Memory

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/pmem20

Factors affecting the forced confabulation effect: a meta-analysis of laboratory studies

Paul Riesthuis, Henry Otgaar, Glynis Bogaard & Ivan Mangiulli

To cite this article: Paul Riesthuis, Henry Otgaar, Glynis Bogaard & Ivan Mangiulli (2023) Factors affecting the forced confabulation effect: a meta-analysis of laboratory studies, Memory, 31:5, 635-651, DOI: 10.1080/09658211.2023.2185931

To link to this article: https://doi.org/10.1080/09658211.2023.2185931

Published online: 08 Mar 2023.

MEMORY
2023, VOL. 31, NO. 5, 635–651
https://doi.org/10.1080/09658211.2023.2185931

Routledge
Taylor & Francis Group

OPEN ACCESS Check for updates

Factors affecting the forced confabulation effect: a meta-analysis of laboratory studies

Paul Riesthuis [a,b], Henry Otgaar[a,b], Glynis Bogaard[b] and Ivan Mangiulli [a,c]

[a]Faculty of Law and Criminology, KU Leuven, Leuven, Belgium; [b]Faculty of Psychology and Neuroscience, Maastricht University, Maastricht, The Netherlands; [c]Department of Education, Psychology, Communication Sciences, University of Bari, Bari, Italy

ABSTRACT
During police investigations, interviewees are sometimes forced to confabulate a response to questions for which they don't know the answer. In this registered report, we conducted a three-level meta-analysis to examine whether forcing people to confabulate an answer to these questions can lead to false memories for the confabulated details and/or events (i.e., forced confabulation effect). Results showed that forced confabulation indeed led to the production of false memories. Qualitative reviews of studies suggested that the forced confabulation effect is enhanced after receiving confirmatory feedback. Moreover, we found evidence that longer delays between the forced confabulation phase and the final memory task are necessary to observe the forced confabulation effect for entire events. However, caution is needed when interpreting the forced confabulation effect. Specifically, our moderator analyses revealed that voluntarily produced confabulation led to more false memories than forced confabulation. Also, our exploratory analysis indicated that the forced confabulation effect was mainly observed in within-subject designs. Taken together, our meta-analysis supports the notion that forcing participants to confabulate can lead them to later report such confabulations as part of the truth. Nonetheless, caution is warranted because this effect might be due to the introduction of misinformation through asking unanswerable questions.

ARTICLE HISTORY
Received 10 November 2020
Accepted 23 February 2023

KEYWORDS
Lying; forced confabulation; memory; false memory; meta-analysis

Post-event misinformation has been shown to exert adverse effects on memory (for a review, see Loftus, 2005; Zaragoza et al., 2006). A typical procedure examining the impact of misinformation on memory is as follows. Participants are shown certain stimuli (e.g., a car stopping in front of a stop sign) followed by questions containing misinformation such as "*Did the car stop in front of a yield sign?*" from an external source (e.g., the experimenter). On a subsequent memory task, the standard finding is that participants report the misinformation as having been presented during the stimulus presentation, thereby creating false memories (Loftus et al., 1978). This finding has also been referred to as the misinformation effect (Frenda et al., 2011; Loftus, 2005). A plethora of research has shown the robustness of the misinformation effect using various stimuli and populations (e.g., Loftus, 2005). Importantly, this research has shown that false memories arise because of *external* misinformation (Scoboria et al., 2017). However, false memories can also be elicited through *internal* misinformation such as lying (Otgaar & Baker, 2018).

To date, much less is known about the impact of lying on false memory formation. One reason is that studies on the effects of lying on memory have used different forms of lying (e.g., feigning amnesia, false denial). This makes it difficult to examine how one specific form of internal misinformation affects false memory proneness (Otgaar et al., 2014; van Oorsouw & Merckelbach, 2004). Interestingly, there exists one research line in this field of lying and memory that has used a roughly similar methodology, namely the forced confabulation paradigm (Ackil & Zaragoza, 1998). A meta-analysis about the forced confabulation literature can provide insights into the size of the effect by which false memories are formed through internal misinformation.

Empirical work on the effects of lying on memory is crucial because in legal cases, witnesses, victims and suspects sometimes lie and this might negatively influence their memory. For example, during investigative interviewing, police interrogators sometimes force victims, witnesses and/or suspects to confabulate answers for questions on which they do not have correct answers

CONTACT Paul Riesthuis Paul.riesthuis@kuleuven.be Faculty of Law and Criminology, KU Leuven, Herbert Hooverplein, 10, 3000, Leuven, Belgium; Faculty of Psychology and Neuroscience, Universiteitssingel 40, 6229 ER, Maastricht University, Maastricht, The Netherlands

(Kassin, 2006). The case of Paul Ingram illustrates how forcedly self-generating untrue statements (i.e., forced confabulation) can impact memory for an experienced event (Wright, 1995). Paul Ingram was accused by his two daughters of having sexually abused them. After months of being repeatedly interrogated, hypnotised and encouraged by police officers and his church pastor to recall the events, he formed false memories for having sexually abused his daughters (Ofshe, 1992). Afterwards, he confessed to the crime and was convicted to 20 years in prison, even though he had recanted his confession later on. Based on this case, several questions might be postulated. First, a vital question is to what extent such forced confabulations can foment the production of false memories and thus affect the reliability of memory-based statements after coming forward with the truth as in the case of Paul Ingram. Second, the question is whether the forced confabulation effect is tantamount to the well-established effect of external misinformation on memory (Loftus, 2005).

Research into the memory effects of forced confabulation has generally used the forced confabulation paradigm (Ackil & Zaragoza, 1998, 2011; Chrobak & Zaragoza, 2008, 2013; Drivdahl & Zaragoza, 2001; Hanba & Zaragoza, 2007; Otgaar et al., 2014; Pezdek et al., 2007, 2009; Zaragoza et al., 2001). In general, studies using this paradigm have shown that forced confabulations can lead to false memories. These elicited false memories can lead to detrimental consequences for the reliability of statements in the courtroom. Establishing the mean effect size helps with informing legal professionals on whether lying can meaningfully contaminate memory. This is vital because false memories can lead to miscarriages of justice such as Paul Ingram's, wherein he formed false memories for his confabulations. Hence, the goal of this meta-analysis is to examine the effect size of the forced confabulation effect across studies.

Forced confabulation and memory

In the first study using the forced confabulation paradigm (Ackil & Zaragoza, 1998), participants first watched a 9-minute fragment of the film "*Looking for Miracles*". This fragment contained three distinct events: a birthday celebration, a boat trip and an altercation. Subsequently, participants were interviewed about the video. Before the interview, participants were assigned to either the forced confabulation or control condition. In the forced confabulation condition, participants were instructed to answer each question and guess when in doubt. In the control condition, participants had to give an honest account and answer only when certain. The crucial part of this paradigm was that the participants in both conditions were asked questions about details that clearly did not take place in the video (i.e., unanswerable questions). After a certain delay, participants had to complete a source memory test wherein they were asked whether they remembered discussing a certain detail during the interview or whether they remember seeing it in the video. Studies using this paradigm have demonstrated that forcing participants to confabulate a response to unanswerable questions causes participants to form false memories for their confabulations, also known as the *forced confabulation effect* (Ackil & Zaragoza, 1998, 2011; Chrobak & Zaragoza, 2008, 2013; Drivdahl & Zaragoza, 2001; Hanba & Zaragoza, 2007; Otgaar et al., 2014; Pezdek et al., 2007, 2009; Zaragoza et al., 2001).

Specifically, the forced confabulation effect entails that forcing participants (vs. truth-telling participants) to confabulate an answer to unanswerable questions leads them to form false memories for their confabulations. This suggests that forcing people to confabulate leads them to remember their lies as the truth. A possible explanation underlying the forced confabulation effect is provided by the Source Monitoring Framework (SMF; Johnson et al., 1993). The SMF stipulates that the mechanisms involved in determining whether the source of a memory is externally (e.g., someone else's story of an experienced event) or internally (e.g., imagination of an event) derived are based on certain attributions. More precisely, truly experienced events are expected to contain more perceptual, affective and/or contextual details whereas imagined events are expected to entail more cognitive details (e.g., references to reasoning) (Johnson et al., 1993). If imagined events contain an abundance of perceptual, affective and/or contextual features, they can be perceived as true memories. In that case, individuals might be prone to making source monitoring errors, thereby confounding imaginations with true memories. A plethora of research has demonstrated that engaging in imagination exercises can increase the amount of perceptual, affective and/or contextual features (e.g., Johnson et al., 1993; Mazzoni & Memon, 2003), and one such form of imagination is forced confabulation. Hence, in light of forcing people to confabulate about an event (or part of it), it increases the likelihood of attributing perceptual, affective and/or contextual features to the confabulations which can be perceived as true, thereby leading to false memory creation (Ackil & Zaragoza, 1998, 2011; Chrobak & Zaragoza, 2008, 2013; Drivdahl & Zaragoza, 2001; Hanba & Zaragoza, 2007; Otgaar et al., 2014; Pezdek et al., 2007, 2009; Zaragoza et al., 2001).

Potential moderators of the forced confabulation effect

Evidence for the forced confabulation effect has been found across multiple studies and it has shown that the effect generalises to different age groups (aged 6–7, 8–10 years and college students) (Ackil & Zaragoza, 1998), occurs for details and entire events (Chrobak & Zaragoza, 2008) and becomes larger with confirmatory feedback (Hanba & Zaragoza, 2007). However, according to our knowledge of the forced confabulation effect there are potential moderators leading to the forced confabulation effect, each discussed in the following sections. Based on our literature search and heterogeneity analyses we decided whether to further examine any of these potential moderators.

Age group

The initial study using the forced confabulation paradigm (Ackil & Zaragoza, 1998) examined three different age groups (aged 6–7, 8–10 years and college students). Findings from that study revealed that the forced confabulation effect was considerably smaller for college students than for children (aged 6–7 and 8–10 years). That is, college students produced fewer false memories for their forced confabulations compared to children. In line with this, follow-up research found the forced confabulation effect in 6- to 8-year-olds but did not find this effect in 10- to 12-year-old children and college students (Otgaar et al., 2014). So, although results suggest that forced confabulations can lead to false memory formation, perhaps this effect is much stronger in children than in adults (Ackil & Zaragoza, 1998; Otgaar et al., 2014); a developmental effect that closely aligns with the finding that children are more susceptible to suggestion than adults (Otgaar et al., 2018). In the current meta-analysis, we examined whether age (children vs. adults) was a moderator underpinning the forced confabulation effect.

