Narratives in CIE paradigms typically describe a specific scenario, which we refer to as an *event* (e.g., a house is on fire). These events typically occur due to a specific reason, which we refer to as a *cause* (e.g., the cause of the fire was a stove). Sometimes, the purported cause of the event can be retracted, and we refer to these retractions as *corrections* (e.g., the cause of the house fire was not a stove). Otherwise, causes can also be confirmed, and we refer to these confirmations as *affirmations* (e.g., the cause of the house fire was a stove).

## 1. Neural correlates of the CIE

Gaining insights into neurocognitive activity underlying the processes of encoding corrections to causal misinformation can help inform interventions aimed at reducing the CIE. Previous neural investigations of the CIE revealed that encoding corrections was associated with reduced brain activity implicating information integration and reading comprehension, such as the activity in the precuneus, posterior cingulate cortex, and left middle temporal gyrus regions, relative to encoding affirmations (Gordon et al., 2017; Jin et al., 2022). This finding suggests that corrections may be more difficult to process than affirmations, and is consistent with accounts that failure to associate misinformation with

* Corresponding author. Department of Psychology, The University of Hong Kong, Pokfulam, Hong Kong SAR, China.
*E-mail address:* xiaoqinghu@hku.hk (X. Hu).

https://doi.org/10.1016/j.neuropsychologia.2026.109428
Received 19 October 2025; Received in revised form 3 February 2026; Accepted 6 March 2026
Available online 10 March 2026

its correction can lead to errant belief in misinformation (Ecker et al., 2022; Gordon et al., 2017; Kendeou et al., 2014). Although brain-behavior correlation can provide hints to underlying processes between subjects, insufficient trials for each participant mean that these findings are silent regarding whether within-subject brain activity during encoding differs between successful and unsuccessful correction recollection. Increasing the number of trials can help provide this evidence and delineate the mechanisms behind effective corrections of causal misinformation.

Examining the subsequent memory effect (SME) can reveal neural activity associated with successful and unsuccessful recollection of corrections. In SME studies, trials during encoding are typically sorted based on subsequent memory performance (remembered or forgotten). Mounting research has identified an early parietal positive deflection known as the P300 and a late positive deflection in the frontal region known as the frontal slow wave (FSW) as important components in the SME (Mecklinger and Kamp, 2023).

The P300 SME is often elicited in studies examining detail-oriented memory for words and images, with greater amplitudes indicating more detailed processing of a stimulus (Forester and Kamp, 2023; Gonsalves and Paller, 2000; Kamp et al., 2017). For example, a larger P300 amplitude while encoding images was linked to better subsequent recognition of the images (Forester and Kamp, 2023). The P300 SME is proposed to support distinctive stimulus encoding and the unitization of stimulus components, such as encoding two different words into one coherent representation (Kamp et al., 2017). Whereas the P300 indexes detailed item encoding, the FSW predicts subsequent associative memory between pairs of words or images, with more positive amplitudes predicting greater associative strength between items (Forester and Kamp, 2023; Kamp et al., 2016, 2017; Kim et al., 2009). The FSW SME is proposed to reflect working memory processes involved in associative encoding, independent of whether familiarity or recollection serve as primary strategies for retrieval (Kamp et al., 2017). For example, when participants had to remember pairs of images, greater FSW amplitude while viewing the second image in the pair predicted stronger associations within the pair (Forester and Kamp, 2023).

Due to the important role of encoding corrections for misinformation belief (Ecker et al., 2022; Gordon et al., 2017; Guo et al., 2025), one of our primary objectives was to examine whether P300 and FSW amplitudes during encoding were associated with subsequent memory for corrections. We had no directional hypotheses about how memory for corrections would relate to the P300 and FSW when participants encoded events and causes. However, we predicted that larger P300 and FSW amplitudes when encoding corrections would reflect improved encoding of corrections and stronger misinformation-correction pairings and thus be linked to improved memory for corrections.

Even if SMEs are not observed, P300 and FSW amplitudes can provide information about mental processes that differ between corrections and affirmations. Past work has shown that attenuated P300 amplitudes index cognitive effort and working memory load (Scharinger et al., 2017), greater recruitment of attentional resources (Polich, 2007), suggesting that more difficult cognitive tasks are associated with reduced P300 amplitude (Ghani et al., 2020). Examining the P300 may thus also shed light onto the relative difficulty of corrections and affirmations, extending previous fMRI work (Gordon et al., 2017; Jin et al., 2022). Although our earlier misinformation study found smaller P300 amplitudes when participants were tasked with updating misinformation with an alternative explanation compared to when no updating was required, repetition of key phrases confounded repetition with updating, which hindered interpretation (Guo et al., 2025). In this study, we examine a similar comparison between corrections and affirmations while controlling for word repetition effects.

Outside of indexing associative memory strength in SMEs, FSW amplitudes have been known to fluctuate with working memory load (Monfort and Pouthas, 2003), cognitive control (West and Travers, 2008), prediction errors (Van Petten and Luka, 2012), and negation processes (Herbert and Kissler, 2014). Examining the FSW outside of the SME context could thus provide hints to the underlying mechanisms that differentiate processing of corrections and affirmations, although precise processes will need to be identified by future research.

### 1.1. Delayed memory for corrections to causal misinformation

Even if people remember corrections immediately after receiving them, belief in corrected misinformation has been shown to increase over time (Carey et al., 2022; Rich and Zaragoza, 2020), and memory for corrections diminishes over time (Swire-Thompson et al., 2023). Increased belief in misinformation after a delay could thus be due to corrections becoming inaccessible or forgotten (Kemp et al., 2024; Swire-Thompson et al., 2023). Stronger memories have been linked to greater resistance to forgetting over time (Radvansky et al., 2022), suggesting that stronger misinformation-correction associations could be associated with improved correction memory durability. One way to measure this association may be through response times. Faster response times (RTs) during recall have been linked to more confident recollection (i.e., "Remember" rather than "Know" judgements; Dewhurst et al., 2006; Wixted and Stretch, 2004) and stronger associative memories between pairs of stimuli (Craddock et al., 2012; Eimas and Zeaman, 1963; Kounios et al., 2001). Because corrections are often linked to misinformation via associative processes (Ecker et al., 2011; Gilbert et al., 1990, 1993), we anticipate that RTs could potentially reflect how strongly corrections are associated with misinformation. In our study, we therefore examine how correction and affirmation memory RTs are associated with memory for corrections after a delay.

Given that strong associations between misinformation and its correction have been linked to reduced misinformation belief (Ecker et al., 2011), we hypothesized that if faster response times reflected stronger misinformation-correction associations, then faster correct memory for corrections would be associated with improved memory for those corrections after a delay.

### 1.2. Simultaneous recognition and exposure paradigm

In addition to studying whether RTs were associated with memory for corrections, we were also interested in whether memory for corrections was affected by re-reading misinformation. Outside the laboratory, people may be exposed to the same misinformation multiple times (Weng et al., 2024). According to the illusory truth effect, merely reading a statement increases its fluency, which can then be misattributed to trustworthiness (Fazio et al., 2015; Gilbert et al., 1990). Although the illusory truth effect suggests that repeated exposure to misinformation may increase belief, it remains unclear whether this effect persists when the misinformation has already been corrected (i.e., encountering misinformation after reading a debunking statement). Importantly, this question is distinct from the so-called 'backfire effect,' which has generally failed to replicate (Wood and Porter, 2019).

To examine this in greater detail, we designed a new task that manipulated whether participants saw intact misinformation (defined here as an event-cause pair, e.g., house on fire-stove). This differed from how the illusory truth effect is typically elicited (i.e., by repeating statements or headlines, Fazio et al., 2015), but was more consistent with how misinformation was initially presented to participants in our experiment. The instructions for this new task were only to identify whether the stimulus presented on screen was previously presented in the experiment (i.e., an old/new recognition task). However, unbeknownst to participants, certain cause-event pairs were presented intact while others were presented separately (i.e., the cause was presented 20 trials after the event). This allowed us to examine whether viewing intact cause-event pairs would increase the accessibility of misinformation and potentially increase reliance on more heuristic-based automatic processes during retrieval.

We were unsure of the direction of this effect. On one hand, seeing

intact misinformation can increase its fluency and strengthen cause-event associations, and this fluency could be misattributed to truth. On the other hand, the fluency and strength of cause-event associations may not influence memory for corrections if people primarily rely on cause-correction associations to recall corrections.

### 1.3. Experiment overview

To examine the neural correlates underlying the encoding of misinformation and corrections, we modified the traditional CIE paradigm for the EEG. On the first day of the experiment, participants first learned a series of events, causes and corrections or affirmations to the causes, and their memory for the veracity of the causes was subsequently tested. On the second day of the experiment, participants completed a recognition task that simultaneously exposed them to intact or separated cause-event misinformation pairs (i.e., simultaneous recognition and exposure task), and their memory for corrections and affirmations was tested identically to the first day of the experiment.

## 2. Method

### 2.1. Participants

We recruited 61 university students (49 female; $M = 22.67$ years, $SD = 2.71$ years). Participants from the age of 18 to 35 were recruited. They reported normal or corrected-to-normal vision, no colorblindness, no chronic medical conditions, and no history of severe mental illness, neurological disorders, or sleep disorders. Because the experiment presented textual stimuli in Chinese, only native Chinese speakers were recruited. Participants received monetary compensation (at approximately 10 USD/hour) for participation. Participants were recruited through mass emails sent to university students. Participants were excluded from encoding EEG and veracity judgement behavioral analysis due to the following reasons: lower than chance veracity judgement accuracy (less than 50%, $n = 4$), failure to follow experiment instructions ($n = 6$), and withdrawal from the experiment ($n = 2$). Participants were further excluded from EEG analysis due to technical issues during recording ($n = 1$). We reported behavioral analyses from 51 participants for the veracity judgement tasks on Day 1 and 2, and 49 participants for the simultaneous recognition and exposure task. We reported EEG encoding analyses from 50 participants.

