AMERICAN PSYCHOLOGICAL ASSOCIATION

Journal of Experimental Psychology: General

ISSN: 0096-3445

2024, Vol. 153, No. 3, 621–655
https://doi.org/10.1037/xge0001527

The Effect of Negative Valence on False Memory Formation in the Deese–Roediger–McDermott Paradigm: A Preregistered Meta-Analysis and Preregistered Replication

Sera Wiechert[1, 2], Dora Proost[3], Emmelie Simoens[3], Gershon Ben-Shakhar[2], Yoni Pertzov[2], and Bruno Verschuere[1]

[1] Department of Clinical Psychology, University of Amsterdam
[2] Department of Psychology, The Hebrew University of Jerusalem
[3] Department of Experimental Psychology, Ghent University

Participants in the Deese–Roediger–McDermott (DRM) paradigm learn lists of words (e.g., bed, tired) associated with a nonpresented lure (i.e., sleep). In subsequent memory tests, individuals tend to report the nonlearned lures, that is, exhibiting false memories. Priorly, the DRM task has been criticized for not capturing the aversive nature of (clinically and forensically relevant) real-life memories. To obtain a robust estimate of the influence of negative versus neutral word lists on the DRM effect, we conducted both a preregistered meta-analysis ($k_{recall} = 49$, $n_{recall} = 2{,}209$, $k_{recognition} = 75$, $n_{recognition} = 3{,}008$, $k_{responsebias} = 31$, $n_{responsebias} = 1{,}128$) and replication ($n_{final} = 278$) predicting increased false memories for negative valence in recall and recognition. For recall, we found significant frequentist evidence in the meta-analysis for a reversed valence effect ($d = -0.18$, i.e., reduced false memories for negative content vs. neutral), whereas the replication displayed null results ($d = 0.03$). For recognition, both the meta-analysis ($d = 0.23$) and replication ($d = 0.35$) showed that negative valence (vs. neutral) increased false memories. However, this effect may be confounded by shifts in response tendencies as controlling for response bias nullified the valence effect in our meta-analysis ($d_{meta} = 0.05$), and we found evidence for differential response bias in our replication ($d_{replica} = 0.39$). Hence, the effect of valence on false memory reports in the DRM may not represent a systematic difference in emotional information but instead depend on how memory is tested, and be partly attributable to differential response tendencies.

Public Significance Statement

Underlined by both meta-analytic and replication study results, this article strongly indicates that negatively valanced word lists (vs. neutral) within the Deese–Roediger–McDermott paradigm increase false memory report in recognition, but not in recall tests. This differentiation may be partly attributable to a response criterion shift for negative valence (vs. neutral) in recognition-like investigation, rather than sole memory differences.

Keywords: Deese–Roediger–McDermott paradigm, false memories, emotion, meta-analysis, replication

This article was published Online First December 7, 2023.

Sera Wiechert https://orcid.org/0000-0002-0260-1702

This research was partly financially supported by the Amsterdam University Fund, making it possible for Sera Wiechert to conduct a research visit at the Hebrew University of Jerusalem and present the project as a poster at the conference on Cognition Research of the Israeli Society for Cognitive Psychology at Akko in February 2023. The authors report no conflicts of interest. The information and data presented in the article have not been published elsewhere. All meta-analytic data, analysis code, and research materials (including our coding scheme) are publicly available on the Open Science Framework at https://osf.io/z6ser/. The meta-analytic project (methodology and analysis) was preregistered at https://osf.io/ft64y. All replication study data, analysis code, and research materials are publicly available on the Open Science Framework at https://osf.io/9hxrg/. The replication project was preregistered at https://osf.io/wqfkh.

Sera Wiechert served as lead for conceptualization, data curation, formal analysis, methodology, writing–original draft, and writing–review and editing. Dora Proost served in a supporting role for writing–original draft and writing–review and editing. Emmelie Simoens served in a supporting role for writing–original draft and writing–review and editing. Dora Proost, Emmelie Simoens, Gershon Ben-Shakhar, Yoni Pertzov, and Bruno Verschuere contributed equally to conceptualization and methodology. Dora Proost and Emmelie Simoens contributed equally to formal analysis and data curation. Gershon Ben-Shakhar, Yoni Pertzov, and Bruno Verschuere contributed equally to writing–original draft and writing–review and editing.