Entire events versus details

The forced confabulation effect has mainly focused on the production of false memories for specific confabulated details. However, studies have also demonstrated this effect for entire fictitious events as well (Chrobak & Zaragoza, 2008, 2013). To study the forced confabulation effect for entire events, a similar forced confabulation paradigm was used with a crucial modification during the interview phase. Specifically, participants in the forced confabulation group received incorrect information about a salient scene in a witnessed video and subsequently had to elaborate upon this incorrect information. For example, in the video participants watched a scene wherein a camp counsellor stood up to make an announcement and lost her balance, fell, knocked over food from the table, and was scolded by the camp leader. In the forced confabulation interview, participants were told that someone pulled a prank on the camp counsellor that made her fall and caused the scene. Following the incorrect information that the entire event was caused by a prank, participants were forced to confabulate about what prank was pulled on the camp counsellor, who did it, and how the prank was done (Chrobak & Zaragoza, 2008).

After a 1-week delay, the authors failed to find robust evidence for the forced confabulation effect for entire fictitious events. However, after an 8-week delay, participants misremembered the entire confabulated event. Participants produced false memories for highly detailed and specific entire fictitious events caused by their forced confabulations. A possible underlying mechanism to form false memories for entire events is given by the explanatory role hypothesis (Chrobak & Zaragoza, 2013). Specifically, this hypothesis holds that forming false memories for entirely fabricated events depends on the causal and explanatory role that fabrications provide in relation to the experienced event. That is, if the fabrications help to better explain the witnessed event, then they are likely to be incorporated into the memory for what truly happened. Studies suggest that false memories can be produced for entire events via the forced confabulation effect (Chrobak & Zaragoza, 2008, 2013). However, false memories for forcedly confabulated details might be elicited more easily than forcedly confabulated events. That is, confabulated details most likely share perceptual, affective and/or contextual features with the truly occurred event, while confabulating an entire event will require more cognitive operations than confabulating details. Hence, it is plausible that the forced confabulation effect is smaller for entire events than details and confabulation event was thus included in the analyses as a potential moderator.

Confirmatory feedback

Several studies have examined the effect of receiving confirmatory feedback on the forced confabulation effect (Hanba & Zaragoza, 2007; Zaragoza et al., 2001). To study the effects of confirmatory feedback on the forced confabulation effect, participants who were forced to confabulate received either confirmatory (e.g., "That's right, ______ is the correct answer") or neutral interviewer feedback (e.g., "_____, O.K.") for their forced confabulations. A week later, participants' memory for the video was examined. Participants who received confirmatory interviewer feedback formed more false memories for their forced confabulations compared to participants who received neutral interviewer feedback. Moreover, the confirmatory interviewer feedback also elevated the confidence in the false memories and increased the probability that the same confabulated detail would be reported at a later memory test as compared with neutral interviewer feedback. Zaragoza et al. (2001) argued that the confirmatory feedback on the forced confabulations led participants to retrospectively reflect on their confabulations and subsequently include them in their memory for what truly occurred. According to the SMF (Johnson et al., 1993), engaging in such imagination can increase the amount of perceptual, affective and contextual details leading to false memory formation. This suggests that receiving confirmatory interviewer feedback on forced confabulations can inflate the forced confabulation effect. Hence, confirmatory interviewer feedback was also examined as a potential moderator to the mean effect size of the forced confabulation effect.

Type of confabulation

A final potential moderator of the forced confabulation effect could be the type of confabulation (e.g., voluntary vs forced confabulations). Pezdek et al. (2007, 2009) examined this using the following procedure: Participants watched a video about a carjacking and subsequently were asked answerable and unanswerable questions (e.g., "what logo was on the suspects' shirt?" when there was no logo on the suspects' shirt) about it. Participants were divided into two groups: forced confabulation and voluntary confabulation. Both groups were instructed to answer all questions. However, participants in the voluntary confabulation group had the option to indicate "I do not know". After a delay, participants received the same answerable and unanswerable questions about the video but now all participants had the option to indicate "I do not know". The authors replicated the forced confabulation effect. Additionally, they found that participants in the voluntary confabulation group were more inclined to repeat the same confabulation for unanswerable questions at the second memory test compared to participants in the forced confabulation group. A possible explanation is that self-generated answers of the participants in the voluntary confabulation group contained similar memory qualities as the experienced event. That is, in line with the SMF (Johnson et al., 1993), self-generated answers of participants in the voluntary confabulation group contained more perceptual, affective and contextual details leading to more repetitions of the same confabulations compared with participants in the forced confabulation group. This suggests that voluntary confabulations can generate more persisting false memories than forced confabulations, which was examined by including type of confabulation as a potential moderator.

The current meta-analysis

Taken together, forced confabulations can result into false memories. However, forced confabulation results vary between studies and estimating the mean population effect size of such an effect is of vital importance to understand the practical implications of using practices like pressuring eyewitnesses, victims and/or suspects to answer questions. That is, it is important to examine to what extent testimonies can be contaminated due to inappropriate interviewing techniques, leading to unreliable testimonies and risky convictions. Moreover, examining the varying effects of moderators provides information concerning specific situations wherein the effects of forced confabulation on memory might be increased (e.g., confirmatory feedback; Zaragoza et al., 2001). Hence, the main goal of our meta-analysis was to examine the impact of forced confabulations on false memory formation, while a secondary aim was to investigate any potential moderators underlying the forced confabulation effect. Potential moderators that were examined were age group, confabulation event, confirmatory feedback and type of confabulation. Our literature search and heterogeneity analyses indicated that potential moderators underlie the forced confabulation effect (i.e., high between-cluster heterogeneity, see Results). This meta-analysis elucidated to what extent forced confabulations could lead to false memories.

Research questions

Research Question 1 (RQ1): We examined to what extent forced confabulations elicited false memories. In other words, how does forcing participants to confabulate a response to unanswerable questions impact false memory formation for such confabulations relative to participants who tell the truth? More specifically, we assessed the mean effect size of the forced confabulation effect.

Hypothesis 1 (*H1*): We predicted that forcing participants to confabulate increased the formation of false memories compared with participants who responded honestly.

Research Question 2 (RQ2): Which demographic or manipulation variable moderated the forced confabulation effect?

2a) Are children more prone to form false memories for their forced confabulations compared with adults?

Hypothesis 2a (H2a): We expected children to produce more false memories for their forced confabulations compared with adults.

2b) Are false memories for entire fictitious events elicited by forced confabulations at the same rate for details?

Hypothesis 2b (H2b): We predicted that the forced confabulation effect for entire fictitious events will be smaller than for details.

2c) Does receiving confirmatory interviewer feedback on the forced confabulations inflate the forced confabulation effect?

Hypothesis 2c (H2c): We expected that receiving confirmatory interviewer feedback resulted in more false memories for their forced confabulations compared with receiving neutral or no interviewer feedback.

2d) Which type of confabulation (i.e., voluntary or forced) leads to more false memories?

Hypothesis 2d (H2d): We predicted that voluntary confabulations increased false memory formation compared to forced confabulations.

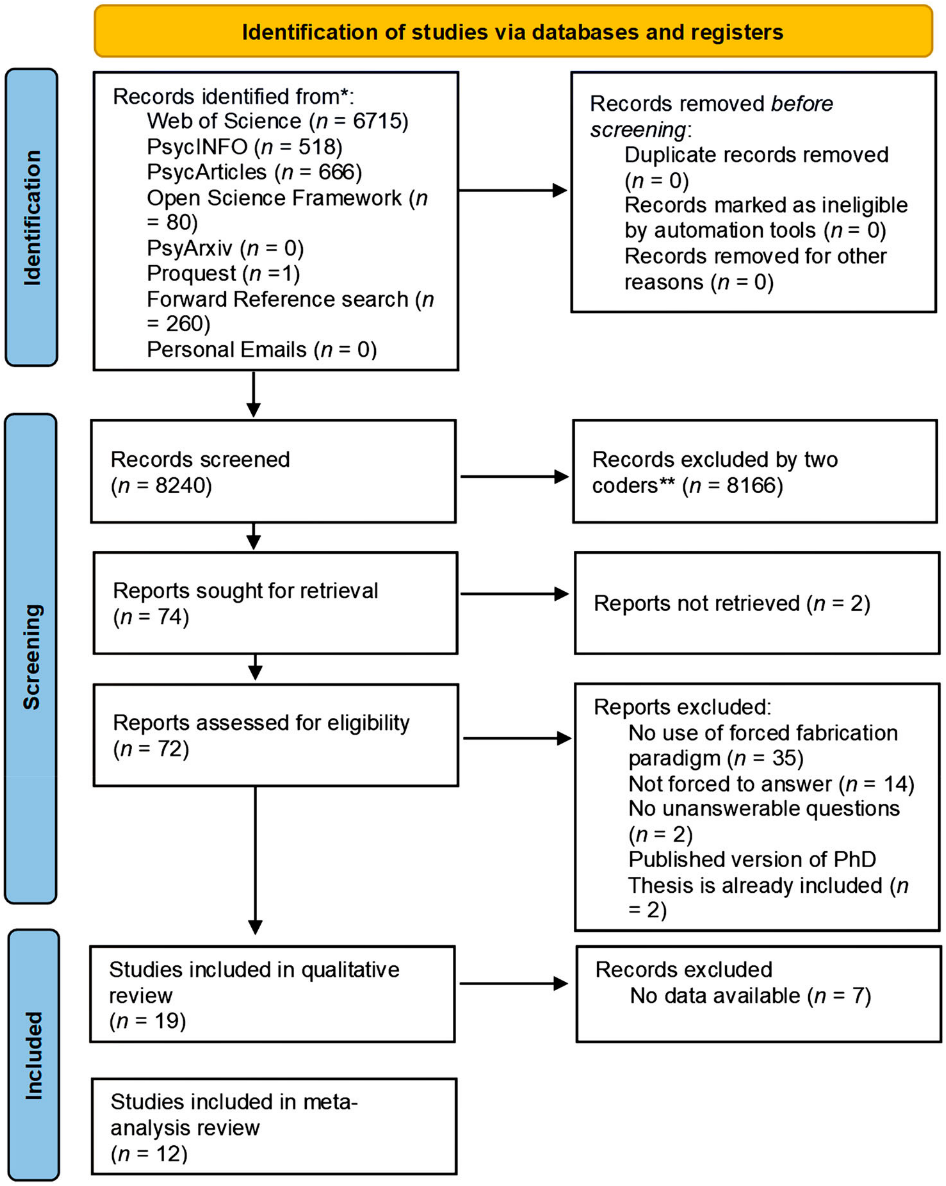

Figure 1. PRISMA workflow chart of literature search.