### 2.2. Design

The experiment used a 2 (Cause veracity: true vs. false) × 2 (Day 2 exposure type: Intact vs. Separated) within-participants design. Participants provided veracity judgements in two time points (Day 1, Day 2).

### 2.3. Experimental materials

One hundred and eighty scenarios were selected from the materials pilot for the experiment. Each scenario consisted of a cause word paired with an image depicting an event. For example, an image of a house on fire was paired with the cause “stove”, suggesting that the fire was caused by a stove in the house. The length of each cause was controlled at two Chinese characters. Images were obtained online.

Two hundred and forty-seven unique scenarios were created for pretesting. The goal of the pretest was to remove any stimuli that had high positive or negative emotional valence, cause-event pairings that were too strongly or weakly related, or causes that were too highly believable or unbelievable. We recruited ten university student volunteers. Each volunteer rated the emotional valence of images and their respective causes, relatedness between the image and cause, and how believable the cause-event relationship was on a scale from 1 (low/negative) to 7 (high/positive). The final images ($M = 3.56$, $SD = 0.75$) and cause words ($M = 3.60$, $SD = 0.87$) were rated as moderately emotional. The causes and events were judged to be moderately related to one another ($M = 4.92$, $SD = 1.01$). The believability of cause-event relationships was relatively high ($M = 4.79$, $SD = 1.05$). For details on how we removed pretesting stimuli, please see Supplementary Material A.

### 2.4. Experimental procedure

The experiment was administered using PsychoPy version 2020.2.10 (Peirce et al., 2019) across two days. On the first day (“Day 1”) of the experiment, participants began with a familiarization task, indicating their familiarity to a series of words that would be used in the following task. Then, they completed the encoding task, in which events (presented as images), causes (presented as words), and corrections or affirmations were presented. They were told to imagine scenarios that linked causes to events, and to remember the correction or affirmation. Afterward, they completed a 3-back task as a distractor. Finally, they completed a veracity judgement task in which they had to determine whether causes from the encoding task were corrected or affirmed. On the second day (“Day 2”), participants first completed the simultaneous recognition and exposure task, followed by another veracity judgement task. Causes and veracities were presented in Simplified Chinese. Next, we provide a detailed description of each of the tasks.

#### 2.4.1. Familiarization task (Day 1)

In the task, participants saw a series of words on a computer screen (max 3s each) and indicated whether they understood the definition of the word using the ‘a’ or ‘l’ keys that represented ‘understand’ and ‘do not understand’, counterbalanced between participants. Outside of testing comprehension, the task also served to attenuate novelty-related EEG responses. Trials containing unrecognized words were excluded from later behavioral and EEG analysis. Participants would have been excluded from the experiment if they scored lower than 80%. However, no participants scored below the threshold ($M = 98.9\%$, $SD = 2.3\%$).

#### 2.4.2. Encoding task (Day 1)

Participants were told that they would encounter a series of images and words, that each image would represent an event, and that each word would represent the cause of the event. They were asked to provide vividness ratings for each cause-event pair (i.e., the extent to which they could imagine how the cause led to the outcome depicted in the event), and were also told not to make preemptive judgements of whether the cause seemed true or false. Next, they were told that they would see corrections or affirmations to the cause. They were asked to remember the event, cause, and correction or affirmation for a subsequent memory test. Participants completed two practice trials to familiarize themselves with the procedure.

In each trial, participants began by viewing an image depicting an event (e.g., a house on fire) in the center of the screen (2s). After a fixation cross (0.3-0.5s), they viewed a word indicating the cause of the previous event (e.g., a stove) for 2 s. Participants then had up to 3 s to rate how vividly they could imagine the relationship between the cause and the event, using the 1 (low) to 5 (high) number keys. Following another fixation cross (0.3-0.5s), they were presented with a correction or affirmation. In the affirmation condition, participants saw the word “True”. In the correction condition, they saw the word “False” (presented as a Simplified Chinese character). The inter-trial-interval was set at 1 s (Fig. 1b). There were 180 unique scenarios (each consisting of an event, cause, and correction or affirmation) which were evenly and randomly split between affirmation and correction conditions. Each scenario appeared twice throughout the task in different halves (i.e., first occurrence in the former 180 trials, second in the latter 180 trials), resulting in a total of 360 trials. Participants took a self-paced break every 20 trials.

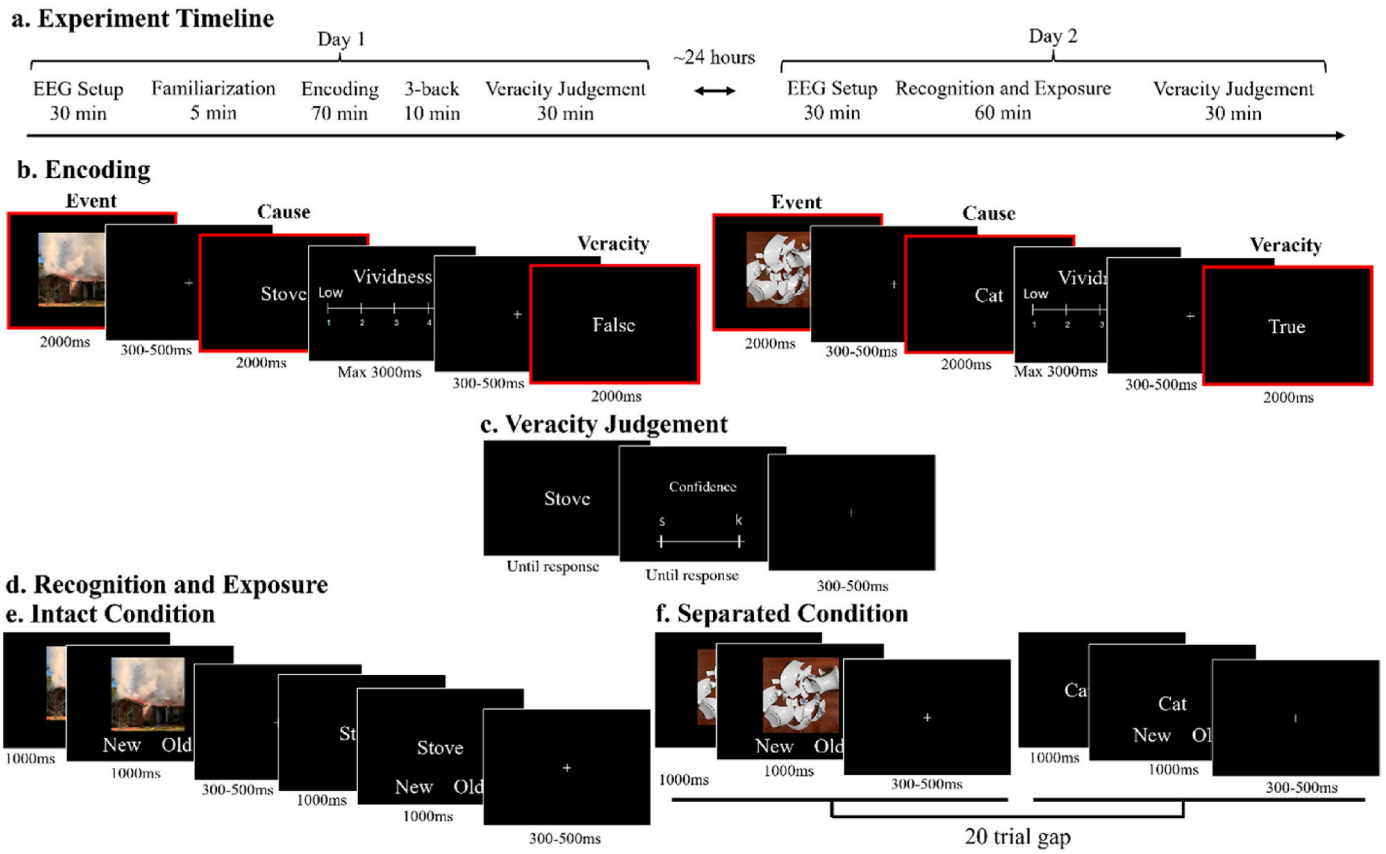



**Fig. 1.** *Procedure Schematic and Example Trials*

*Note.* **a)** Experiment timeline showing the order of task administration and duration of each task. **b)** Exemplar encoding trials in the correction condition. Each trial began with an image depicting an event. Then, the cause of the event was shown as a word. After vividness rating, the veracity of the cause was shown (left, false condition; right, true condition). Red outlines indicate ERP epochs of interest. **c)** Exemplar veracity judgement trial**.** Each trial began with a cause word from an earlier encoding trial, displayed until a response was made, followed by a confidence rating ('s' and 'k' indicate which keys to press for low and high confidence respectively). **d)** Participants completed a recognition and exposure task in which they made old/new judgements to words and images. Unbeknownst to participants, this was divided into the Intact (**e**) and Separated (**f**) conditions. In the intact condition, events and causes were presented in their original order. In the separated condition, events and causes were separated by a 20 trial gap. Text was translated to English from Simplified Chinese for this figure. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

### *2.4.3. 3-Back task (Day 1)*

After the encoding task, participants completed a 3-back working memory task (approximately 10 min long) in which they had to determine if the current number shown on the screen was the same or different to the number shown three trials ago. There were 240 trials, and participants took a self-paced break every 60 trials.