Correspondence concerning this article should be addressed to Sera Wiechert, Department of Clinical Psychology, University of Amsterdam, Nieuwe Achtergracht 129B, 1018 WT Amsterdam, The Netherlands. Email: s.wiechert@uva.nl

Memory can be fallible and constructive (Roediger & McDermott, 1995). Especially in clinical and legal contexts, false memories can have immense consequences. Specifically, when objective evidence is unavailable or goes unnoticed due to missing resources, individuals' memories play a crucial role in legal proceedings (Brainerd, 2013). In such cases, erroneous memories could cause investigators to pursue wrong clues and pose incorrect accusations, potentially leading to wrongful convictions (Howe & Knott, 2015). This further emphasizes the importance of studying factors underlying false memory formation.

False memories can be defined as either remembering events that never happened or remembering them differently from how they actually happened (Roediger & McDermott, 1995). The Deese–Roediger–McDermott (DRM) paradigm (Roediger & McDermott, 1995) is one of the most frequently used paradigms to study false memories. In the DRM task, participants are typically presented with word lists (e.g., "chair," "legs," "desk") that are semantically associated with nonpresented words, the critical lures ("table"). In subsequent recall or recognition tests, the report of the lure as a word presented in the learned list is treated as evidence for false memory (i.e., the so-called DRM effect).

The DRM paradigm is a fast and straightforward approach to investigate false memories and has been studied widely (see Gallo, 2010 for a review). The DRM effect has been replicated across different age groups (e.g., children in Howe et al., 2004, 2010, 2011; elderly in Gallo & Roediger, 2003; Lo et al., 2014), and clinical populations (e.g., Moritz et al., 2004). Even studies implementing false memory warnings (e.g., Calvillo & Parong, 2016; Neuschatz et al., 2003) have found significant DRM effects across conditions. However, the DRM paradigm has been criticized for lacking ecological validity (e.g., DePrince et al., 2004; Freyd & Gleaves, 1996; Pezdek & Lam, 2007), especially in representing memories of clinical or legal relevance. The authenticity of memories that matter in real life (e.g., legal or clinical contexts) is commonly associated with negative emotions, and the emotional nature of the memory may affect its retrieval accuracy.

To increase ecological validity and measure false memory production under more realistic circumstances, research has investigated the effect of emotion in the DRM paradigm in three ways. First, participants' moods have been experimentally manipulated before commencing the DRM task (e.g., Knott et al., 2014; Storbeck & Clore, 2005; van Damme, 2013; Wright et al., 2005). Second, emotional word lists have been used (e.g., Baugerud et al., 2016; Brainerd et al., 2008, 2010; El Sharkawy et al., 2008). Third, combining the first two lines, studies have manipulated mood congruency (i.e., the matching between the valence of the stimuli and the experimentally induced mood; e.g., Bland et al., 2016; Ruci et al., 2009; Zhang et al., 2017). Over time, it has been argued that mood induction fails to capture the central feature of real-life situations, in which the negative valence of the to-be-remembered event induces a negative mood rather than the opposite direction (Brainerd et al., 2008). Therefore, in the current study, we focus on emotional word lists. More specifically, while mimicking real-life contexts, we aim to examine the effect of negative valence on the DRM effect. We note that not only negative but also the effect of positive valence has been studied previously. While not the focus of our project, existing literature suggests that positive valence, compared to neutral, can heighten false memories (Dehon et al., 2010; Piguet et al., 2008; Zhang, 2017; Zhang et al., 2017, 2019; but see also Brainerd et al., 2008). However, the effect may be less pronounced than that of negative valence (Brainerd & Bookbinder, 2019; Brainerd et al., 2010; Chang et al., 2021; Ruci et al., 2009). In addition to negative valence, we also considered the potential effects of arousal, as prior research has indicated that higher arousal within emotional conditions can further heighten false memory production (e.g., Brainerd et al., 2010; Chang et al., 2021; for a review, see Kensinger, 2004); however, arousal effects are only secondary to our project.

So, how would negative valence influence false memory production? Several theoretical frameworks have been proposed to explain the formation of false memories in recall and recognition tests within the DRM paradigm. In the following, we provide a concise overview of prominent approaches and their corresponding predictions, with particular emphasis on the fuzzy-trace theory due to its prominence in prior literature (e.g., Bookbinder & Brainerd, 2016), in line with our preregistered rationale.

According to the fuzzy-trace theory (Bookbinder & Brainerd, 2016; Reyna & Brainerd, 1995), two opponent memory processes are at play: verbatim and gist. Verbatim traces store the exact details of an experience, whereas gist traces are involved in processing the underlying meaning of the event (Reyna & Brainerd, 1995). Retrieval of verbatim traces is the basis for what is usually called recollection (Abadie et al., 2021; Brainerd et al., 2009), hereby supporting true and suppressing false memory report. In contrast, gist retrieval is the basis for what is commonly called familiarity (Bookbinder & Brainerd, 2017; Gomes et al., 2013). Constructing gist traces in the DRM learning phase may trigger critical lures in recall and recognition as the representation summarizes common semantic features of the studied words rather than the specific item (Gallo, 2010). Negative content is considered to have increased semantic connectedness, which may strengthen gist encoding, ultimately increasing false memories (Bookbinder & Brainerd, 2016; Gallo et al., 2009; Howe et al., 2010).