Method

Literature search

We used the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) statement for our meta-analysis (Moher et al., 2009). The completed PRISMA checklist for this registered report is available on the Open Science Framework (OSF; https://osf.io/4mqpx/). Moreover, a PRISMA workflow chart was implemented to organise our search results and showed the selection of studies (see Figure 1; Moher et al., 2009). A review paper concerning the effects of lying on memory (Otgaar & Baker, 2018) identified twelve articles that examined the effects of forced fabrication. However, following this paper, there was a surge of research on the topic of fabrication and memory (e.g., Battista et al., 2020; Vo et al., 2021). Hence, to identify all studies examining the forced confabulation effect we queried the databases PsycINFO, PsycArticles and all editions of the Web of Science Core Collection for our literature review. Via Google Scholar, we performed a forward reference search for the original study by Ackil and Zaragoza (1998). Moreover, we searched for dissertations and theses via ProQuest Dissertations & Theses Global and we explored OSF and PsyArXiv for any unpublished data. We used the following keywords to appear anywhere in a full article: (confabulation OR fabrication AND memory OR interview). To retrieve additional or unpublished data we advertised an open call via social media, emailed authors who published about the forced confabulation effect, and advertised via associations related to the forced confabulation effect (e.g., Society for Applied Research in Memory and Cognition, European Association for Psychology and Law) (see appendix for email). We did not receive any additional data.

Our literature search focused on articles published from: 31 December 1997, until 31 January 2022. This initial search led to 8240 potential articles. To determine which articles were to be considered for the meta-analysis, two raters (1st and 2nd author) conducted the literature search and in – or excluded studies based on the criteria below. During the title and abstract screening round of our literature search process, each rater independently decided which articles were to be included in the full-text screening round. To include as many articles as possible, the selected articles of both raters moved on to the full-text screening round. That is, if one rater indicated that an article should be included, then it would automatically be included in the full-text screening round. During the abstract and screening round, 8166 articles were excluded.[1] Then, the selected articles ($n = 74$) went through a full-text screening round and each rater decided independently whether the article had to be included in the meta-analysis. If both raters agreed, the article was included. For articles that the raters disagreed upon, a final decision was made by all authors. If the majority of the authors voted in favour of including an article, the paper would be included. Otherwise, the article was excluded from the meta-analysis. However, there were no disagreements among the two raters which led to the inclusion of 19 studies. Afterwards, the two raters (1st and 2nd author) independently coded all the 19 studies and extracted the data and conducted the quality assessments from the included articles.

Inclusion and exclusion criteria

After our initial literature search, our first step was to eliminate all duplicates. Next, review articles, commentaries and books were excluded. We only included articles published after the year 1997. The decision for 1997 was because the first work studying the forced confabulation paradigm was conducted in 1998 (Ackil & Zaragoza, 1998). Only articles published in English were included. We only included studies that adopted the procedure of the forced confabulation paradigm (Ackil & Zaragoza, 1998). Specifically, the procedure of the included studies had to adhere to the following criteria: (i) participants watched a particular stimulus (e.g., video, photo, or virtual reality) or performed a certain act, (ii) included questions that were unanswerable (e.g., about details or events never presented) and (iii) some participants were forced to answer such unanswerable questions. The outcome variable was the forced confabulation effect which referred to the number of details and/or events recalled in response to unanswerable questions after coming forward with the truth. If studies were missing the necessary statistics (e.g., mean, standard deviation, sample size), we (1st and 2nd author) contacted the corresponding authors. In total, we contacted all authors and sent five reminders over a 3-month period if the authors did not respond. All corresponding authors responded after the 5th reminder, except for one author, which led to the exclusion of that article. If necessary information was not provided, the study was excluded from the meta-analysis ($n = 7$). In total, we retrieved sufficient data from 12 studies extracting 25 effect sizes ($N = 1387$).

Coding

The two raters (1st and 2nd author) independently coded all 19 studies and extracted the data which was documented in an Excel spreadsheet (see OSF; https://osf.io/4mqpx/). The following study characteristics were extracted: Study design (i.e., within, between, or mixed), independent variables, type of stimuli, randomised assignment, double or single blinded experiment, sample size justification, peer reviewed or not, the assumptions of statistical tests, and the above-mentioned potential moderators. Agreement between the two raters was assessed, but there were no discrepancies. Final decisions were documented in the Excel spreadsheet.

Meta-analytic procedure

To perform the meta-analysis, we needed either the means, standard deviations, number of participants in each group for each study or the test statistics such as the, t or F values, degrees of freedom and sample sizes which allowed us to calculate or convert the necessary information for the meta-analysis. Because some studies used different experimental designs (e.g., within and/or between subjects), we transformed all effect sizes into a common metric (Hedges g') but calculated the design specific estimates of sampling variance (Morris & DeShon, 2002). We only calculated or converted test statistics for independent, dependent t-tests and one-way ANOVAs when the underlying assumption of these tests was to examine the forced confabulation effect directly. We decided to include such calculations and/or conversion to maximise information about the forced confabulation effect. Results of two-way ANOVAs or Chi-squared tests were excluded because the conversion to the common metric (i.e., Hedges g') was impossible or would not provide direct information about the forced confabulation effect (e.g., covariates complicating interpretation).

Additionally, p-values were extracted to perform a z-curve analysis (Bartoš & Schimmack, 2020; Brunner & Schimmack, 2020). If any specific information was lacking in articles based on the specified criteria, we sent a specific data request (see appendix for email). The effect size indicator of this meta-analysis was Hedges g'. Hedges g' was calculated in the statistical software environment R (R Development Core Team, 2016) via the "metafor" package (Viechtbauer, 2015) (see R code/script on OSF; https://osf.io/4mqpx/):

$$Hedges\ g' = \frac{M_1 - M_2}{SD^{*}pooled}$$

Where $M_1 - M_2$ expresses the difference in means which is divided by the pooled and weighted standard deviations ("*SD**pooled"). SD*pooled was calculated as follows:

$$SD^{*}pooled = \sqrt{\frac{(n_1 - 1)SD_1^2 + (n_2 - 1)SD_2^2}{n_1 + n_2 - 2}}$$

Where n stands for sample size and SD stands for standard deviation. The subscript indicates the group.

The meta-analysis was performed on the mean scores of false memories for confabulated items. Furthermore, the meta-analysis was performed in the statistical software environment R using the "metafor" package (Viechtbauer, 2015). Our proposed analysis was to conduct a random effect model to estimate the mean effect size and evaluate the different moderators (e.g., confirmatory feedback) on the forced confabulation effect size. However, because we extracted multiple effect sizes from the same study, we decided to conduct a three-level random effect meta-analysis to account for the non-independency of the effect sizes (Cheung, 2019).[2] We reported 95% confidence and prediction intervals. The former indicates whether the effect sizes were statistically significant (95%CI does not include 0) and its variation, while the latter indicates the interval in which next studies can expect to observe true effect sizes (IntHout et al., 2016).

To visually analyse variability between and within study effect sizes, we used forest plots and inspected specific influence of individual studies on the mean effect size via a baujat plot and the "influence" function of the "metafor" package. Also, we examined the total variability between studies via statistical heterogeneity analyses by calculating the Tau^2, I^2, the between and within cluster heterogeneity and the Q statistic with corresponding 95% confidence intervals through the "metafor" package in the statistical software environment R. All information was presented in the forest plots. Tau^2 consists of the variation among effect sizes between studies (Borenstein et al., 2009). The I^2 statistic can be interpreted on range from 0 to 100% (Higgins et al., 2003). Benchmarks for the I^2 statistic of 25%, 50% and 75% which are identified as low, medium and high heterogeneity, respectively. The I^2 statistic provides the amount of variance on a relative scale (Borenstein et al., 2009). Hence, the I^2 indicates whether the observed variance of a meta-analysis is real variance between studies or simply a sampling error (Higgins et al., 2003).

For the current meta-analysis, we performed subgroup analyses to examine the possible moderators when the I^2 statistic was above 40% to explain the variation between studies. To perform subgroup analyses, it is recommended to have a minimum of four studies in each subgroup when the independent variable is categorical and six to ten studies when it is continuous (Fu et al., 2011). When such requirement is not met, an in-depth qualitative review was provided. Because we only had one study with sufficient information about the moderator whether confirmatory feedback affected the forced confabulation effect and three studies for the moderator about the difference between confabulating a detail or event, we provided qualitative reviews of these two moderators in the Discussion section. Moreover, a test for heterogeneity was performed via the Q statistic and examines whether the dispersion of effect sizes is zero (Borenstein et al., 2009).

Publication Bias. To address potential publication bias in the forced confabulation literature, funnel plots and a z-curve analysis were carried out. Funnel plots were inspected visually and the trim-and-fill procedure was implemented to establish potential publication bias (Duval & Tweedie, 2000) (see OSF for code). A z-curve analysis was performed to address publication bias when heterogeneity is present (Bartoš & Schimmack, 2020; Brunner & Schimmack, 2020). Recent research demonstrated that a z-curve analysis outperformed other publication bias detection methods (e.g., p-curve) when the effect sizes were different between studies and when the distribution of effects sizes is unknown (Brunner & Schimmack, 2020). A z-curve analysis estimates the Expected Discovery Rate (EDR) and the Expected Replicability Rate (ERR) derived from the test statistics provided in the articles (e.g., *F, r, t*, etc.). When the Observed Discovery Rate (ODR) differs substantially from the EDR, the z-curve analysis suggests that publication bias is present. Additionally, the z-curve analysis gives insights into the probabilities of successful future replication studies via the ERR. Hence, to evaluate the publication bias via z-curve we used the package "z-curve" (Bartoš & Schimmack, 2020, 2022) in the statistical software environment R. The R code/script for the complete analysis is available on OSF (https://osf.io/4mqpx/).