### *2.4.4. Veracity judgement (Day 1)*

Participants were told that they would see a series of causes from the encoding task, and were instructed to try and remember, quickly and accurately, whether that cause was corrected or affirmed. Participants completed two practice trials before the task to familiarize themselves with the procedure. In each trial, a cause word from the encoding task was shown on screen. Participants had unlimited time to indicate using the 'a' or 'l' keys (counterbalanced between participants) whether they thought the cause was previously corrected or affirmed. After participants responded, they rated how confident they were in their answer using the 's' (low confidence) or 'k' (high confidence) keys without a time limit. Finally, a fixation cross (0.3-0.5s) was presented before the next cause word appeared on screen (Fig. 1c). Each cause word was presented twice in different halves of the task (i.e., first occurrence in the former 180 trials, second in the latter 180 trials), for a total of 360 trials. Response times (RTs) were recorded.

### *2.4.5. Simultaneous recognition and exposure task (Day 2)*

On Day 2 of the experiment, participants completed a recognition task, in which they were instructed to determine if the images or words that appeared on screen had previously been seen in the experiment on Day 1. For each trial, a single stimulus (image or cause word) was presented in the center of the screen (1s) followed by a text prompt (1s) signifying that a response could be made. When this text prompt appeared, participants indicated whether they thought the stimulus was old or new with the 'a' or 'l' keys, counterbalanced between participants. A fixation cross was presented between trials (0.3-0.5s). All images and cause words from Day 1 were shown twice, resulting in 720 trials containing old stimuli. An additional 360 trials of new stimuli contained new stimuli (i.e., lures, 180 new words and 180 new images), created to resemble old stimuli both visually and semantically. A break was given to participants every 180 trials.

Unbeknownst to the participants, the recognition task included two experimental conditions. In the Intact condition (Fig. 1e), cause words were presented immediately after their corresponding event images (i.e., preserving the original event-cause order from Day 1). In the Separated condition (Fig. 1f), cause words were presented after a temporal lag (i.e., 20 trials after the corresponding event image). We included the Separated condition to ensure that all stimuli were presented an equal number of times to participants to control for absolute levels of stimulus familiarity. Because we were primarily interested in false causes and corrections, false causes trials were classified into three accuracy

categories: Day 1 correct with fast RT (at or below the participant's median RT), Day 1 correct with slow RT (above the median RT), and Day 1 incorrect. The cause-event pairs in these categories were split evenly and randomly between Intact and Separated conditions. For true causes, trials were classified into two categories: Day 1 correct and Day 1 incorrect, and were evenly and randomly assigned to Intact and Separated conditions.

#### 2.4.6. Veracity judgement (Day 2)

Immediately after the simultaneous recognition and exposure task, participants completed a veracity judgement task identical to the veracity judgement task on Day 1.

### 2.5. EEG processing

#### 2.5.1. EEG acquisition

We recorded continuous EEG data using a 64-channel Waveguard cap linked to an EEGO amplifier (10/20 system; ANT Neuro, Enschede, Netherlands). The online sampling rate was 500Hz. The ground electrode was set at AFz, and the online reference was set at CPz. A horizontal electrooculogram (EOG) was placed 1.5 cm below the left canthus. Before beginning EEG recording, impedance of all electrodes was below 20 kΩ. EEG recordings were obtained on both Day 1 and 2. However, due to potential confounds with response time, EEG data for the veracity judgement tasks are not analyzed, and we focus on EEG during the encoding task only.

#### 2.5.2. EEG preprocessing

EEG data were preprocessed with MATLAB 2021b, EEGLAB 2022.0 and ERPLAB 9.00 (Delorme and Makeig, 2004; Lopez-Calderon and Luck, 2014). Three electrodes (EOG, M1, M2) were removed before analysis, as EOG electrodes were used solely for artifact monitoring and correction, and M1/M2 electrodes served as references during data acquisition. The raw data were bandpass filtered between 0.1 and 40Hz using EEGLAB's zero-phase FIR filter. To remove any line noise, a 50Hz notch filter was applied. Channels that displayed abnormally high EEG amplitudes or inactivity throughout the experiment were visually detected and removed before interpolating ($M$ = 4.00, $SD$ = 2.66 channels removed). Segments containing large and persistent amplitudes of noise (e.g., broadband deflections inconsistent with physiological EEG activity) were then visually detected and removed from the continuous data. Stimuli (including image, cause or veracity) that were removed did not differ between true ($M$ = 28.4, $SE$ = 4.7 stimuli) and false ($M$ = 30.1, $SE$ = 4.70 stimuli) conditions, $t(49) = -1.26$, $p = .212$. To facilitate Independent Component Analysis (ICA), a 1Hz high pass filter was applied and the data were downsampled to 250Hz. ICA was performed on the continuous data, and resulting independent components were labeled with the ICLabel toolbox (Pion-Tonachini et al., 2019). Components that contributed highly to eye or muscle movements were identified, and corrections were applied onto the dataset prior to ICA ($M$ = 3.82, $SD$ = 1.62 components removed). After re-referencing the data to the common average across all electrodes, epochs from −200ms to 2000ms were created. Automatic artifact rejection was applied to the epoched data on regions of interest (ROI) - any epoch containing peak to peak differences over 100 μV (sliding time window of 200ms, step size of 100ms), or amplitudes greater than ±75 μV was rejected. ERPs were baseline corrected from −200 to 0ms (stimulus onset).

#### 2.5.3. ERP quantifications

The FSW channels (Fz, F1, F2, F3, F4) and time window (1000-2000ms) were based on the FSW's description as a frontal and central component that manifested at around 1000ms (Forester et al., 2020; Forester and Kamp, 2023; Kamp et al., 2017; Mecklinger and Kamp, 2023), and a visual inspection of grand mean waveforms (Fig. S1) supported this characterization. The P300 channels (CP1, CP2, Pz, P3, P4) was similarly defined based on prior descriptions of a centroparietal positivity (Forester and Kamp, 2023; Kamp et al., 2017; Mecklinger and Kamp, 2023). However, examination of the grand mean waveforms (Fig. S1) revealed that the positivity began at around 200ms, earlier than previously reported latencies (Mecklinger and Kamp, 2023). Therefore, we adapted the time window here to 200-500ms to capture most of the parietal positivity. Adaptive means were calculated as the mean amplitude spanning 50ms in each direction of the peak value within these predefined time windows (Nielsen and Gonzalez, 2020). The minimum number of trials required in each condition was 8, and all participants fulfilled this requirement (Moran et al., 2013).

## 3. Results

We first present behavioral results followed by ERP results. Standard errors of the mean (SE) have been corrected for within-subjects comparisons for all following results (Morey, 2008).

### 3.1. Day 1 veracity judgements

During veracity judgement on Day 1, each stimulus appeared two times in total. Responses were only counted as correct if both instances of the stimuli were answered correctly. Otherwise, the response was counted as incorrect.

First, we characterized the effects of corrections and affirmations on memory. A paired $t$-test showed that accuracy for true causes ($M$ = 66.4%, $SE$ = 1.9%) was significantly greater than for false causes ($M$ = 59.4%, $SE$ = 1.9%), $t(50) = 2.54$, $p = .014$, $d = 0.35$ (Fig. 2a), suggesting that affirmations were remembered to a greater extent than corrections.

We next examined RTs during veracity judgments. There was a significant main effect of accuracy, $F(1,50) = 53.5$, $p < .001$, $\eta_p^2 = 0.517$, such that correct responses ($M$ = 2.36s, $SE$ = 0.08s) had faster RTs than incorrect responses ($M$ = 3.19s, $SE$ = 0.08s). There was no main effect of veracity, $F(1,50) = 1.3$, $p = .262$, $\eta_p^2 = 0.025$. Importantly, there was a significant interaction between accuracy and veracity, $F(1,50) = 13.8$, $p < .001$, $\eta_p^2 = 0.216$. Post hoc tests revealed that for correct answers, responses for true causes ($M$ = 2.22s, $SE$ = 0.04s) were faster than for false causes ($M$ = 2.50s, $SE$ = 0.04s), $t(50) = 4.52$, $p < .001$, $d = 0.37$. However, there was no difference for incorrect answers between true ($M$ = 3.27s, $SE$ = 0.07s) and false causes ($M$ = 3.11s, $SE$ = 0.07s), $t(50) = 1.74$, $p = .087$, $d = 0.13$ (Fig. 2b). Results suggest that participants took longer to accurately remember corrections than affirmations.

### 3.2. Day 1 confidence responses

Regarding confidence responses, there was a significant main effect of accuracy, $F(1,50) = 150.8$, $p < .001$, $\eta_p^2 = 0.751$, such that correct answers ($M$ = 73.9%, $SE$ = 1.8%) had a larger proportion of high confidence ratings than incorrect answers ($M$ = 42.5%, $SE$ = 1.8%). There was no main effect of veracity, $F(1,50) = 2.0$, $p = .163$, $\eta_p^2 = 0.039$. There was a significant interaction, $F(1,50) = 10.3$, $p = .002$, $\eta_p^2 = 0.171$. Post hoc tests revealed that for correct answers, true causes ($M$ = 77.0%, $SE$ = 2.2%) had significantly more high confidence responses than false causes ($M$ = 70.7%, $SE$ = 2.2%), $t(50) = 2.07$, $p = .044$, $d = 0.28$. For incorrect answers, false causes ($M$ = 47.3%, $SE$ = 1.7%) had more high confidence responses than true causes ($M$ = 37.7%, $SE$ = 1.7%), $t(50) = 4.01$, $p < .001$, $d = 0.46$ (Fig. 2c). These results show that correct memory of affirmations was associated with greater confidence than memory for corrections, while incorrectly remembering affirmations to false causes was associated with higher confidence than incorrectly remembering corrections to true causes.