In addition to the fuzzy-trace theory, single-process accounts based on spreading activation (Anderson, 1983) are considered when studying false memory. Spreading activation assumes that memory is comprised of representations (i.e., nodes) interconnected by associative links. Regarding the DRM task, the activation of list nodes may spread to other nodes close in the associative network, thus, activating the critical node as well (Roediger et al., 2001). According to associative-activation theory (Howe et al., 2009), individuals acquire an elaborated, interrelated, and dense network of nodes throughout their lives. Therefore, associative activation becomes more potent and automatic as new knowledge is acquired. In addition to the above, the activation-monitoring theory (Roediger & McDermott, 1995, 2000) also assumes active source monitoring processes at retrieval (Johnson et al., 1993), operating to examine whether the activated representation was internally generated or originated from an external source. Notably, both activation-based theories may differ merely in their emphasis. While the former theory focuses on spreading activation in explaining false memory occurrence (i.e., highlighting semantic associations and activation), the latter additionally emphasizes the failure to monitor the source of activation accurately (i.e., source monitoring and misattribution). The activation theories suggest that negative valence (vs. neutral) increases false memories for critical content due to its increased semantic

connectedness (Bookbinder & Brainerd, 2016) and by inducing a negative context exacerbating source monitoring errors (Zhang et al., 2019).

While the discussed theories suggest increased false memories for negative valence (vs. neutral), they may also indirectly influence true memory rates. According to spreading activation accounts, if negative valence (vs. neutral) enhances activation throughout the network, both true and false nodes may be highly activated. Consequently, true memories may also be more prominent in the context of negative valence (vs. neutral). In contrast, the fuzzy-trace theory suggests that elevated gist processing underlies false memory propensity for negative content (vs. neutral). In turn, heightened gist processing may potentially suppress or leave unaffected the verbatim processing, which promotes the formation of true memories.

Overall, most research supports the notion that negative valence (vs. neutral) increases false memory formation in the DRM paradigm. However, alternative accounts within the DRM (and in the broader false memory literature) argue that negative material may, in fact, be more attention-grabbing, enhance recollection elaboration and distinctiveness, and increase deeper processing; thus, supporting item-specific processing (i.e., verbatim traces in fuzzy-trace theory) and source monitoring (i.e., retrieval monitoring effectiveness in activation-monitoring theory; Doss et al. (2020); Howe et al., 2010; Palmer & Dodson, 2009). Consequently, negative lists would induce fewer false memories than neutral lists. Notably, this directionality of effect may be particularly pronounced when valence conditions are balanced in associative strength and gist (Howe et al., 2010). This assumption is based on the idea that negative material might generally be more densely integrated and richly interconnected. Hence, when controlling for associative strength and gist, the effect of negative valence (vs. neutral) on false memory formation may be reduced or even reversed.

When reviewing empirical evidence, many studies demonstrate a significant increase in false memory production for negative compared to neutral critical lures in both recall and recognition tests of the DRM (e.g., Brainerd et al., 2008; Buratto et al., 2013; El Sharkawy et al., 2008; Norris et al., 2019; Shah & Knott, 2018). However, even though qualitative literature reviews state that this is a consistent finding, specifically for recognition tests (e.g., Bookbinder & Brainerd, 2016), some studies report no significant difference in false memory production between negative and neutral critical lures in recognition tests (e.g., Bland et al., 2016; Yuvruk et al., 2019; Zhang et al., 2019; also see Choi et al., 2013 for slightly different operationalization). Furthermore, as Bookbinder and Brainerd (2016) indicated, results for recall tests seem to be more mixed. For example, Howe (2007), Howe et al. (2010), and Palmer and Dodson (2009) found decreased false memory recall for negative compared to neutral lures, while others found nonsignificant differences between neutral and negative recall (Bauer et al., 2009; Joormann et al., 2009). In addition, several studies suggest that false memories in the DRM may be confounded by a systematic shift in the decision criterion, a so-called response bias (e.g., Dehon et al., 2010; El Sharkawy et al., 2008; Howe et al., 2010). Specifically, liberal response bias can be defined as an overall tendency to make an "old" response to targets (studied items) as well as false alarms (nonstudied critical or unrelated items) regardless of the actual memory representation (Macmillan & Creelman, 2005). Response bias measures and the criterion-shift theory have been previously criticized for not accurately representing false memory report in the DRM paradigm (Wixted & Stretch, 2000). Still, there is research that adequately investigated several response bias measures, specifically examining criterion differences between valence conditions. These indicate that liberal response bias may be further heightened by the negative valence of material (e.g., Brainerd et al., 2008; Dehon et al., 2010; El Sharkawy et al., 2008; Howe et al., 2010).