Exploratory analysis

Besides the subgroup analyses for the potential moderators, we also conducted exploratory subgroup analyses for study quality using the following quality moderators: Study design, randomised assignment, double or single blinded experiment, sample size justification, peer reviewed or not, external validity and the assumptions of statistical tests. For each specific item of study quality, we answered a yes/no question to indicate whether the specific study quality item is present (e.g., "Does the study justify their sample size?"). Sample size justifications were not reported because only one study performed a power analysis, and no justifications were provided for any of the other study quality characteristics, except for external validity. Assessing external validity is a complex matter for which no consensus exists on how to evaluate it (Dekkers et al., 2010; Valentine & Cooper, 2008). Hence, we decided to assess the external validity of individual studies using a 5-point Likert scale (*1 = low external validity*, *5 = high external validity*) and we reported our explanations for why a study received a certain score (see excel spreadsheet on OSF).

Based on the results of all quality moderators, two raters decided independently whether a study had low or high quality using a 5-point Likert scale (*1 = low quality*, *5 = high quality*). These decisions and justifications for the allocation of a study to either category of low or high study quality was reported in the excel spreadsheet (see OSF). This entire procedure avoided the use of quality composition scores which have been criticised for its possible misleading information and allowed for transparency in our decisions on study quality (Feeley, 2020; Valentine & Cooper, 2008). We examined whether individual or a combination of quality moderators influenced the effect size of the forced confabulation effect as recommend by Feeley (2020).

Moreover, we proposed to explore whether the content of stimuli (e.g., criminal nature vs. other) or different delay periods between lying and final memory test moderated the forced confabulation effect. Examining the content of stimuli can provide information on whether the forced confabulation effect is also detected when more ecologically valid stimuli are used (e.g., watching a mock crime). However, this was not possible for the content of stimuli because only one study used content that was not of criminal nature. Additionally, in police investigations, witnesses, suspects and/or victims are sometimes interviewed after different delay periods (e.g., Griffiths, 2008). We conducted an exploratory moderator analysis to examine whether delay moderated the forced confabulation effect.

Results

Forced confabulation effect

Our proposed random effects model without considering the dependency of effect sizes revealed a mean effect size (Hedges' *g*) of .39, 95%CI [.10; .69], 95%PI [−1.04; 1.83] (see our R script and analyses on OSF). The heterogeneity values were: $Tau^2 = .52$, 95%CI [.29; 1.06], $I^2 = 92.3\%$, 95%CI [87.2; 96.1], Cochran's Q $(24) = 298.94$, $p < .001$ showing high levels of heterogeneity. Visual inspection of the baujat plot and influence output showed that four effect sizes were influential (see R script for visualisations and additional information).[3] However, as previously mentioned, we extracted multiple effect sizes from individual studies which violated the assumption of independency (Cheung, 2019).

To account for the dependency among effect sizes, we conducted a three-level random effects meta-analysis which showed a mean effect size (Hedges' *g*) of .61, 95% CI [.21; 1.00], 95%PI [−.84; 2.05] (see Figure 2). The overall I^2 was 92.1% of which 19.7% was attributed to within cluster heterogeneity and 72.4% between cluster heterogeneity, Cochran's Q $(24) = 298.94$, $p < .001$. Results showed that the three-level meta-analysis model was indeed a better fit compared with the two-level meta-analysis model (see Table 1).

Moderator analyses

Age Group. To examine whether children were more susceptible to the forced confabulation effect than adults, we conducted sub-group analyses and fitted a meta-regression model to test for subgroup differences (see Figure 3; for an overview of all studies and moderators see Table 2).

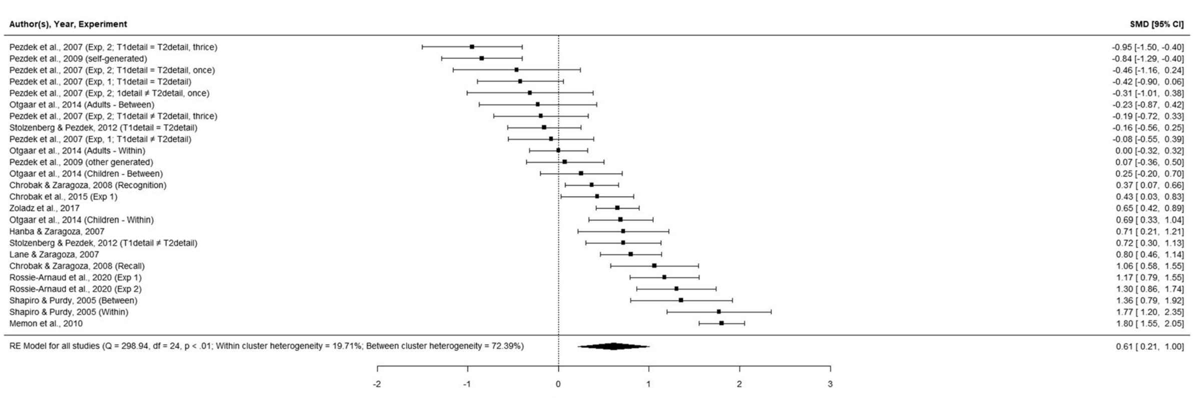

Figure 2. Forest plot of all studies.

Table 1. Model comparison between three-level model including between and within cluster heterogeneity and two-level models.

Model	df	AIC	BIC	AICc	LogLik	LRT	*p*	QE
3-level	3	48.6	52.2	49.8	−21.3			298.9
2-level without between cluster variance	2	58.9	61.3	59.5	−27.5	12.3	.0005	298.9
2-level without within cluster variance	2	60.8	63.1	61.4	−28.4	14.1	.0002	298.9

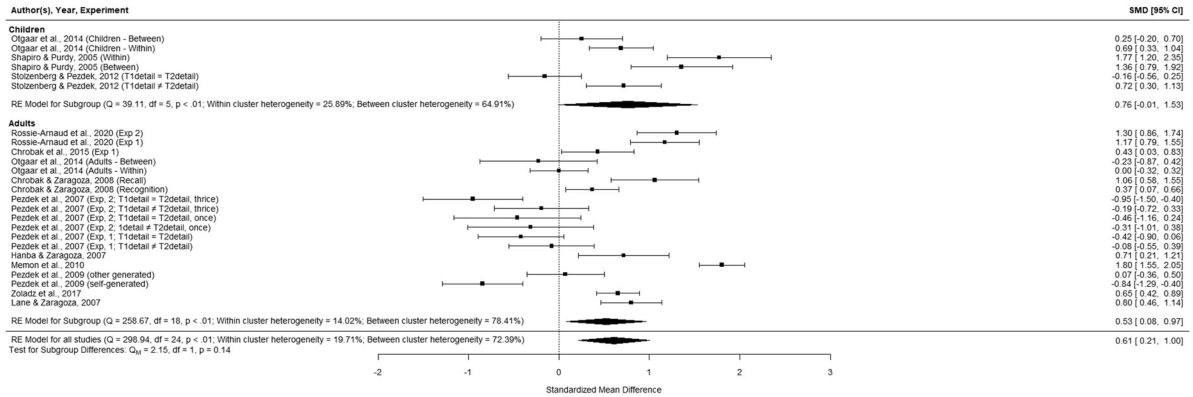

Figure 3. Forest plot with the analysis of the moderator "Age".

Results showed that age group was not a statistically significant moderator of the forced confabulation effect, Q_m (1) = 2.15, $p = .14$. Interestingly, the three-level random effects model for children showed a mean effect size (Hedges *g*′) of .76, 95%CI [−.01; 1.53], 95%PI [−83; 2.36] which was not statistically significant (95% confidence interval includes 0). Overall heterogeneity (I^2) of the mean effect size for children was 90.8% of which 25.89% was within cluster heterogeneity and 64.91% was between cluster heterogeneity, Cochran's *Q* (5) = 39.12, $p < .001$. For adults the mean effect size was .53, 95%CI [.08; .97], 95%PI [−.93; 1.98]. Overall heterogeneity (I^2) of the mean effect size for adults was 92.43% of which 14.02% was within cluster heterogeneity and 78.41% was between cluster heterogeneity, Cochran's *Q* (18) = 258.68, $p < .001$.

Entire Events versus Details. There were only three articles (Chrobak et al., 2015; Chrobak & Zaragoza, 2008, 2013) that examined the forced confabulation effect for entire events, with a total of six experiments being reported. From these six experiments, we were able to extract, calculate or convert only three effect sizes. This was insufficient to perform subgroup or moderator analyses (Fu et al., 2011). Hence, we conducted a qualitative review of these three studies and provided an interpretation of the results in the Discussion section.

Confirmatory Feedback. In total, there were four studies with five experiments (Frost et al., 2003, 2006; Hanba & Zaragoza, 2007; Zaragoza et al., 2001) that examined the effect of confirmatory feedback on the forced confabulation effect. Of these five experiments, only one effect size could be calculated. Thus, we qualitatively reviewed these four studies and their results and implications are analysed in the Discussion section.

Type of Confabulation. We also examined whether forced or voluntary confabulation differentially affected the forced confabulation effect. To do so, we performed subgroup analyses and fitted a meta-regression model (see Figure 4). Our analyses showed that the type of confabulation moderated the forced confabulation effect, Q_m (1) = 16.62, $p < .001$. That is, when participants voluntarily confabulated a response to unanswerable questions, the effect seemed to be stronger than when their confabulations were forced. This effect was demonstrated by the mean effect size of the forced confabulation versus voluntary confabulation which was −.36, 95%CI [−.60; −.12], 95%PI [−.91; .19], indicating a reversed effect. Overall heterogeneity (I^2) of the mean effect size for voluntary fabrications was 49.91% which was completely attributed to the within cluster heterogeneity, Cochran's *Q* (8) = 15.74, $p = .05$. The mean effect size of forced confabulation versus control items was .87, 95%CI [.55; 1.20], 95%PI [−.18; 1.93]. Overall, heterogeneity (I^2) of the mean effect size for forced fabrications was 87.86% of which 25.15% was within cluster heterogeneity and 62.71% was between cluster heterogeneity, Cochran's *Q* (15) = 138.47, $p < .01$.