To examine the association between confidence and RT, we used a linear mixed-effect model with log RT as the outcome variable, fixed effects of confidence and veracity, and random intercept effects of participant and scenario (Table S1). Confidence was negatively associated with RT ($b = -0.40$, $SE$ = 0.01, $p < .001$), while veracity was not

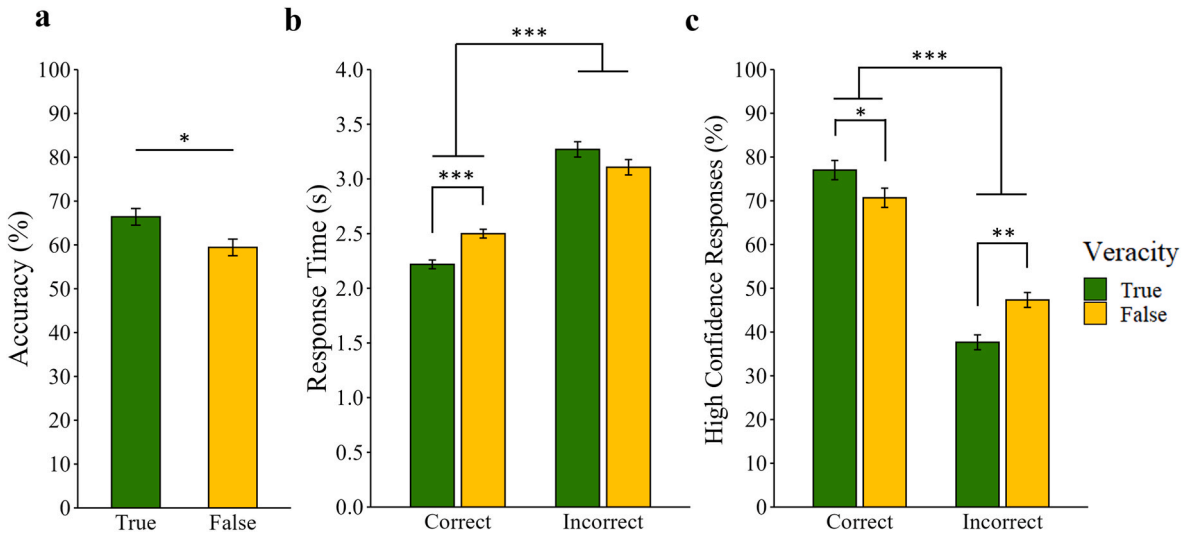



**Fig. 2.** Day *1 Behavioral Results*
*Note.* **a)** Veracity judgement accuracy for true and false causes on Day 1. **b)** response time and **c)** percentage of 'high confidence' responses in correct or incorrect answers on Day 1, split by true and false causes. Error bars denote standard error. $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$.

associated with RT ($b = 0.02$, $SE = 0.02$, $p = .367$). There was a significant interaction between confidence and veracity ($b = -0.14$ $SE = 0.02$, $p < .001$). Simple slopes analysis showed that the effect of confidence on RT was more negative for true causes ($\beta = -0.47$, $SE = 0.02$) than false causes ($\beta = -0.33$, $SE = 0.02$), $z = 6.39$, $p < .001$. These results suggest that while greater confidence is related to faster RTs, confidence was more strongly related to RT for true causes.

### 3.3. Recognition accuracy

We next examined recognition accuracy during the simultaneous recognition and exposure task as a manipulation check. If cause words in the Intact condition had a higher recognition accuracy than those in the Separated condition, this would indicate that presenting event images before causes facilitated processing of the causes, suggesting that participants were more likely to comprehend the misinformation as an event-cause unit. During the simultaneous recognition and association task, events and causes were shown twice in total. Responses were counted as correct only when both occurrences were correct.

A 2 (Day 2 exposure type: Intact, Separated) × 2 (Cause veracity: true, false) ANOVA revealed a significant main effect of exposure type, $F(1,48) = 110.3$, $p < .001$, $\eta_p^2 = 0.697$. Intact conditions had significantly higher accuracy than Separated conditions (Fig. 3a). There was also a significant main effect of veracity, $F(1,48) = 4.2$, $p = .046$, $\eta_p^2 = 0.080$, such that true cause words had higher recognition accuracy compared to false cause words. There was also a significant interaction between exposure type and veracity, $F(1,48) = 4.3$, $p = .044$, $\eta_p^2 = 0.082$. Post hoc tests revealed that Intact true causes ($M = 93.1\%$, $SE = 0.9\%$) had higher accuracy than Intact false causes ($M = 90.9\%$, $SE = 1.1\%$), $t(48) = 3.49$, $p = .001$, $d = 0.32$. There were no significant differences between Separated true ($M = 80.9\%$, $SE = 1.5\%$) and Separated false ($M = 80.8\%$, $SE = 1.4\%$) causes, $t(48) = 0.05$, $p = .963$, $d = 0.0$. This suggests that our manipulation was successful and that event images facilitated recognition of true causes more than false causes.

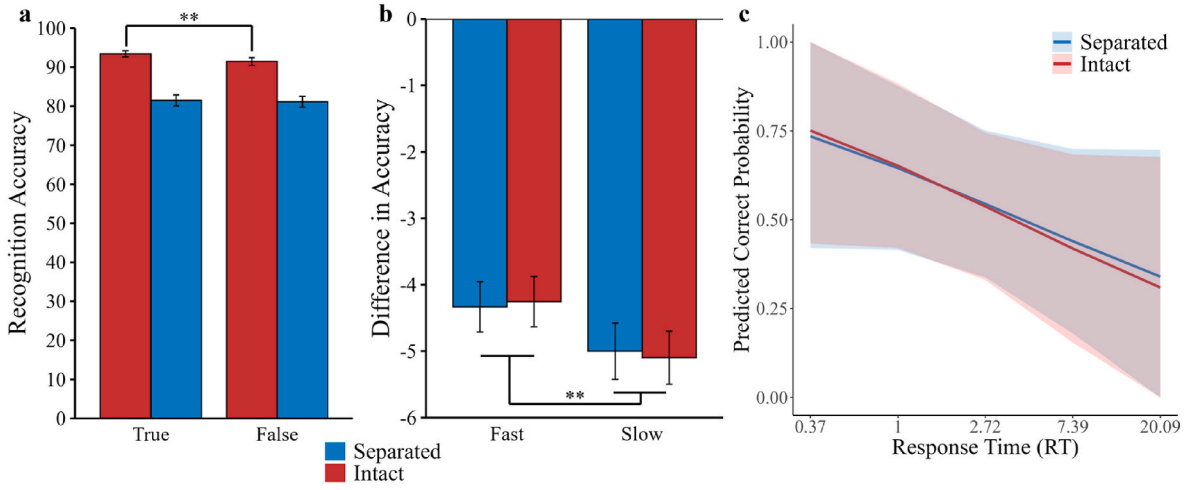



**Fig. 3.** Day *2 Recognition Accuracy and Difference in Accuracy*
*Note.* **a)** Cause recognition accuracy in the Intact and Separated conditions, split across true and false causes. **b)** Difference in accuracy between Day 1 and Day 2 for false causes assigned to intact or separated presentation conditions. Fast: below or equal median reaction time, Slow: above median reaction time when answering correctly during retrieval Day 1, calculated individually for each participant. **c)** predicted probability of retaining correct memory for correction on day 2 based on day 1 response time. Although RT was log-transformed, the x-axis shows actual (untransformed) response time values for easier interpretation. Error bars and shaded regions denote standard error. $^{**}p < .01$.

### 3.4. Difference in memory for corrections between day 1 and day 2

We next examined the relationship between cause-correction associations and memory for corrections, and whether seeing an intact misinformation pairing in the simultaneous recognition and exposure task modulated this relationship. To examine how memory for corrections changed from Day 1 to Day 2, we calculated the difference in accuracy between days. This was calculated by subtracting the number of items correctly answered on Day 1 from the number of items correctly answered on Day 2 that were also correct on Day 1. This can be represented as $Correct_{Day2} \cap Correct_{Day1} - Correct_{Day1}$. For example, if a participant correctly answered 10 items on Day 1, but only correctly answered 5 of those 10 items on Day 2, their difference in accuracy was −5.

To examine the relationship between cause-correction associations and delayed memory for corrections, we divided correct Day 1 veracity judgements into slow (RT above the median) and fast (RT below or equal to the median) trials for each participant individually (see Table S2 for the mean and SD of fast, slow and incorrect judgements). We further divided the analysis into Intact and Separated conditions, resulting in four conditions (i.e., fast Intact, fast Separated, slow Intact and slow Separated) for false causes. We did not consider RTs for true causes and thus do not report these results.

A 2 (Day 2 exposure type: Intact, Separated) × 2 (RT: fast, slow) ANOVA showed no main effect of exposure type, $F(1,50) = 0.00$, $p = .967$, $\eta_p^2 = 0.00$, and no significant interaction between exposure type and Day 1 response time, $F(1,50) = 0.14$, $p = .709$, $\eta_p^2 = 0.003$. However, there was a main effect of Day 1 response time, $F(1,50) = 9.49$, $p = .003$, $\eta_p^2 = 0.159$ (Fig. 3b). This suggests that faster response times on Day 1 resulted in a less negative difference score (i.e., better accuracy on Day 2) compared to slow Day 1 response times, but there was no difference between Intact and Separated conditions.