It becomes apparent that the effect of negative valence on false memory production in the DRM paradigm is inconclusive and may depend partially on the type of memory test conducted (recall vs. recognition) or arise from potential shifts in response tendencies. To investigate these potential explanations thoroughly, we conducted a preregistered meta-analysis and a preregistered replication to examine the effect of negative versus neutral word lists on false memories as captured by recall and recognition tests in the DRM paradigm.

Bookbinder and Brainerd (2016) published a narrative review on the current topic. Narrative reviews reflect an evidence-based synthesis to summarize studies for interpretation or to find a connection within the literature. Still, we argue that a statistical summary of the literature is necessary. Meta-analyses provide higher quality of evidence by summarizing effect sizes across populations, designs, and operationalizations and thus can provide a valuable statistical quantification of the current evidence on a particular topic (Esterhuizen & Thabane, 2016), as well as the generalizability of the results (i.e., external validity; Garg et al., 2008), also controlling for sample size differences in the literature. Specifically, we included moderation and sensitivity analyses as well as assessing potential publication bias. Yet, we know that high heterogeneity, publication bias, and questionable research practices within the literature can bias or even invalidate a meta-analysis (Esterhuizen & Thabane, 2016). For these reasons, a meta-analysis can be a valuable step in evaluating the literature but cannot substitute a preregistered replication, which is less affected by publication bias and questionable research practices (Van Elk et al., 2015). Replication studies are designed to validate prior findings on a specific effect (Nosek & Errington, 2020), and their results shed light on the reliability of the investigated effect under a specific operationalization (Nosek & Errington, 2020). However, due to their relatively constrained designs, replications typically provide less information about the effect's generalizability in comparison to meta-analyses (Allen & Preiss, 1993).

Meta-analyses and replication studies offer complementary information on the validity and reliability of the findings. Consequently, we believe that adopting and comparing both approaches is of great value. Therefore, to shed light on the directionality of the effect of stimulus valence on false memory production in the DRM paradigm, we conducted a preregistered meta-analysis and a preregistered replication and compared their outcomes systematically. In line with our preregistered rationale based on the more dominant account of the fuzzy-trace theory, we predicted for both the meta-analysis and the replication study that negative stimuli increase false memory production compared to neutral stimuli. Importantly, this prediction was tested separately for the recall and recognition test outcomes. Additionally, we investigated the effect of potential response bias shifts for negative valence (vs. neutral) in both approaches. Lastly, as the effect of valence on true memory in the DRM has not been a focus

theoretically or empirically, we exploratorily examined the effect of negative valence (vs. neutral) on true memory formation in both recall and recognition tests but refrained from making explicit predictions about its directionality.

Study 1: Meta-Analysis

Transparency and Openness

We adhered to the meta-analysis reporting standards guidelines for meta-analytic reporting (Appelbaum et al., 2018). All meta-analytic data, analysis code, and research materials (including our coding scheme) are available at https://osf.io/z6ser/. Data were analyzed using RStudio (Version 1.2.5019; using the package metafor, Version 3.0-2; Viechtbauer, 2010) and JASP (Version 0.14.1). This project was preregistered following a literature search and a pretest of a mini meta-analysis including a small sample of studies to verify the planned procedures and analyses. The preregistration can be found at https://osf.io/ft64y. We followed the PRISMA-P checklist when preparing the protocol, and PRISMA reporting guidelines for the final report.

Paradigm

The DRM paradigm containing neutral and negative lists represents a quasi-experimental within-subject design (i.e., word lists are not randomly allocated to valence conditions but are inherently neutral or negative). In the DRM paradigm, subjects are commonly presented with lists of words assembled with close semantic associates of a nonpresented critical lure (Roediger & McDermott, 1995). For example, the DRM variant designed to study the effect of negative valence, consists of distinct negative and neutral lists, semantically associated with negative and neutral lures, respectively (e.g., "tears," "sad," and "tissue" related to "cry"; "apple," "vegetable," and "orange" related to "fruit"; Howe et al., 2010). Lists are presented visually (e.g., Brainerd et al., 2008) or auditorily (e.g., Wright et al., 2005) in the learning phase, and participants are instructed to learn the words as accurately as possible for a later memory test (Roediger & McDermott, 1995). Subsequently, participants' memory is tested in a free recall (i.e., recall out loud or write down all items they can remember within a specified timeframe) and/or a recognition test (i.e., make old/new judgments for studied items, nonstudied critical lure words, and nonstudied unrelated words). Typically, DRM studies reveal that participants recall and recognize significantly more nonstudied critical items than nonstudied unrelated items, indicating that the DRM paradigm induces false memories for related critical content (Roediger & McDermott, 1995).