Table 2. Overview of included studies in meta-analysis.

	Studies	ES	*N*	WS	BS	LQ	MQ	HQ	SD	LD
All studies	12	25	1387	7	18	9	8	8	5	20
Moderator Age										
Children	3	6	231	2	4	0	2	4	0	6
Adults	9	19	1156	5	14	9	6	4	5	14
Moderator type of Confabulation										
Forced vs Control	9	16	308	7	9	1	7	8	5	11
Forced vs Voluntary	3	9	1079	0	9	8	1	0	0	9

Note. ES = number of effect sizes, *N* = total sample size, WS = number of effect sizes from within subject designs, BS = number of effect sizes from between subject designs, LQ = number of effect sizes from lower quality studies, MQ = number of effect sizes from medium quality studies, HQ = number of effect sizes from higher quality studies, SD = number of effect sizes from studies with shorter delays, LD = number of effect sizes from studies with longer delays.

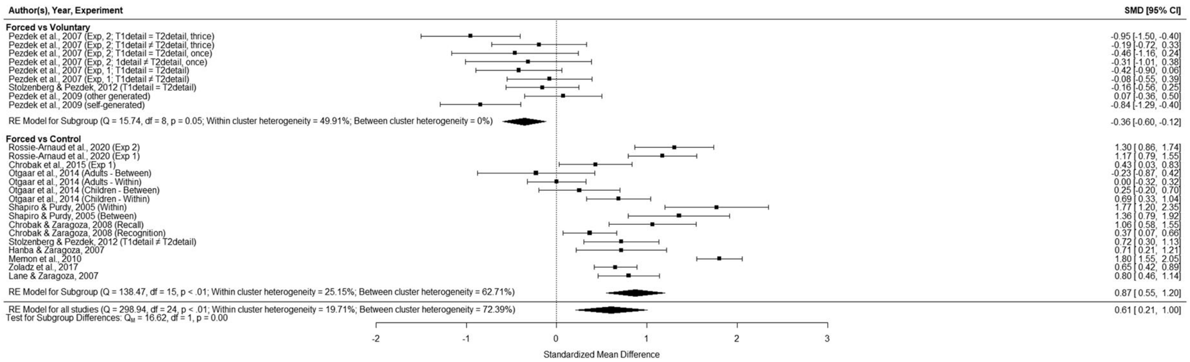

Figure 4. Forest plot with the analysis of the moderator "Type of Confabulation".

Publication bias

Visual inspection of the funnel plot and the trim and fill procedure did not find indications of asymmetry (no missing studies on right side; see R script for visualisations). Moreover, a z-curve analysis (Bartoš & Schimmack, 2022) was performed using 33 *p*-values of which 30 were statistically significant. The z-curve analyses showed the following: Observed discovery rate = .91, 95%CI [.75; .98], expected discovery rate = .89, 95%CI [.27; .99] and an expected replicability rate = .90, 95%CI [.67; .98] (see Figure 5).

Exploratory analyses

We explored whether study quality, study design, content of stimuli (criminal nature vs. other) and delay periods between lying and final memory test moderated the forced confabulation effect. We were unable to conduct a subgroup analysis and meta-regression for the content of stimuli because only one study used content not criminal in nature.

Study Quality. Study quality rating ranged from 2 to 4 because there were no studies deemed 1 = low quality or 5 = high quality. Hence, we performed a subgroup analysis and meta-regression wherein we examined the differences between lower quality (= 2), medium quality (= 3) and higher quality (= 4). Study quality did not statistically significantly moderate the forced confabulation effect, Q_m (2) = 5.00, *p* = .08.

Study Design. We also examined whether the forced confabulation effect was different for within-subject designs compared with between-subject designs. A test of moderator showed that study design moderated the forced confabulation effect, Q_m (1) = 5.05, *p* = .03. Subgroup analyses showed that the mean effect size for within-subject designs was 1.12, 95%CI [.55; 1.70], 95%PI

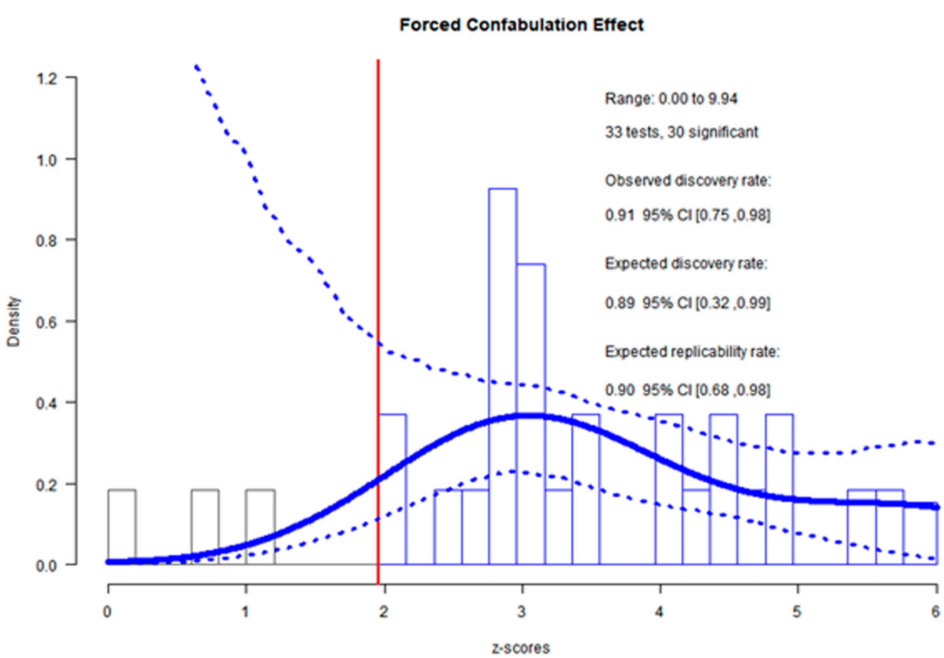

Figure 5. Z-curve analysis of the forced confabulation literature.

[−28; 2.53], with a total heterogeneity (I^2) of 93.71% of which 21.76% was attributed to within cluster heterogeneity and 71.94% to between cluster heterogeneity, Cochran's Q (6) = 97.34, $p < .0001$. The mean effect size for between-subject designs was .32, 95%CI [−.06, .70], 95%PI [−.89; 1.53], which was not statistically significant (95%CI includes 0), meaning that for between-subjects designs, no sign of the forced confabulation effect emerged. Total heterogeneity (I^2) was 86.42% of which 31.88% was attributed to within cluster heterogeneity and 54.54% to between cluster heterogeneity, Cochran's Q (17) = 111.56, $p < .001$.

Delay. Most studies used a 7-day delay between the forced fabrication phase and the final memory task. One study used an 8-week delay and four studies used a shorter delay. We dichotomised the variables into shorter delays (delays less than 7 days) and longer delays (7-days or longer). A test of moderator did not find any statistically significant result, Q_m (1) = 0.02, $p = 0.89$.

Study-Level Statistical Power. Using the "metameta" package (Quintana, 2022), we calculated the study level statistics power of the studies included in the meta-analysis had to detect the observed mean effect size of Hedges $g' = .61$. Studies had a power .76 to detect the mean effect size (see R script/code on OSF for visualisation and power for a range of effect sizes).

Discussion

We conducted a meta-analysis and systematic review to examine the forced confabulation effect. This effect refers to the finding that forcing participants to confabulate in response to unanswerable questions can lead them to later report such confabulated details as part of the experienced event. Furthermore, we investigated whether certain demographic or manipulation variables (i.e., age group, details vs entire events, confirmatory feedback and type of confabulation) moderated the forced confabulation effect.

In line with our main hypothesis (H1), and the initial study of Ackil and Zaragoza (1998), we found evidence of a forced confabulation effect with an average Hedges ǵ of .61. This means that participants who were forced to confabulate reported more details and/or events in response to unanswerable questions than participants who told the truth. In terms of the common language effect size (McGraw & Wong, 1992), a Hedges g' of .61 indicates that 72.9% of participants in the forced confabulation group will have a higher mean of confabulated details/events than participants in the control group (Cohen's U3). Moreover, there is a 66.7% probability that a participant randomly selected from the "forced confabulation" group will have reported more confabulated details/events than a participant randomly selected from the control group. However, the forced confabulation effect should be interpreted in the light of our moderator analyses.

Age group

For instance, in contrast to our prediction (H2a), the moderator analysis did not show that children were more susceptible to the forced confabulation effect than adults. However, subgroup analyses revealed that no forced confabulation effect existed for children, while there was an effect for adults. The forced confabulation effect for children might have not emerged because of the low number of studies that could be included in the meta-analysis (3 studies, 6 effect sizes), which is reflected in its large 95%CI (Hedges' $g = .76$, 95%CI [−.01, 1.53]). Moreover, the results of the original study by Ackil and Zaragoza (1998) were missing in the current meta-analysis which, based on visual inspection of their results, showed a stronger forced confabulation effect for children than adults (Figure 1; Ackil & Zaragoza, 1998). Interestingly, in their study, the forced confabulation effect seemed to be stronger for younger children (aged 6–8) than for older children (age 8–11). However, as we only had limited studies examining the forced confabulation effect in children, we dichotomised our moderator variable into children (< 18 years) and adults (> 18 years) and, thus, we were unable to examine this developmental trend. Furthermore, because our moderator analysis indicated that there was no statistically significant difference in the forced confabulation effect between children and adults, the subgroup analyses should be interpreted with caution. A direct preregistered replication of the original study (Ackil & Zaragoza, 1998) should be conducted to examine whether forced confabulation effect is developmentally invariant.