Although difference scores are widely used to examine memory change (Roheger et al., 2020), they may also be unreliable (Edwards, 2001). To address this potential concern, and to examine whether using continuous RT data would result in a similar pattern of results, we used a logistic mixed-effects regression model (glmer function, lme4 package, Bates et al., 2015) with the likelihood of a correct response on Day 2 as the outcome variable, fixed effects of log-transformed day 1 RT (log RT) and day 2 exposure (Intact vs. Separated), and their interaction. Random intercept effects for participants and scenario were included (for full model specification and results, see Table S3). We used a sum contrast to code the day 2 exposure condition, which allowed us to examine main effects of coefficients similar to an ANOVA. As seen in Fig. 3c, we found that log RT negatively predicted day 2 accuracy, with each 1-unit increase associated with 36% lower odds of a correct response on day 2 (OR = 0.64, $p < .001$). Neither day 2 exposure condition (OR = 0.99, $p = .863$) nor its interaction with log RT (OR = 1.03, $p = .750$) significantly predicted day 2 accuracy. These results are consistent with the ANOVA analysis and suggest that RT could predict delayed memory for corrections.

### 3.5. ERP results

For the following ERP analyses, we first examined the subsequent memory effect (SME) by dividing ERPs during encoding into subsequent correct and incorrect veracity judgements. Then, motivated by observations that faster correct veracity judgements resulted in less difference in accuracy (i.e., better accuracy) after a delay compared to slow correct veracity judgements, we conducted exploratory analyses comparing ERPs between subsequent fast and slow veracity judgements *within* subsequently correct veracity judgements. For mean and standard deviation of remaining trials in each condition, refer to Table S5.

#### 3.5.1. Encoding - event

We first examined the P300 when participants encountered an image depicting an event. A repeated measures ANOVA between veracity condition (true, false) and subsequent accuracy (correct, incorrect) with P300 amplitude as the outcome measure revealed no significant effect of subsequent accuracy ($F(1,49) = 3.53$, $p = .066$, $\eta_p^2 = 0.067$), veracity ($F(1,49) = 0.65$, $p = .424$, $\eta_p^2 = 0.013$), or interaction ($F(1,49) = 1.57$, $p = .216$, $\eta_p^2 = 0.031$). This suggested that there was no subsequent memory effect for the P300 during event encoding.

We next compared ERPs within subsequently correct answers between fast and slow responses. A repeated measures ANOVA between veracity (true, false) and subsequent RT (fast correct, slow correct) showed a significant main effect of RT, $F(1,49) = 5.54$, $p = .023$, $\eta_p^2 = 0.102$, such that subsequent fast correct answers ($M = 6.20$ μV, $SE = 0.12$ μV) were associated with greater P300 than subsequent slow correct ($M = 5.82$, $SE = 0.12$) answers. However, after correcting for multiple comparisons across image, cause and veracity stages for the P300 using the FDR method, the difference was not significant ($p = .069$). There was no main effect of veracity, $F(1,49) = 0.10$, $p = .750$, $\eta_p^2 = 0.002$, and no interaction, $F(1,49) = 0.40$, $p = .533$, $\eta_p^2 = 0.008$.

To fully account for the continuous nature of RT, we also conducted linear mixed models analysis with continuous log RT as the outcome variable, fixed effects of P300 amplitude and veracity, and random intercept effects of subject and stimulus (Table S6). A similar pattern of results emerged: although P300 amplitudes were initially a significant predictor of faster RTs ($p = .040$), they were not significant after correcting for multiple comparisons across image, cause and veracity stages ($p = .063$).

#### 3.5.2. Encoding – cause

We next examined P300 when participants viewed the cause word. A repeated measures ANOVA between veracity condition (true, false) and subsequent accuracy (correct, incorrect) with P300 amplitude as the outcome measure showed no significant main effect of subsequent accuracy, $F(1,49) = 1.90$, $p = .174$, $\eta_p^2 = 0.037$, no significant main effect of veracity, $F(1,49) = 0.12$, $p = .733$, $\eta_p^2 = 0.002$, and no significant interaction, $F(1,49) = 1.24$, $p = .271$, $\eta_p^2 = 0.025$. A follow-up analysis with RTs showed no significant main effect of subsequent RT, $F(1,49) = 0.53$, $p = .469$, $\eta_p^2 = 0.011$, no significant main effect of condition, $F(1,49) = 1.55$, $p = .219$, $\eta_p^2 = 0.031$, and no significant interaction, $F(1,49) = 3.15$, $p = .082$, $\eta_p^2 = 0.060$.

To examine whether associative encoding between cause and event would aid subsequent veracity judgement, we analyzed the FSW when participants viewed the cause word. A repeated measures ANOVA between veracity condition (true, false) and subsequent accuracy (correct, incorrect) with FSW amplitude as the outcome measure showed no significant main effect of subsequent accuracy, $F(1,49) = 0.03$, $p = .863$, $\eta_p^2 = 0.001$, no significant main effect of veracity, $F(1,49) = 0.17$, $p = .680$, $\eta_p^2 = 0.004$, and no significant interaction, $F(1,49) = 0.17$, $p = .680$, $\eta_p^2 = 0.004$. A follow-up analysis between subsequently fast and slow correct answers showed no significant main effect of subsequent RT, $F(1,49) = 2.25$, $p = .140$, $\eta_p^2 = 0.044$, no significant main effect of veracity, $F(1,49) = 0.58$, $p = .450$, $\eta_p^2 = 0.012$, and no significant interaction, $F(1,49) = 0.67$, $p = .418$, $\eta_p^2 = 0.013$. Overall, there was no evidence that in-depth encoding or associative memory processes during presentation of the cause word differed according to subsequent accuracy or RT. Analyses with continuous RT values using linear mixed models showed the same pattern of results for the P300 and FSW (Table S7 and S8; $ps > 0.340$).

#### 3.5.3. Encoding - veracity

To examine whether in-depth encoding of a correction or affirmation would be linked to subsequent veracity judgements, the P300 was analyzed when participants encountered veracity information. A repeated measures ANOVA between veracity condition (true, false) and subsequent accuracy (correct, incorrect) with P300 amplitude as the outcome measure showed no significant main effect of subsequent accuracy, $F(1,49) = 2.10$, $p = .154$, $\eta_p^2 = 0.041$, and no significant

interaction, $F(1,49) = 3.11$, $p = .084$, $\eta_p^2 = 0.060$. There was a significant main effect of veracity, such that affirmations ($M = 0.47$ μV, $SE = 0.09$ μV) were associated with greater P300s than corrections ($M = 0.15$ μV, $SE = 0.09$ μV), $F(1,49) = 6.05$, $p = .017$, $\eta_p^2 = 0.110$ (Fig. 4a–c). A follow-up analysis between subsequent fast and slow correct answers showed that there was no significant main effect of subsequent RT, $F(1,49) = 0.49$, $p = .487$, $\eta_p^2 = 0.010$, no main effect of veracity, $F(1,49) = 0.82$, $p = .368$, $\eta_p^2 = 0.017$, and no significant interaction, $F(1,49) = 0.08$, $p = .780$, $\eta_p^2 = 0.002$. Although analyses with continuous RT values initially showed that P300 amplitudes were a significant predictor of RT ($p = .042$), they were not significant after correcting for multiple comparisons (Table S9; $p = .063$).

Although it is unclear whether associative memory processes elicited during encoding of corrections or affirmations would reflect associations with the event or the cause specifically, examining the FSW SME can inform us about whether general associative processes are associated with subsequent memory. To this end, we analyzed the FSW when participants read corrections or affirmations. A repeated measures ANOVA between veracity condition (true, false) and subsequent accuracy (correct, incorrect) with FSW amplitude as the outcome measure revealed no significant main effect of subsequent accuracy, $F(1,49) = 0.33$, $p = .569$, $\eta_p^2 = 0.007$, and no main effect of veracity, $F(1,49) = 2.43$, $p = .126$, $\eta_p^2 = 0.047$. The interaction was also non-significant, $F(1,49) = 3.51$, $p = .067$, $\eta_p^2 = 0.067$.

A follow-up analysis between subsequent fast and slow correct answers revealed a significant effect of veracity, $F(1,49) = 7.85$, $p = .007$, $\eta_p^2 = 0.138$, such that corrections ($M = 2.23$ μV, $SE = 0.36$ μV) elicited a greater FSW compared to affirmations ($M = 0.82$ μV, $SE = 0.36$ μV), as seen in Fig. 4b and d. Correcting for multiple comparisons with the FDR method (including FSW-RT analyses in the veracity and cause stages)