Exclusion Criteria

We applied four criteria to evaluate study eligibility. First, we excluded all papers that did not have a quasi-experimental design (i.e., excluding theoretical, qualitative literature, or [systematic] literature reviews; tagged as "nonexperimental"). Second, we excluded papers that did not use the DRM paradigm to elicit semantic–associative memory illusions (tagged as "not DRM"). DRM variants were included if the main characteristics were met (i.e., learning lists of items semantically associated with critical lures and a recall and/or recognition test). Third, we excluded papers that did not report a within-subject valence manipulation: no neutral or negative valence condition (tagged as "wrong/no list"). Fourth, as a post hoc addition to the preregistered exclusion criteria, we decided against the inclusion of studies using mood manipulations if there was no mood control group, as differential mood may bias valence results (see mood congruency studies; e.g., Bland et al., 2016).

Literature Search and Screening

We started with a literature search on published reports. Next, we searched the Open Science Framework (OSF) server for preprints and ProQuest Dissertation Abstracts for unpublished PhD theses. Following this, we also launched a call for (unpublished) studies on Twitter and ResearchGate and asked corresponding authors of included publications for additional (unpublished) studies. We elaborate on the detailed procedure in the following.

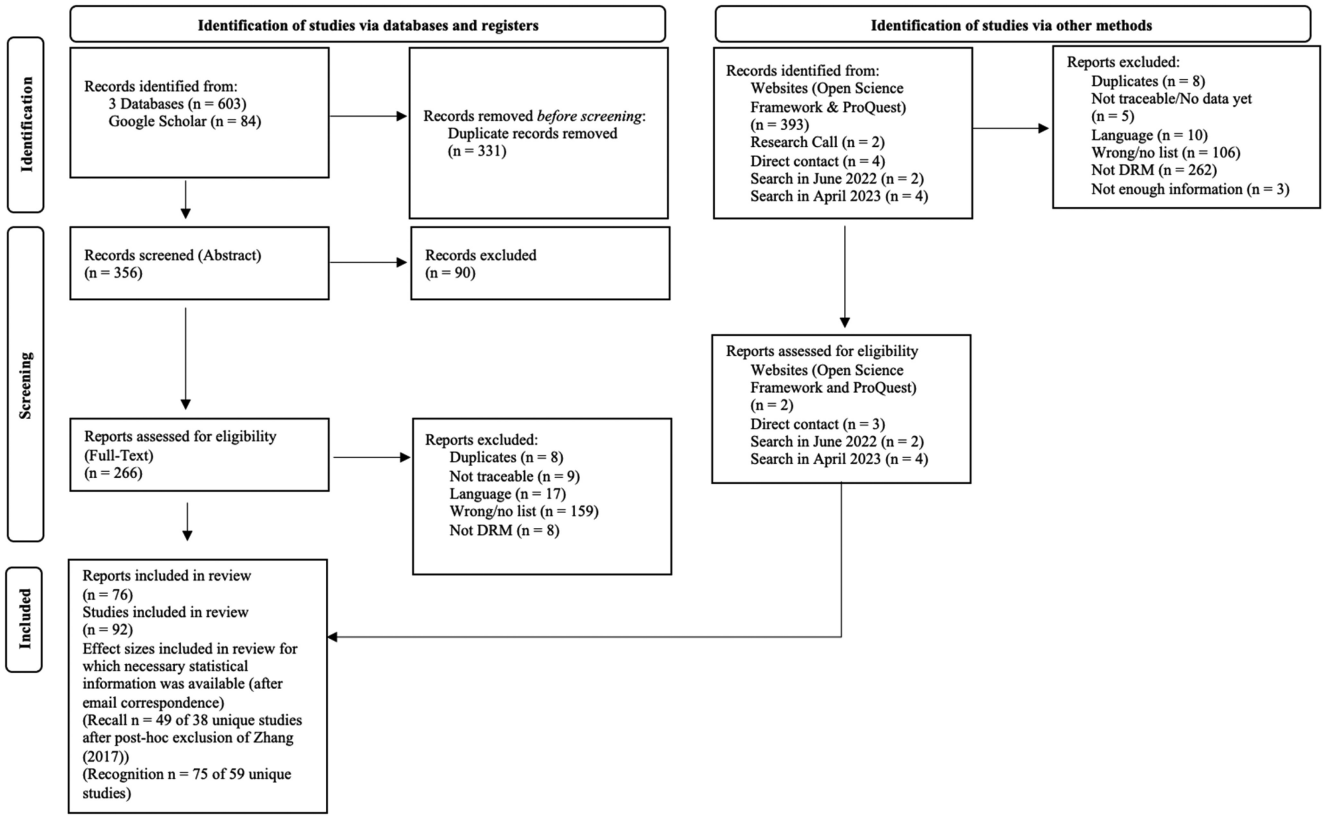
The literature search, which was limited to articles published since 1995 (i.e., year of the first DRM publication; Roediger & McDermott, 1995) was conducted in March 2021. 603 publications were retrieved from the databases Web of Science, Medline, and PsycInfo. In addition to the 603 articles, we searched Google Scholar for relevant articles ranked by Google's relevancy (i.e., March 22, 2021). This yielded 250 articles of which 84 were selected. The search terms can be found in Appendix A.