Entire events versus details

We also examined whether the forced confabulation effect was stronger for details than entire events (H2b). We were unable to perform reliable subgroup analyses and/or meta regression, because we were only able to use data from two studies, resulting in three effect sizes (Chrobak & Zaragoza, 2008, 2013; Fu et al., 2011). Therefore, this moderator was examined through a qualitative review. In their first work, Chrobak and Zaragoza (2008) studied the effects of forced confabulation about entire events on memory after a 1-week and 8-week delay. In these studies, some participants were warned before the final memory task that unanswerable questions were included while others were not. Interestingly, the authors observed an overall forced confabulation effect for entire events after a 1-week delay using a recognition task. However, follow-up analyses revealed that this effect was mainly caused by the experimental and control groups that were not warned about the unanswerable questions. So, they failed to find evidence for a statistically significant difference between the warned groups (forced confabulation versus control) in terms of responses to unanswerable questions about entire confabulated events. However, after an 8-week delay and using a free recall task, they found robust evidence for the forced confabulation effect for both warned and unwarned groups, even when participants initially rejected the confabulated events after a 1-week delay (Chrobak & Zaragoza, 2008). Similar results were found after a 6-week delay and when using a free recall task (Chrobak & Zaragoza, 2013, Exp. 1) and recognition task (Chrobak & Zaragoza, 2013; Exp. 2).

These results indicated that longer delays between the forced confabulation phase and the final memory task are necessary to observe a reliable forced confabulation effect for entire events. Why delay influences the confabulation effect for entire statements can be explained using the Source Monitoring Framework (Johnson et al., 1993). This framework postulates that the initial act of confabulation increases the amount of attributed perceptual, affective and/or contextual features of the confabulated event. The more perceptual, affective and/or contextual features ascribed to a confabulated event, the more likely it is misremembered as an experienced event. Moreover, with the passage of time and weakening of the memory trace, it becomes more difficult to monitor the specific features of an event and, thus, the source of the information (Ebbinghaus, 1885; Murre & Dros, 2015; Radvansky et al., 2022). Consequently, as memory traces fade, it can become more difficult to distinguish between whether the event was confabulated or experienced.

Interestingly, our exploratory analysis did not find evidence that delay between the forced fabrication phase and the final memory task moderated the overall forced confabulation effect. However, due to the low number of studies, this comparison was only between short delays (less than 7 days) and longer delays (7 days or more) wherein only one study had a delay period longer than 7 days (most studies used a 7-day delay). Moreover, this analysis was based on the overall forced confabulation effect and did not distinguish between the effect for details and entire events. Future studies could directly examine the role of delay in the forced confabulation effect for details and entire events.

Confirmatory feedback

Another potential moderator variable that we examined was the effect of confirmatory feedback on the forced confabulation effect (H2c). However, we identified four studies of which only one had the necessary information to be included in the meta-analysis. Hence, this moderator was also examined through a qualitative review (Frost et al., 2003, 2006; Hanba & Zaragoza, 2007; Zaragoza et al., 2001). Three studies found evidence for the typical within-subject forced confabulation effect (Frost et al., 2003, 2006; Zaragoza et al., 2001).[4] That is, participants endorsed more details that they initially were forced to fabricate on the final recognition task compared with completely new false details. As predicted, it was demonstrated that the forced confabulation effect was stronger after receiving confirmatory feedback compared with neutral feedback.

The qualitative review also revealed that confirmatory feedback increased participants' confidence in the confabulated items. The general notion is that when participants' confidence in a (true or false) detail and/or event is high, the more likely they report this information later on (Smalarz & Wells, 2020). Interestingly, providing participants with confirmatory feedback led to higher confidence in confabulated details than neutral feedback (Zaragoza et al., 2001). Moreover, confirmatory feedback indirectly increased confidence in confabulated details because it led participants to repeat the same confabulated details more often, to more quickly generate the confabulated details, and decreased their doubts (Hanba & Zaragoza, 2007). The remaining two studies did not examine whether confirmatory feedback increased the confidence in the confabulated details (Frost et al., 2003, 2006). Taken together, the results are in line with related research examining factors affecting eyewitness identification research showing that post-identification confirmatory feedback can increase witnesses' confidence in their identification, especially when they are mistaken (Steblay et al., 2014; Wixted & Wells, 2017). Considering the forced confabulation effect, it seems that confirmatory feedback after forcedly confabulating in response to unanswerable questions, increased participants' confidence in these confabulations leading them to report such confabulated details as part of the truth later on.

Thus, receiving confirmatory feedback after forcedly confabulating made participants more inclined to endorse false details that they initially forcefully confabulated and did so more consistently (i.e., repeated the same answer) than when they received neutral feedback. These effects were even observed after longer delays (1–2 months; Zaragoza et al., 2001). Even though we were unable to examine the effect of feedback meta-analytically, our qualitative review suggested that giving confirmatory feedback seems to amplify the forced confabulation effect.

Type of confabulation

We further examined whether the group or items that were used as comparison (i.e., control group) played an important role into the forced confabulation effect (H2d). Indeed, our analyses showed that the type of confabulation (confirmatory analysis) and study design (exploratory analysis) moderated the forced confabulation effect. For example, in the studies by Pezdek et al. (2007, 2009), the control group only consisted of participants who had voluntarily confabulated a response to unanswerable questions during the first interview. Our moderator analyses showed that compared against such a control group, the confabulation effect was even reversed (Hedges $g' = -.36$). Specifically, when comparing participants who were *forced* to fabricate with those who did so *voluntarily*, results showed that voluntary confabulations were more likely to elicit false memories than forced fabrications.

However, the comparison between voluntary and forced fabrications is not a fair one to assess the forced confabulation effect. More specifically, in the control groups of Pezdek and colleagues' studies (2007, 2009), many participants were excluded from the data analysis because they did not (voluntarily) confabulate a response to unanswerable questions during the first interview. Not only did the exclusion lead to lower statistical power and unbalanced comparisons (different sample sizes in each group), it also excluded a group of participants who are of great interest for examining the forced confabulation effect. Comparing participants who did not confabulate in response to unanswerable questions with participants who were forced can give valuable insights in how and why false memories occur. This comparison can provide evidence to what extent forcing participants to confabulate leads to false memories, above and beyond the false memories possibly elicited due to the presentation of misinformation (i.e., unanswerable questions).

Finally, participants who voluntarily confabulated might have mistakenly believed that these confabulations actually occurred. For these participants, the act of posing unanswerable questions might have served as some form of misinformation that they could falsely mention in later memory reports (Loftus, 2005; Zaragoza et al., 2006). If unanswerable questions were indeed a source of misinformation, a fairer comparison would be to compare participants in the forced and voluntary confabulation groups who provided an answer immediately. In conclusion, due to the selective comparisons in Pezdek et al. (2007, 2009) analyses, their forced confabulation effect is unlikely to provide a valid estimate.

Forced confabulation and misinformation effect

Interestingly, our exploratory analyses showed that the forced confabulation effect was stronger in within-subject designs compared with between-subject designs. An important element of within-subject design studies was the comparison between items that were forcefully fabricated during the initial interview and items that were never presented during the initial interview, but only appeared on the final memory task. However, asking unanswerable questions may be a confounding variable in within-subject design studies. More specifically, it might well be the case that presenting participants with misinformation in the form of unanswerable questions during an initial interview will lead them to report this misinformation later on, even when these questions did not receive a forced confabulation response (Loftus, 2005). If true, then such items could lead to misinformation effects in which no forced confabulation is needed. To test this possibility, it is crucial to examine studies in which control groups were added wherein no forced confabulation instructions were given. An analysis of the data of Otgaar et al. (2014) and Shapiro and Purdy (2005), who used both a between (i.e., forced confabulation group vs control group) and within-subject (i.e., forced confabulated items vs control items) designs, indeed showed a misinformation effect within the control groups. Specifically, participants in the control group responded more often in response to previously asked unanswerable questions, than new unanswerable questions. This finding suggests that presenting participants with unanswerable questions (i.e., misinformation) may partially account for the observed findings in the forced confabulation literature. Hence, perhaps the forced confabulation effect is not completely caused by the forced confabulation, but also by the presentation of unanswerable questions which act as misinformation.

One potential explanation for the similar findings between the forced confabulation and misinformation effects stems from the *discrepancy detection* principle (Tousignant et al., 1986). This principle posits that misinformation is included into memory reports when people are unable to detect the discrepancy between experienced events and post-event misinformation. When the trace of a memory is strong, participants will be more likely to detect such discrepancies, while less so when the memory trace is weak. Moreover, if the post event misinformation is rather different or completely new, it might be more easily detected and therefore rejected from inclusion in the memory report. In light of the forced confabulation effect, it is possible that participants were more likely to detect that completely new unanswerable questions were misinformation in comparison with the repeated unanswerable questions.

Publication bias and limitations

Concerning the publication bias, our funnel plots, trim and fill procedures and z-curve analysis (Bartoš & Schimmack, 2022) did not detect problematic evidence in the forced confabulation literature. Moreover, we did not receive any unpublished findings when requesting data from authors, which shows that publication bias is likely less of an issue in this line of research. Additionally, we found that studies had generally sufficient statistical power to detect the observed mean effect of the current meta-analysis. Taken together, this suggests that the current meta-analysis was based on robust and reliable studies.

Some limitations need to be addressed regarding the interpretation of our results. First, we were only able to retrieve necessary data of 12 out of the 19 studies. This prevented us from performing several moderator analyses but also led to a less precise estimate of the forced confabulation effect. For future work in this field, we urge researchers to publicly store their data and materials, or at least provide full statistical results including the summary data (i.e., cell means, standard deviations and sample sizes) of statistical significant and null findings.

This will benefit future meta-analyses, replication studies, checks of reproducibility (reproduce same analysis with same data) and robustness (reproduce results using same data with different analysis), but also yields personal advantages (e.g., more citations; McKiernan et al., 2016). Also, we were unable to retrieve the data or extract the necessary statistics of the primary study by Ackil and Zaragoza (1998), which hampered us in further investigating why our meta-analytic results contrasted their findings. Hence, to further examine the robustness of the forced confabulation effect, a direct preregistered replication of the original study is necessary.