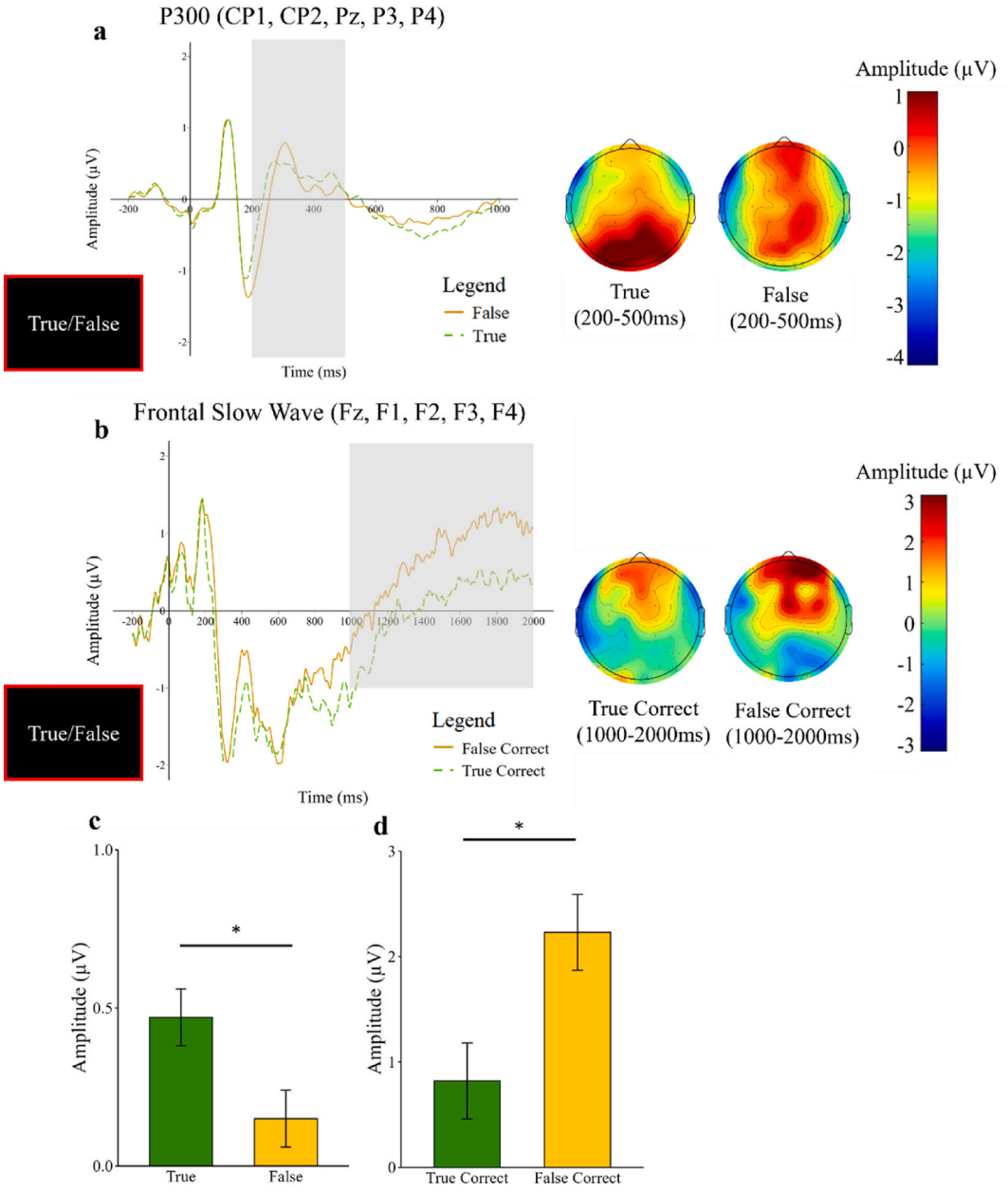



**Fig. 4.** *P300 and FSW Associated with Corrections and Affirmations*
*Note.* ERP and scalp topography showing main effects of veracity in **a)** Parietal P300 and **b)** FSW, time-locked to corrections or affirmations (epoch outlined in red). Adaptive mean amplitude in **c)** Parietal P300 and **d)** FSW. Error bars denote standard error. *$p < .05$. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

did not alter the significance of this finding ($p = .014$). However, there was no main effect of RT, $F(1,49) = 0.10$, $p = .754$, $\eta_p^2 = 0.002$, and no interaction, $F(1,49) = 0.07$, $p = .795$, $\eta_p^2 = 0.001$. FSW amplitude similarly did not predict continuous RTs (Table S10, $ps > 0.131$).

Overall, results indicated that when learning about veracity, in-depth encoding and associative memory processes did not differ for subsequent correct and incorrect veracity judgements. However, in-depth encoding during affirmations was greater than during corrections, and within subsequent correct answers, corrections elicited a greater FSW than affirmations.

#### 3.5.4. Relationship between day 1 encoding and day 2 behavior

To examine whether Day 1 encoding ERPs showed an SME for veracity judgement accuracy on Day 2, we conducted 2 (Cause veracity: true, false) × 2 (Day 2 accuracy: correct, incorrect) repeated measures ANOVAs on the same components and epochs as before. No significant main effect of accuracy on Day 2 or interactions were observed for all components and epochs ($ps > 0.05$). For detailed statistics, please refer to Supplementary Material F.

## 4. Discussion

Misinformation often continues to influence judgement and decision making even after its correction (Ecker et al., 2022). In this study, we sought to uncover the mechanisms behind successful and unsuccessful corrections to causal misinformation by examining neural activity while participants encoded events, causes and corrections or affirmations. We also examined how cause-correction associations related to delayed memory for corrections, and whether this relationship was modulated by viewing intact and separated cause-event pairs. We next discuss our results in greater detail.

### 4.1. Veracity judgement accuracy

Our results are consistent with prior work showing that corrections are remembered to a lesser extent than affirmations (Brydges et al., 2020; Gordon et al., 2017). Participants also took a longer time to remember corrections than affirmations and were more confident when failing to remember corrections than when failing to remember affirmations. These results suggest that it may be more difficult to retrieve corrections compared to affirmations, and that people may be confident even when they do not remember corrections.

### 4.2. Delayed memory for corrections

Memory for corrections to misinformation diminishes over time (Swire-Thompson et al., 2023). There may be substantial variation in how different people believe in misinformation (Porter and Wood, 2024; Walter and Murphy, 2018), but less is known about the heterogeneity in memory for corrections. We hypothesized that the strength of misinformation-correction associations would be linked to delayed veracity judgements. We attempted to measure the strength of this association by recording participants' response times (RTs) when recalling corrections or affirmations originally associated with a cause (Dewhurst et al., 2006; Wixted and Stretch, 2004; Eimas and Zeaman, 1963). Some support for our hypothesis was found: faster RTs when correctly remembering corrections on Day 1 predicted a greater likelihood of remembering those corrections on Day 2, and greater confidence was linked to faster RT judgements. However, given that RT was only measured during retrieval, it is also possible that RTs instead reflect enhanced retrieval processes that could covary with the distinctiveness of cause words (i.e., reduced interference from similar causes). This was not investigated in our study, and the relationship between distinctiveness and correction recollection could be an interesting avenue for future research.

Another variable that may influence delayed memory for corrections is whether people are exposed to misinformation again after it is corrected. The illusory truth effect posits that reading or comprehending statements can increase their perceived truthfulness (Fazio et al., 2015; Gilbert et al., 1990). However, it is unclear if the effect persists when statements have already been corrected. We attempted to address this gap by showing participants intact or separated cause-event pairings during a simultaneous recognition and exposure task. Viewing intact cause-event pairs did not seem to impact memory for corrections relative to viewing separated pairs, which suggests that the illusory truth effect may be weak if statements have already been corrected. On the other hand, it can be argued that presenting events and causes without explicitly instructing participants to evaluate them propositionally may elicit different mental processes than reading sentences, headlines or narratives in traditional illusory truth effect experiments. Although this is true, we note that participants were previously instructed to interpret words following images propositionally on Day 1 (i.e., interpret the word as the cause for the preceding image), which may have resulted in similar processing of causes and events during the simultaneous recognition and exposure task. Indeed, greater recognition performance in the Intact than Separated condition suggested that viewing event images facilitated the recognition of cause words at a higher rate for intact than separated pairs. Interestingly, we also found that causes in the Intact condition were recognized more frequently for true than false causes. Although this was not our primary focus, this suggests that either corrections weakened the strength of cause-event pairs, or affirmations strengthened the pair. Future research could examine this at a deeper level for greater insight into correction mechanisms.

### 4.3. Electrophysiological findings

In addition to behavioral results, analyzing electrophysiological activity can provide insight into how encoding processes differ between subsequent accuracy and veracity condition. We examined how the P300 and FSW during event, cause, and veracity encoding related to subsequent veracity judgment accuracy and RTs, which reflected how well participants could retrieve corrections and affirmations. The P300 and FSW SME index in-depth encoding and associative encoding processes respectively (Forester and Kamp, 2023; Johnson, 1995; Kamp et al., 2016, 2017; Kim et al., 2009). We did not observe SMEs during event, cause, or veracity encoding. This was surprising, as existing research finds that associations between misinformation corrections are important in veracity judgements (Gilbert et al., 1990, 1993), suggesting that SMEs should arise when encoding corrections. However, strong conclusions to the contrary cannot be made based on our non-significant findings alone, highlighting the need for further work in this area. We provide a few speculations on why we did not find the SME here. Firstly, previous work on the continued influence effect has shown that the effectiveness of corrections could depend on the extent to which incomplete event models cause discomfort (Susmann and Wegener, 2021), or the extent to which participants generated an alternative explanation for an event (Lewandowsky et al., 2012), processes that ERPs may struggle to capture. Secondly, SMEs that reflect successful encoding of corrections and affirmations may manifest in longer time-windows. For example, Forester and Kamp (2023) discovered SMEs in 6000ms time-windows, covering an entire trial in which participants encoded a pair of images one after the other. Finally, when participants encountered a repeated scenario during encoding, they may have paid less attention or adjusted their strategy in a non-systematic way, resulting in increased noise in the final ERPs.