We used *Zotero* (Version 5.0.96.2) to deduplicate the total of retrieved published literature (n before deduplication = 687, n after deduplication = 356). Following this, we used the systematic review web app *Rayyan* to screen reports in English, Dutch, German, and Turkish (Ouzzani et al., 2016); the reviewer was not blind to authors, institutions, journals, and results. The screening process consisted of two phases, that is, title/abstract and full-text screening in which two researchers screened each report blinded to each other's decisions. Both phases were piloted on 10 randomly selected articles and screening only commenced if researchers agreed on all articles (initially or after discussion) of the pilot.

An overview of the screening procedure and exclusions is displayed in Figure 1. In the initial screening based on papers' titles and abstracts, all 356 articles were screened, and 90 were excluded (n = 266). Conflicts were resolved through discussion, and screeners showed interrater reliability (Cronbach's κ) of .95 (McHugh, 2012). Next, in the full-text screening of 266 articles, 65 eligible studies were identified (12 studies less than stated in the preregistration). The reasons for the post hoc exclusions in comparison to the preregistration were: duplicate = 1, not satisfying the criteria to classify as a DRM task = 2, and not including a pure neutral condition = 9. Here, screeners displayed a Cronbach's κ of .96 (based on the 77 inclusions stated in the preregistration, before the 12 post hoc exclusions).

When searching for reports on the OSF server and ProQuest Dissertation Abstracts (see Appendix A for search terms), we identified 393 potential studies. In the additional call for (unpublished) studies on Twitter and ResearchGate and upon contacting all authors of included publications to ask for additional (unpublished) studies, we identified six potential reports. After screening these 399 reports, five eligible reports were identified, leaving the meta-analysis with a sample of 70 reports. In a final inspection of the published literature in June 2022 and again in April 2023, we

Figure 1
PRISMA Flowchart



Note. As soon as the reviewer found one violation of an inclusion criterion, screening for other violations stopped. Difference to preregistration: 65 instead of 77 inclusions from databases due to additional exclusions while extracting: duplicate ($n = 1$), wrong/no list ($n = 9$), not DRM ($n = 2$). PRISMA = Preferred Reporting Items for Systematic Reviews and Meta-Analyses; DRM = Deese–Roediger–McDermott.

found six new reports, leaving the meta-analysis with a final sample of 76 eligible reports.

To obtain all necessary data, authors were contacted when studies reported insufficient statistical information. Studies could not be included if the authors did not reply after two emails or did not send the data within the communicated time frame. In total, 76 reports (92 studies) were extracted, but full statistical information for the computations was available for 50 effect sizes based on recall (of 39 studies), 75 effect sizes based on recognition (of 59 studies), and 31 effect sizes (of 25 studies) based on the recognition outcome, including the correction for potential response bias.

Coding System in Extraction

We used a standard coding system for each study's identification information (e.g., authors, title, journal), sample characteristics (e.g., sample size, mean and *SD* of age in the sample, percentage of female gender in the sample), and design characteristics (e.g., word list presentation order, number of lists, number of words in lists). Before the main data extraction phase, two researchers independently extracted the data from five articles to ensure a collaborative understanding of the data extraction procedure. The remaining articles were then extracted by one of the two researchers.