A second issue that hinders the interpretation of the current meta-analytic findings is the variety among studies in terms of study design (i.e., between-subject, within-subject designs, mixed), control groups (e.g., all honest participants, only voluntary confabulating participants, or control items), types of analyses (e.g., signal detection analyses, consistent vs inconsistent reporting), but also a lack of reporting of all measured details (e.g., amount of unanswerable questions or control questions). Our moderator analyses indicated that the forced confabulation effect depends on the type of control group and subject design. Even though we were able to convert many effect sizes to a common metric to maximise the amount of information included following the recommendations of Morris and DeShon (2002), it is possible that the mean effect size is over- or underestimated due to the exclusion of several relevant studies (Bartoš et al., 2022).

Furthermore, most studies examined the forced confabulation effect from a witness perspective. However, recent research examining the effects of confabulation on memory for a mock crime from a perpetrator perspective revealed that it can yield different mnemonic consequences such as forgetting (Riesthuis et al., 2022). Because victims and suspects are sometimes also forced to answer unanswerable questions during legal proceedings (Kassin, 2006), future research could examine the mnemonic effects of forced confabulation by comparing the two different perspectives. Moreover, the stimuli typically used in forced confabulation studies are videos of mock crimes. To assess the generalizability of the forced confabulation effect, future research could investigate the robustness of the effect when stimuli more reflective of real-life situations are used (e.g., performing a mock crime, virtual reality scenes).

Conclusion and practical implications

Taken together, our meta-analysis and systematic review showed support for the forced confabulation effect. That is, people can form false memories for their confabulations, especially after receiving confirmatory feedback (vs neutral feedback). Interestingly, children did not seem to form more false memories for their forced confabulations as compared with adults. However, caution is needed when interpreting our meta-analytical results. That is, the results showed that the forced confabulation effect was moderated by the type of control group. Specifically, voluntary confabulations led to more false memory formation for such confabulations than forced confabulations. Moreover, our exploratory analysis suggested that a forced confabulation effect might even emerge when there is no instruction to forcedly confabulate (Otgaar et al., 2014; Shapiro & Purdy, 2005), likely suggesting a role of the misinformation effect when asking unanswerable questions (Loftus, 2005).

Hence, our meta-analysis suggests that forced confabulation can lead to false memories for such confabulations and that its underlying mechanism might be closely connected to the misinformation effect. These results highlight the need for police investigators to avoid closed/suggestive questions. These types of questions can lead to false memories especially when people are forced to answer questions they don't know the answer to, and even more so when interviewees receive confirmatory feedback afterwards. Moreover, open questions should be used because they lead to longer, more detailed and more accurate reports than closed questions (Oxburgh et al., 2010). Hence, in line with previous recommendations (Ackil & Zaragoza, 1998), people should not be forced to answer unanswerable questions because this may lead to false memories for such confabulations resulting in unreliable testimonies.

Notes

1. Many articles were excluded in the abstract and screening round because "fabrication" is a commonly used term in physics and therefore many articles were not eligible.
2. We also presented the proposed random effect model in the Results section.
3. Removing influential studies did not alter the pattern of results. Hence, we decided to include the studies to maximise information about the forced confabulation effect.
4. The study by Hanba and Zaragoza (2007) did not report these results.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

Fonds Wetenschappelijk Onderzoek, Grant/Award Number: g0d3621n; Onderzoeksraad, KU Leuven, Grant/Award Number: zkd5384. Both were awarded to the second author.

ORCID

Paul Riesthuis http://orcid.org/0000-0001-6520-2453
Ivan Mangiulli http://orcid.org/0000-0002-5409-7325

References

Ackil, J. K., & Zaragoza, M. S. (1998). Memorial consequences of forced confabulation: Age differences in susceptibility to false memories. *Developmental Psychology*, *34*(6), 1358–1372. https://doi.org/10.1037/0012-1649.34.6.1358

Ackil, J. K., & Zaragoza, M. S. (2011). Forced fabrication versus interviewer suggestions: Differences in false memory depend on how memory is assessed. *Applied Cognitive Psychology*, *25*(6), 933–942. https://doi.org/10.1002/acp.1785

Bartoš, F., Maier, M., Shanks, D., Stanley, T. D., Sladekova, M., & Wagenmakers, E. J. (2022). *Meta-Analyses in psychology often overestimate evidence for and size of effects*. https://doi.org/10.31234/osf.io/tkmpc

Bartoš, F., & Schimmack, U. (2020). *zcurve: An R Package for Fitting Z-curves*. R package version 1.0.0, https://www.github.com/FBartos/zcurve.

Bartoš, F., & Schimmack, U. (2022). Z-curve 2.0: Estimating replication rates and discovery rates. *Meta-Psychology*, *6*. https://doi.org/10.15626/MP.2021.2720

Battista, F., Mangiulli, I., Herter, J., Curci, A., & Otgaar, H. (2020). The effects of repeated denials and fabrication on memory. *Journal of Cognitive Psychology*, *32*(4), 369–381. https://doi.org/10.1080/20445911.2020.1767626

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. (2009). *Introduction to meta-analysis*. Wiley.

Brunner, J., & Schimmack, U. (2020). Estimating population mean power under conditions of heterogeneity and selection for significance. *Meta-Psychology*, *4*. https://doi.org/10.15626/MP.2018.874

Cheung, M. W. L. (2019). A guide to conducting a meta-analysis with non-independent effect sizes. *Neuropsychology Review*, *29*(4), 387–396. https://doi.org/10.1007/s11065-019-09415-6

Chrobak, Q. M., Rindal, E. J., & Zaragoza, M. S. (2015). The impact of multifaceted questions on eyewitness accuracy following forced fabrication interviews. *The Journal of General Psychology*, *142*(3), 150–166. https://doi.org/10.1080/00221309.2015.1060188

Chrobak, Q. M., & Zaragoza, M. S. (2008). Inventing stories: Forcing witnesses to fabricate entire fictitious events leads to freely reported false memories. *Psychonomic Bulletin & Review*, *15*(6), 1190–1195. https://doi.org/10.3758/PBR.15.6.1190

Chrobak, Q. M., & Zaragoza, M. S. (2013). When forced fabrications become truth: Causal explanations and false memory development. *Journal of Experimental Psychology: General*, *142*(3), 827–844. https://doi.org/10.1037/a0030093

Dekkers, O. M., Elm, E. V., Algra, A., Romijn, J. A., & Vandenbroucke, J. P. (2010). How to assess the external validity of therapeutic trials: A conceptual approach. *International Journal of Epidemiology*, *39*(1), 89–94. https://doi.org/10.1093/ije/dyp174

Drivdahl, S., & Zaragoza, M. (2001). The role of perceptual elaboration and individual differences in the creation of false memories for suggested events. *Applied Cognitive Psychology*, *15*(3), 265–281. https://doi.org/10.1002/acp.701

Duval, S., & Tweedie, R. (2000). Trim and fill: A simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics*, *56*(2), 455–463. https://doi.org/10.1111/j.0006-341X.2000.00455.x

Ebbinghaus, H. (1885). *Über das gedächtnis: untersuchungen zur experimentellen psychologie*. Duncker & Humblot.

Feeley, T. H. (2020). Assessing study quality in meta-analysis. *Human Communication Research*, *46*(2-3), 334–342. https://doi.org/10.1093/hcr/hqaa001

Frenda, S. J., Nichols, R. M., & Loftus, E. F. (2011). Current issues and advances in misinformation research. *Current Directions in Psychological Science*, *20*(1), 20–23. https://doi.org/10.1177/0963721410396620

Frost, P., Lacroix, D., & Sanborn, N. (2003). Increasing false recognition rates with confirmatory feedback: A phenomenological analysis. *Psychology Faculty Papers*. http://hdl.handle.net/10474/583.

Frost, P., Sparrow, S., & Barry, J. (2006). Personality characteristics associated with susceptibility to false memories. *The American Journal of Psychology*, *119*(2), 193–204. https://doi.org/10.2307/20445334

Fu, R., Gartlehner, G., Grant, M., Shamliyan, T., Sedrakyan, A., Wilt, T. J., Griffith, L., Oremus, M., Raina, P., Ismaila, A., Santaguida, P., Lau, J., Trikalinos, T. A., & Trikalinos, T. A. (2011). Conducting quantitative synthesis when comparing medical interventions: AHRQ and the effective health care program. *Journal of Clinical Epidemiology*, *64*(11), 1187–1197. https://doi.org/10.1016/j.jclinepi.2010.08.010

Griffiths, A. (2008). *An examination into the efficacy of police advanced investigative interview training?* [Doctoral dissertation]. University of Portsmouth.

Hanba, J. M., & Zaragoza, M. S. (2007). Interviewer feedback in repeated interviews involving forced confabulation. *Applied Cognitive Psychology*, *21*(4), 433–455. https://doi.org/10.1002/acp.1286

Higgins, J., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *BMJ*, *327*(7414), 557–560. https://doi.org/10.1136/bmj.327.7414.557

IntHout, J., Ioannidis, J. P., Rovers, M. M., & Goeman, J. J. (2016). Plea for routinely presenting prediction intervals in meta-analysis. *BMJ Open*, *6*(7), e010247. https://doi.org/10.1136/bmjopen-2015-010247

Johnson, M. K., Hashtroudi, S., & Lindsay, D. S. (1993). Source monitoring. *Psychological Bulletin*, *114*(1), 3–28. https://doi.org/10.1037/0033-2909.114.1.3

Kassin, S. M. (2006). A critical appraisal of modem police interrogations. In T. Williamson (Ed.), *Investigative interviewing: Rights, research, regulation* (pp. 207–228). Willan.