When examining neural activity associated with encoding corrections and affirmations, we found that corrections elicited a weaker P300. Outside of SME studies, lower P300 amplitudes have been linked to increased task difficulty and memory load (Kok, 2001; Scharinger et al., 2017), while greater P300 amplitude has been linked to more effective encoding of misinformation corrections (Guo et al., 2025). Together, this suggests that corrections in our experiment may have been more

difficult than encoding affirmations and thus less effectively encoded. By reinforcing existing event-cause associations, affirmations may have elicited greater fluency compared to corrections that prompt revision of prior beliefs, thus being easier to process. Indeed, greater working memory capacity and central executive function have been shown to correlate with reduced CIE (Brydges et al., 2018; Jia et al., 2020), suggesting that the capacity to update information in working memory is an important component of correcting misinformation.

Within subsequently remembered corrections and affirmations, corrections were associated with a greater FSW amplitude than affirmations. However, a nuanced interpretation of this effect is difficult, given that FSW amplitudes have been known to fluctuate with working memory load (McEvoy, 1998; Monfort and Pouthas, 2003; Rämä et al., 2000), cognitive control (West and Travers, 2008), prediction error (Van Petten and Luka, 2012), and negation processes (Herbert and Kissler, 2014). Thus, while the more positive FSW for corrections suggests that corrections and affirmations are processed differently, the present data do not permit a more fine-grained attribution of this effect to any single underlying mechanism.

### 4.4. Limitations and future directions

Limitations should be noted. During veracity judgments, only the cause of an event was used as a cue, whereas outside the laboratory, we tend to encounter causes with their corresponding events. However, this was necessary for our experiment to obtain a sensitive measure of response time, as presenting event cues before or after the cause would elicit additional recollective processes before participants were able to respond. Additionally, to increase trial counts for SME analyses, we presented each trial twice during encoding. Neural activity in response to trial repetitions may differ from first exposure, and observed effects may be confounded with recognition or recollection processes.

Future research on refuting misinformation should consider analyzing response times to assess whether interventions are more effective for certain types of misinformation. In cases where collecting longitudinal data is difficult, response times may also be a viable indicator to predict delayed memory for corrections. The present study did not examine retrieval-related ERP components, which limits our insight into how retrieval mechanisms influence memory for misinformation corrections. Future research could examine these components (e.g., late positive component, FN400, or feedback and error related components) to explore how the neural correlates of correction retrieval relate to long-term correction retention.

### 4.5. Conclusion

Given the potential dangers of misinformation and its persistent influence after correction, the present experiment recorded electrophysiological activity while participants encoded events, causes and corrections, and examined their relationships with subsequent memory for corrections. We found no evidence that ERPs were related to subsequent memory for corrections, but found that corrections may have been more difficult to process compared to affirmations. Re-exposure to misinformation did not modulate delayed memory for corrections. However, stronger initially correct memories for corrections predicted more accurate memory for corrections after the delay, suggesting that future interventions could consider measuring response times to target weaker correction memories and improve the long-term effectiveness of corrections.

## CRediT authorship contribution statement

**Sean Guo:** Writing – review & editing, Writing – original draft, Visualization, Software, Resources, Methodology, Investigation, Formal analysis, Conceptualization. **Danni Chen:** Writing – review & editing, Methodology, Conceptualization. **Wanrou Hu:** Investigation, Formal analysis. **Xiaoqing Hu:** Writing – review & editing, Supervision, Methodology, Funding acquisition, Conceptualization.

## Ethics

This research was approved by the Human Research Ethics Committee of the University of Hong Kong (EA210341).

## Open practices statement

Data are available at https://osf.io/j9evm/. The experiment was not preregistered, and materials are available upon request.

## Consent to participate

Participants provided informed consent prior to participation.

## Consent for publication

Not applicable.

## Code availability

Code will be available upon request.

## Funding

The research was supported by the Ministry of Science and Technology of China STI2030-Major Projects (No. 2022ZD0214100), National Natural Science Foundation of China (No. 32171056), General Research Fund (No. 17614922) of Hong Kong Research Grants Council to X. H.

## Declaration of competing interest

The authors have no conflicts of interest to report.

## Acknowledgements

The authors would like to thank Xibo Zuo and Ruoying Zheng for their assistance in data collection.

## Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.neuropsychologia.2026.109428.

## References

Bates, D., Mächler, M., Bolker, B.M., Walker, S.C., 2015. Fitting linear mixed-effects models using lme4. J. Stat. Software 67 (1). https://doi.org/10.18637/jss.v067.i01.

Brydges, C.R., Gignac, G.E., Ecker, U.K.H., 2018. Working memory capacity, short-term memory capacity, and the continued influence effect: a latent-variable analysis. Intelligence 69 (December 2017), 117–122. https://doi.org/10.1016/j.intell.2018.03.009.

Brydges, C.R., Gordon, A., Ecker, U.K.H., 2020. Electrophysiological correlates of the continued influence effect of misinformation: an exploratory study. J. Cognit. Psychol. 32 (8), 771–784. https://doi.org/10.1080/20445911.2020.1849226.

Carey, J.M., Guess, A.M., Loewen, P.J., Merkley, E., Nyhan, B., Phillips, J.B., Reifler, J., 2022. The ephemeral effects of fact-checks on COVID-19 misperceptions in the United States, Great Britain and Canada. Nat. Hum. Behav. 6 (2), 236–243. https://doi.org/10.1038/s41562-021-01278-3.

Chan, M.P.S., Jones, C.R., Hall Jamieson, K., Albarracín, D., 2017. Debunking: a meta-analysis of the psychological efficacy of messages countering misinformation. Psychol. Sci. 28 (11), 1531–1546. https://doi.org/10.1177/0956797617714579.

Craddock, P., Molet, M., Miller, R.R., 2012. Reaction time as a measure of human associative learning. Behav. Process. 90 (2), 189–197. https://doi.org/10.1016/j.beproc.2012.01.006.

Cushman, F., 2008. Crime and punishment: distinguishing the roles of causal and intentional analyses in moral judgment. Cognition 108 (2), 353–380. https://doi.org/10.1016/j.cognition.2008.03.006.