Additional data extraction and coding decisions were made when studies used multiple between-participant conditions per outcome for different ages (e.g., younger and older samples), gender groups (e.g., female and male samples), language groups (e.g., English and German samples), or health groups (e.g., sample with internalizing disorders and healthy control). In these cases, we extracted multiple effect sizes per outcome measure from each study. Otherwise, we extracted the between-subject condition most closely resembling the common operationalization used in the DRM literature. If a study used multiple within-participant conditions in one study, we extracted the effect per outcome measure that most closely resembled the common operationalization used in the DRM literature. Lastly, in case of multiple within-subject arousal conditions for negative valence in one study, we used the lower arousal condition for negative valence.

Meta-Analytic Procedures

Effect Size Calculation

Initially, we intended to control for potential response bias in our effect size calculations by including the recall and recognition rates of nonstudied unrelated words per valence. However, we soon realized that most studies did not report nonstudied unrelated recalled

words separately by valence conditions.[1] Therefore, we used two different effect size estimates (see Table 1 for an overview). The first did not include potential response bias (i.e., nonstudied unrelated words) and was calculated for both recall and recognition. The second effect size estimate was calculated only for recognition and incorporated potential response bias by including the rate of nonstudied unrelated recognized words. Comparing the two effect size estimates for the recognition test may indicate whether results are affected by potential response bias while also allowing direct comparisons with the recall outcome. For all estimates, standardized paired differences with a positive sign (+) indicated a larger number of false memories in the neutral compared to the negative valence condition, whereas standardized paired differences with a negative sign (−) indicated effects in the opposite direction. Following the preregistration, as the two recognition outcomes revealed different conclusions for the primary analysis, we conducted all additional analyses (moderator, publication bias, and sensitivity analyses) on all three effect size estimates.

Effect sizes were controlled for within-subject correlations either by using $SD_{pooled} = \sqrt{(SD_1^2 + SD_2^2 - 2 \times r \times SD_1 \times SD_2)}$, or by using the *SD* of the difference score if reported (see Table 1 for more information). As for some studies, the necessary correlation measures were unavailable, we calculated the average correlation computed across all studies that reported correlations and imputed it for all effect sizes for which the respective correlations (or *SD* of the difference score) could not be retrieved. Additionally, when necessary, *SD*s were calculated from *SE*s or confidence intervals (CI; $SD = SE \times \sqrt{n}$; $SD = \sqrt{n} \times$ (upper 95% CI threshold − lower 95% CI threshold)/3.92). If estimates were not reported (and not provided to us) but visualized in graphs, we imputed the numbers with *WebPlotDigitizer* (Rohatgi, 2021).

Study Quality and Open Science Characteristics

We evaluated stimulus and design quality of included studies in three ways (see Tables 2 and 3 for an overview). First, studies reporting using lists matched for mean backward associative strength were categorized as having lower risk of bias (0; 1 = low quality/not reported − higher risk of bias). Mean backward associative strength represents the average tendency, according to norms of association, for list words to evoke critical distractors (Brainerd et al., 2008). If lists differ in mean backward associative strength, resulting difference in responding may bias observed valence effects. Second, studies explicitly reporting a sample justification (e.g., power analysis, Bayesian stopping rule) were categorized as having lower risk of bias (0; 1 = low quality/not reported − higher risk of bias). Third, we checked whether the presentation order of the lists was counterbalanced or randomized (0 = counterbalanced or random − lower risk of bias; 1 = fixed/not reported − higher risk of bias). We also reflected upon the use of open science practices and results' accuracy in three ways. First, we examined whether it was explicitly reported that the study was preregistered (0; 1 = not preregistered/not reported). Second, we examined whether studies openly shared their data with publication (0 = immediately openly available vs. 1 = available under conditions vs. 2 = no mentioning). As open science practices (i.e., preregistration and data sharing) only became normative following the replication crisis in 2011, we did not assess study quality based on these characteristics but solely reflect on their use in our sample of studies. Finally, we used *Statcheck* (Rife et al., 2016) to code the number of inconsistencies within a research paper (i.e., continuous variable; blank row if program could not retrieve any statistics). *Statcheck* examines whether *p* values match their accompanying statistics. It searches for null hypothesis significance tests (i.e., recognizing *r* and *t*, *F*, χ^2, *Z*, and *Q* tests) and recalculates the *p* value using the reported test statistics and degrees of freedom. Nuijten et al. (2017) investigated its validity and identified an accuracy between 96.2% and 99.9% in discriminating between inconsistent and consistent results.