Lane, S. M., & Zaragoza, M. S. (2007). A little elaboration goes a long way: The role of generation in eyewitness suggestibility. *Memory & Cognition*, *35*(6), 1255–1266. https://doi.org/10.3758/BF03193599

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., Miller, D. G., & Burns, H. J. (1978). Semantic integration of verbal information into a visual memory. *Journal of Experimental Psychology: Human Learning and Memory*, *4*(1), 19–31. https://doi.org/10.1037/0278-7393.4.1.19

Mazzoni, G., & Memon, A. (2003). Imagination can create false autobiographical memories. *Psychological Science*, *14*(2), 186–188. https://doi.org/10.1046/j.1432-1327.2000.01821.x

McGraw, K. O., & Wong, S. P. (1992). A common language effect size statistic. *Psychological Bulletin*, *111*(2), 361. https://doi.org/10.1037/0033-2909.111.2.361

McKiernan, E. C., Bourne, P. E., Brown, C. T., Buck, S., Kenall, A., Lin, J., McDougall, D., Nosek, B. A, Ram, K., Soderberg, C. K., Spies, J. R., Thaney, K., Updegrove, A., Woo, K. H., & Yarkoni, T. (2016). How open science helps researchers succeed. *eLife*, *5*. e16800. https://doi.org/10.7554/eLife.16800

Memon, A., Zaragoza, M., Clifford, B. R., & Kidd, L. (2010). Inoculation or antidote? The effects of cognitive interview timing on false memory for forcibly fabricated events. *Law and Human Behavior*, *34*(2), 105–117. https://doi.org/10.1007/s10979-008-9172-6

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & The PRISMA Group. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLoS Medicine*, *6*(7), e1000097. https://doi.org/10.1371/journal.pmed.1000097

Morris, S. B., & DeShon, R. P. (2002). Combining effect size estimates in meta-analysis with repeated measures and independent-groups designs. *Psychological Methods*, *7*(1), 105. https://doi.org/10.1037/1082-989X.7.1.105

Murre, J. M., & Dros, J. (2015). Replication and analysis of Ebbinghaus' forgetting curve. *PloS one*, *10*, e0120644. https://doi.org/10.1371/journal.pone.0120644

Ofshe, R. J. (1992). Inadvertent hypnosis during interrogation: False confession due to dissociative state; mis-identified multiple personality and the satanic cult hypothesis. *International Journal of Clinical and Experimental Hypnosis*, *40*(3), 125–156. https://doi.org/10.1080/00207149208409653

Otgaar, H., Howe, M. L., Memon, A., & Wang, J. (2014). The development of differential mnemonic effects of false denials and forced confabulations. *Behavioral Sciences & Law*, *32*(6), 718–731. https://doi.org/10.1002/bsl.2148

Otgaar, H., Howe, M. L., Merckelbach, H., & Muris, P. (2018). Who Is the better eyewitness? Sometimes adults but at other times children. *Current Directions in Psychological Science*, *27*(5), 378–385. https://doi.org/10.1177/0963721418770998

Otgaar, H., & A. Baker. (2018). When lying changes memory for the truth. *Memory*, *26*, 2–14. https://doi.org/10.1080/09658211.2017.1340286

Oxburgh, G. E., Myklebust, T., & Grant, T. (2010). The question of question types in police interviews: A review of the literature from a psychological and linguistic perspective. *International Journal of Speech, Language & the Law*, *17*(1). https://doi.org/10.1558/ijsll.v17i1.45

Pezdek, K., Lam, S. T., & Sperry, K. (2009). Forced confabulation more strongly influences event memory if suggestions are other-generated than self-generated. *Legal and Criminological Psychology*, *14*(2), 241–252. https://doi.org/10.1348/135532508X344773

Pezdek, K., Sperry, K., & Owens, S. M. (2007). Interviewing witnesses: The effect of forced confabulation on event memory. *Law and Human Behavior*, *31*(5), 463–478. https://doi.org/10.1007/s10979-006-9081-5

Quintana, D. (2022). *A guide for calculating study-level statistical power for meta-analyses*. https://osf.io/js79t/.

Radvansky, G. A., Doolen, A. C., Pettijohn, K. A., & Ritchey, M. (2022). A new look at memory retention and forgetting. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *48*(11), 1698–1723. https://doi.org/10.1037/xlm0001110

R Development Core Team. (2016). *R: A language and environment for (version 3.3)*. R Foundation for Statistical Computing. http://www.R-project.org/

Riesthuis, P., Otgaar, H., De Cort, A., Bogaard, G., & Mangiulli, I. (2022). Creating a false alibi leads to errors of commission and omission. *Applied Cognitive Psychology*, *36*(4), 936–945. https://doi.org/10.1002/acp.3982

Rossi-Arnaud, C., Spataro, P., Bhatia, D., Doricchi, F., Mastroberardino, S., & Cestari, V. (2020). Long-lasting positive effects of collaborative remembering on false assents to misleading questions. *Acta Psychologica*, *203*, 102986. https://doi.org/10.1016/j.actpsy.2019.102986

Scoboria, A., Wade, K. A., Lindsay, D. S., Azad, T., Strange, D., Ost, J., & Hyman, I. (2017). A mega-analysis of memory reports from eight peer-reviewed false memory implantation studies. *Memory (Hove, England)*, *25*(2), 146–163. https://doi.org/10.1080/09658211.2016.1260747

Shapiro, L. R., & Purdy, T. L. (2005). Suggestibility and source monitoring errors: Blame the interview style, interviewer consistency, and the child's personality. *Applied Cognitive Psychology*, *19*(4), 489–506. https://doi.org/10.1002/acp.1093

Smalarz, L., & Wells, G. L. (2020). Do multiple doses of feedback have cumulative effects on eyewitness confidence? *Journal of Applied Research in Memory and Cognition*, *9*(4), 508. https://doi.org/10.1037/h0101857

Steblay, N. K., Wells, G. L., & Douglass, A. B. (2014). The eyewitness post identification feedback effect 15 years later: Theoretical and policy implications. *Psychology, Public Policy, and Law*, *20*(1), 1. https://doi.org/10.1037/law0000001

Stolzenberg, S., & Pezdek, K. (2013). Interviewing child witnesses: The effect of forced confabulation on event memory. *Journal of Experimental Child Psychology*, *114*(1), 77–88. https://doi.org/10.1016/j.jecp.2012.09.006

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

Valentine, J. C., & Cooper, H. (2008). A systematic and transparent approach for assessing the methodological quality of intervention effectiveness research: The study design and implementation assessment device (study DIAD). *Psychological Methods*, *13*(2), 130. https://doi.org/10.1037/1082-989X.13.2.130

van Oorsouw, K., & Merckelbach, H. (2004). Feigning amnesia undermines memory for a mock crime. *Applied Cognitive Psychology*, *18*(5), 505–518. https://doi.org/10.1002/acp.999

Viechtbauer, W. (2015). *Meta-analysis package for R*. https://cran.r-project.org/web/packages/metafor/metafor.pdf.

Vo, T. V. A., Gunderson, C. A., & Ten Brinke, L. (2021). How deception and believability feedback affect recall. *Memory, 30(6)*, 1–9. https://doi.org/10.1080/09658211.2021.1883064

Wixted, J. T., & Wells, G. L. (2017). The relationship between eyewitness confidence and identification accuracy: A new synthesis. *Psychological Science in the Public Interest*, *18*(1), 10–65. https://doi.org/10.1177/1529100616686966

Wright, L. (1995). *Remembering satan*. Vintage Books.

Zaragoza, M. S., Belli, R. F., & Payment, K. E. (2006). Misinformation effects and the suggestibility of eyewitness memory. In M. Garry, & H. Hayne (Eds.), *Do justice and let the sky fall: Elizabeth Loftus and her contributions to science, law, and academic freedom* (pp. 35–63). Erlbaum.

Zaragoza, M. S., Payment, K. E., Ackil, J. K., Drivdahl, S. B., & Beck, M. (2001). Interviewing witnesses: Forced confabulation and confirmatory feedback increase false memories. *Psychological Science*, *12*(6), 473–477. https://doi.org/10.1111/1467-9280.00388

Zoladz, P. R., Cadle, C. E., Dailey, A. M., Fiely, M. K., Peters, D. M., Nagle, H. E., Mosley, B. E., Scharf, A. R., Brown, C. M., Duffy, T. J., Earley, M. B., Rorabaugh, B. R., & Payment, K. E. (2017). Blunted cortisol response to acute pre-learning stress prevents misinformation effect in a forced confabulation paradigm. *Hormones and Behavior*, *93*, 1–8. https://doi.org/10.1016/j.yhbeh.2017.04.003

Appendix

Open call – Email

Subject: Call for unpublished data for a meta-analysis: "XXXX"
Dear Prof/Dr/Ms/Mr XXXX,

I am a XXXX at the University of XXXX, and I'm conducting a meta-analysis on XXXX, along with co-authors XXXX and XXXX.

The pre-registered protocol for this meta-analysis is publicly available on the Open Science Framework (OSF) at osf.io/XXXX; it is also attached as a pdf to this email.

As you have published studies relevant to this topic, we are getting in touch to see if you have any unpublished/file-drawer data, or papers in-press, which we may have missed through database searching, and which you would like to have included in the meta-analysis.

Feel free to email either the raw data (from which we will calculate summary scores) or the summary scores themselves. While any raw data emailed to us will of course remain confidential, please know that summary scores included in the meta-analysis will be made publicly available in a dataset on the OSF.

The latest we will be able to include your data in our meta-analysis is XXth of XXX, XXXX. We are hoping to include as many relevant studies as possible, so any additional data is greatly appreciated.

Thank you very much,
XXXX

Specific data request – Email

Subject: Requesting data for a meta-analysis, from your paper: "XXXX"

Dear Prof/Dr/Ms/Mr XXXX,

I am a XXXX at the University of XXXX, and I'm conducting a meta-analysis on XXXX, along with co-authors XXXX and XXXX.

The pre-registered protocol for this meta-analysis is publicly available on the Open Science Framework (OSF) at osf.io/XXXX; it is also attached as a pdf to this email.

We think your study "XXXX" meets inclusion criteria for our meta-analysis. However, the effect size we're interested in (i.e., the correlation/difference between XXX and XXX) does not seem to be reported in the published paper.

We would be grateful if you could send either the summary scores or the raw data themselves (from which we can calculate the effect size). While any raw data emailed to us will of course remain confidential, please know that summary scores included in the meta-analysis will be made publicly available in a dataset on the OSF.

The latest we will be able to accept your data for inclusion is XXth of XXX, XXXX.

We are hoping to include as many relevant studies as possible, so any additional data is greatly appreciated.

Thank you very much,

XXXX