Delorme, A., Makeig, S., 2004. EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. J. Neurosci. Methods 134. http://www.sccn.ucsd.edu/eeglab/.
Dewhurst, S.A., Holmes, S.J., Brandt, K.R., Dean, G.M., 2006. Measuring the speed of the conscious components of recognition memory: remembering is faster than knowing. Conscious. Cognit. 15 (1), 147–162. https://doi.org/10.1016/j.concog.2005.05.002.
Ecker, U.K.H., Lewandowsky, S., Cook, J., Schmid, P., Fazio, L.K., Brashier, N., Kendeou, P., Vraga, E.K., Amazeen, M.A., 2022. The psychological drivers of misinformation belief and its resistance to correction. Nat. Rev. Psychol. 1 (1), 13–29. https://doi.org/10.1038/s44159-021-00006-y.
Ecker, U.K.H., Lewandowsky, S., Swire, B., Chang, D., 2011. Correcting false information in memory: manipulating the strength of misinformation encoding and its retraction. Psychonomic Bulletin and Review 18 (3), 570–578. https://doi.org/10.3758/s13423-011-0065-1.
Edwards, J.R., 2001. Ten difference score myths. Organ. Res. Methods 4 (3), 265–287.
Eimas, P.D., Zeaman, D., 1963. Response speed changes in an Estes' paired-associate "miniature" experiment. J. Verb. Learn. Verb. Behav. 1 (5), 384–388. https://doi.org/10.1016/S0022-5371(63)80022-5.
Fazio, L.K., Brashier, N.M., Keith Payne, B., Marsh, E.J., 2015. Knowledge does not protect against illusory truth. J. Exp. Psychol. Gen. 144 (5), 993–1002. https://doi.org/10.1037/xge0000098.
Forester, G., Kamp, S.M., 2023. Pre-associative item encoding influences associative memory: behavioral and ERP evidence. Cognit. Affect Behav. Neurosci. https://doi.org/10.3758/s13415-023-01102-7.
Forester, G., Kroneisen, M., Erdfelder, E., Kamp, S.M., 2020. Survival processing modulates the neurocognitive mechanisms of episodic encoding. Cognit. Affect Behav. Neurosci. 20 (4), 717–729. https://doi.org/10.3758/s13415-020-00798-1.
Ghani, U., Signal, N., Niazi, I.K., Taylor, D., 2020. ERP based measures of cognitive workload: a review. In: Neuroscience and Biobehavioral Reviews, vol 118. Elsevier Ltd, pp. 18–26. https://doi.org/10.1016/j.neubiorev.2020.07.020.
Gilbert, D.T., Krull, D.S., Malone, P.S., 1990. Unbelieving the unbelievable: some problems in the rejection of false information. J. Pers. Soc. Psychol. 59 (4), 601–613. https://doi.org/10.1037/0022-3514.59.4.601.
Gilbert, D.T., Tafarodi, R.W., Malone, P.S., 1993. You can't not believe everything you read. J. Pers. Soc. Psychol. 65 (2), 221–233. https://doi.org/10.1037/0022-3514.65.2.221.
Gonsalves, B., Paller, K.A., 2000. Neural events that underlie remembering something that never happened. Nat. Neurosci. 3 (12), 1316–1321. https://doi.org/10.1038/81851.
Gordon, A., Brooks, J.C.W., Quadflieg, S., Ecker, U.K.H., Lewandowsky, S., 2017. Exploring the neural substrates of misinformation processing. Neuropsychologia 106 (September), 216–224. https://doi.org/10.1016/j.neuropsychologia.2017.10.003.
Guo, S., Chen, D., Hu, X., 2025. Providing an alternative explanation improves misinformation rejection and alters event-related potentials during veracity judgements. Brain Cognit. 186, 106290. https://doi.org/10.1016/j.bandc.2025.106290.
Herbert, C., Kissler, J., 2014. Event-related potentials reveal task-dependence and inter-individual differences in negation processing during silent listening and explicit truth-value evaluation. Neuroscience 277, 902–910. https://doi.org/10.1016/j.neuroscience.2014.07.043.
Jia, L., Shan, J., Xu, G., Jin, H., 2020. Influence of individual differences in working memory on the continued influence effect of misinformation. J. Cognit. Psychol. 32 (5–6), 494–505. https://doi.org/10.1080/20445911.2020.1800019.
Jin, H., Jia, L., Yin, X., Yan, S., Wei, S., Chen, J., 2022. The neural basis of the continued influence effect of misinformation. Acta Psychol. Sin. 54 (4), 343. https://doi.org/10.3724/sp.j.1041.2022.00343.
Johnson, R., 1995. Event-related potential insights into the neurobiology of memory systems. In: Handbook of Neuropsychology, vol. 10, pp. 135–163. https://www.researchgate.net/publication/312950087.
Kamp, S.M., Bader, R., Mecklinger, A., 2017. ERP subsequent memory effects differ between inter-item and unitization encoding tasks. Front. Hum. Neurosci. 11. https://doi.org/10.3389/fnhum.2017.00030.
Kamp, S.M., Lehman, M., Malmberg, K.J., Donchin, E., 2016. A buffer model account of behavioral and ERP patterns in the Von Restorff paradigm. AIMS Neuroscience 3 (2), 181–202. https://doi.org/10.3934/Neuroscience.2016.2.181.
Kemp, P.L., Goldman, A.C., Wahlheim, C.N., 2024. On the role of memory in misinformation corrections: repeated exposure, correction durability, and source credibility. In: Current Opinion in Psychology, vol. 56. Elsevier B.V. https://doi.org/10.1016/j.copsyc.2023.101783.
Kendeou, P., Butterfuss, R., Kim, J., Van Boekel, M., 2019. Knowledge revision through the lenses of the three-pronged approach. Mem. Cognit. 47 (1), 33–46. https://doi.org/10.3758/s13421-018-0848-y.
Kendeou, P., Walsh, E.K., Smith, E.R., O'Brien, E.J., 2014. Knowledge revision processes in refutation texts. Discourse Process. 51 (5–6), 374–397. https://doi.org/10.1080/0163853X.2014.913961.
Kim, A.S.N., Vallesi, A., Picton, T.W., Tulving, E., 2009. Cognitive association formation in episodic memory: evidence from event-related potentials. Neuropsychologia 47 (14), 3162–3173. https://doi.org/10.1016/j.neuropsychologia.2009.07.015.
Kok, A., 2001. On the utility of P3 amplitude as a measure of processing capacity. Psychophysiology 38 (3), 557–577. https://doi.org/10.1017/S0048577201990559.
Kounios, J., Smith, R.W., Yang, W., Bachman, P., Esposito, M.D.', 2001. Cognitive Association Formation in human memory revealed by spatiotemporal brain imaging. In: Neuron, vol. 29.
Lewandowsky, S., Ecker, U.K.H., Seifert, C.M., Schwarz, N., Cook, J., 2012. Misinformation and its correction: continued influence and successful debiasing. Psychological Science in the Public Interest, Supplement 13 (3), 106–131. https://doi.org/10.1177/1529100612451018.
Lombrozo, T., 2010. Causal-explanatory pluralism: how intentions, functions, and mechanisms influence causal ascriptions. Cogn. Psychol. 61 (4), 303–332. https://doi.org/10.1016/j.cogpsych.2010.05.002.
Lopez-Calderon, J., Luck, S.J., 2014. ERPLAB: an open-source toolbox for the analysis of event-related potentials. Front. Hum. Neurosci. 8 (1 APR). https://doi.org/10.3389/fnhum.2014.00213.
McEvoy, L., 1998. Dynamic cortical networks of verbal and spatial working memory: effects of memory load and task practice. Cerebr. Cortex 8 (7), 563–574. https://doi.org/10.1093/cercor/8.7.563.
Mecklinger, A., Kamp, S.M., 2023. Observing memory encoding while it unfolds: functional interpretation and current debates regarding ERP subsequent memory effects. In: Neuroscience and Biobehavioral Reviews, vol. 153. Elsevier Ltd. https://doi.org/10.1016/j.neubiorev.2023.105347.
Monfort, V., Pouthas, V., 2003. Effects of working memory demands on frontal slow waves in time-interval reproduction tasks in humans. Neurosci. Lett. 343 (3), 195–199. https://doi.org/10.1016/S0304-3940(03)00385-9.
Moran, T.P., Jendrusina, A.A., Moser, J.S., 2013. The psychometric properties of the late positive potential during emotion processing and regulation. Brain Res. 1516, 66–75. https://doi.org/10.1016/j.brainres.2013.04.018.
Morey, R.D., 2008. Confidence Intervals from Normalized Data: a correction to Cousineau (2005). Tutorials in Quantitative Methods for Psychology 4 (2), 61–64. https://doi.org/10.20982/tqmp.04.2.p061.
Nielsen, K., Gonzalez, R., 2020. Comparison of common amplitude metrics in event-related potential analysis. Multivariate Behav. Res. 55 (3), 478–493. https://doi.org/10.1080/00273171.2019.1654358.
Peirce, J., Gray, J.R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman, E., Lindeløv, J.K., 2019. PsychoPy2: experiments in behavior made easy. Behav. Res. Methods 51 (1), 195–203. https://doi.org/10.3758/s13428-018-01193-y.
Pion-Tonachini, L., Kreutz-Delgado, K., Makeig, S., 2019. ICLabel: an automated electroencephalographic independent component classifier, dataset, and website. Neuroimage 198, 181–197. https://doi.org/10.1016/j.neuroimage.2019.05.026.
Polich, J., 2007. Updating P300: an integrative theory of P3a and P3b. In: Clinical Neurophysiology, vol 118, pp. 2128–2148. https://doi.org/10.1016/j.clinph.2007.04.019. Issue 10.
Porter, E., Wood, T.J., 2024. Factual corrections: concerns and current evidence. In: Current Opinion in Psychology, vol. 55. Elsevier B.V. https://doi.org/10.1016/j.copsyc.2023.101715.
Radvansky, G.A., Doolen, A.C., Pettijohn, K.A., Ritchey, M., 2022. A new look at memory retention and forgetting. J. Exp. Psychol. Learn. Mem. Cognit. 48 (11), 1698–1723. https://doi.org/10.1037/xlm0001110.
Rämä, P., Paavilainen, L., Anourova, I., Alho, K., Reinikainen, K., Sipilä, S., Carlson, S., 2000. Modulation of slow brain potentials by working memory load in spatial and nonspatial auditory tasks. Neuropsychologia 38 (7), 913–922. https://doi.org/10.1016/S0028-3932(00)00019-1.
Rich, P.R., Zaragoza, M.S., 2020. Correcting misinformation in news stories: an investigation of correction timing and correction durability. Journal of Applied Research in Memory and Cognition 9 (3), 310–322. https://doi.org/10.1016/j.jarmac.2020.04.001.
Roheger, M., Folkerts, A.-K., Krohm, F., Skoetz, N., Kalbe, E., 2020. Prognostic factors for change in memory test performance after memory training in healthy older adults: a systematic review and outline of statistical challenges. Diagn. Progn. Res. 4 (1). https://doi.org/10.1186/s41512-020-0071-8.
Scharinger, C., Soutschek, A., Schubert, T., Gerjets, P., 2017. Comparison of the working memory load in N-back and working memory span tasks by means of EEG frequency band power and P300 amplitude. Front. Hum. Neurosci. 11. https://doi.org/10.3389/fnhum.2017.00006.
Susmann, M.W., Wegener, D.T., 2021. The role of discomfort in the continued influence effect of misinformation. Mem. Cognit. https://doi.org/10.3758/s13421-021-01232-8.
Swire-Thompson, B., Dobbs, M., Thomas, A., DeGutis, J., 2023. Memory failure predicts belief regression after the correction of misinformation. Cognition 230. https://doi.org/10.1016/j.cognition.2022.105276.
Van Petten, C., Luka, B.J., 2012. Prediction during language comprehension: benefits, costs, and ERP components. In: International Journal of Psychophysiology, vol. 83, pp. 176–190. https://doi.org/10.1016/j.ijpsycho.2011.09.015. Issue 2.
Walter, N., Murphy, S.T., 2018. How to unring the bell: a meta-analytic approach to correction of misinformation. Commun. Monogr. 85 (3), 423–441. https://doi.org/10.1080/03637751.2018.1467564.
Weng, O., Johnson, K.J., Kreuter, M.W., 2024. Repeated exposure to COVID-19 misinformation: a longitudinal analysis of prevalence and predictors in a community sample. J. Publ. Health Manag. Pract. 30 (5), E211–E214. https://doi.org/10.1097/PHH.0000000000002019.
West, R., Travers, S., 2008. Tracking the temporal dynamics of updating cognitive control: an examination of error processing. Cerebr. Cortex 18 (5), 1112–1124. https://doi.org/10.1093/cercor/bhm142.
Wilkes, A.L., Leatherbarrow, M., 1988. Editing episodic memory following the identification of error. The Quarterly Journal of Experimental Psychology Section A 40 (2), 361–387. https://doi.org/10.1080/02724988843000168.
Wixted, J.T., Stretch, V., 2004. In defense of the signal detection interpretation of remember/know judgments. Psychon. Bull. Rev. 11 (4), 616–641.
Wood, T., Porter, E., 2019. The elusive backfire effect: mass attitudes' steadfast factual adherence. Polit. Behav. 41 (1), 135–163.