Moderators

Clinical Group. Literature suggests that individuals with internalizing psychological disorders (e.g., posttraumatic stress disorder in Brennen et al., 2007; major depressive disorder in Joormann et al., 2009) report more negative false memories than healthy participants. Therefore, we hypothesized that individuals with internalizing psychological disorders might generally show a stronger tendency of increased false memories for negative material. We included the clinical "internalizing disorders" groups with the following internalizing disorders: depressive disorder, anxiety disorder, obsessive-compulsive disorder, trauma-related disorder including posttraumatic stress disorder, dissociative disorder, bulimia, anorexia, and dysthymia (i.e., 0 = healthy vs. 1 = clinical "internalizing disorders" sample).

Mean Arousal of Negative Lures. Even though research has more strongly focused on the effect of valence than arousal, literature suggests that arousing negative material can further enhance the report of false memories (Brainerd et al., 2010; Chang et al., 2021). Arousal levels are commonly measured in stimulus validation studies in which participants rate words' arousal with self-assessment manikins (Bradley & Lang, 1994). We propose that differences in mean arousal levels of negative lists may influence false memory production and, consequently, differences in false memory production between neutral and negative valence. Therefore, we calculated the mean arousal of critical lures of negative lists for each study (i.e., summing up reported arousal levels of the negative critical lures divided by the total number of negative lures) and investigated this continuous variable as a moderator. For consistency and feasibility, we only extracted mean arousal ratings for negative critical stimuli and only used the exhaustive database by Warriner et al. (2013) to extract arousal levels. If the stimuli were presented in

[1] Valence ratings of items (e.g., affective norms for English words by Bradley & Lang, 1999) are continuous scores with no prespecified cutoffs per valence. Hence, when using experimental manipulations of valence, researchers decide a priori on the average valence ratings' threshold in order to categorize conditions as neutral versus negative valence. However, this a priori valence categorization is problematic when recall is used because new unrelated words that participants recall (i.e., intrusions) cannot be categorized (there is no valence cutoff score for individual words). Thus, it is difficult to put recalled intrusions (i.e., new unrelated words) into valence conditions post hoc. Therefore, most studies report intrusions in the recall test as a general intrusion measure, neither providing valence ratings of recalled intrusions nor categorizing intrusions into valence conditions. Intrusions would be more easily reported within valence conditions if the recall test were conducted after each list or block of lists (i.e., the intrusion is tied to a specific valence of a list or a block of lists), but this operationalization is used only rarely. For the recognition outcome, categorizing new unrelated items into valence conditions is more readily applicable but still not adapted systematically in the literature.

Table 1
Measures for the Effect Size Calculation per Outcome

Measure	Recall outcome	Recognition outcome	Recognition outcome controlling for response bias
Effect size calculation	Standardized paired difference: neutral nonstudied critical lures − negative nonstudied critical lures	Standardized paired difference: neutral nonstudied critical lures − negative nonstudied critical lures	Standardized paired difference: (neutral nonstudied critical lures − neutral nonstudied unrelated words) − (negative nonstudied critical lures − negative nonstudied unrelated words).
Within-subject correlation considered	Correlation between neutral nonstudied critical and negative nonstudied critical words	Correlation between neutral nonstudied critical and negative nonstudied critical words	1. Correlation between neutral nonstudied critical and neutral nonstudied unrelated words 2. Correlation between negative nonstudied critical and negative nonstudied unrelated words 3. Correlation between (neutral nonstudied critical lures − neutral nonstudied unrelated words) and (negative nonstudied critical lures − negative nonstudied unrelated words)
Within-subject correlation retrieved	$k = 52$, average $r = .31$	$k = 52$, average $r = .31$	1. $k = 22$, average $r = .21$ 2. $k = 20$, average $r = .16$ 3. $k = 20$, average $r = .23$

another language, we translated the critical lures to English and extracted their arousal ratings from the stated database.

Delay. Longer delays between presentation and retrieval have been associated with increased false memory rates overall (e.g., McDermott, 1996; Seamon et al., 2002). However, limited prior research suggests that increased retention intervals may also influence the relationship between valence and false memory formation in the DRM paradigm. We investigated the effect of delay in the current meta-analysis exploratorily and found no significant effects in any outcome (see https://osf.io/z6ser/ for more information).

Statistical Analyses

All analyses were conducted with a frequentist approach. In addition, we conducted a Bayesian analysis on the main effect of valence on all three outcomes. We implemented a Bayesian approach to all analyses excluding moderator and publication bias. Analyses were conducted separately for all three effect measures to guard against within-subject dependency.

For null hypothesis significance testing, we used the standard $p < .05$ criteria (null hypothesis significance test) to determine whether the analysis reached statistical significance. For Bayesian statistics, we calculated Bayes factors (BF_{01}), which indicates how likely the data are under the null hypothesis compared to the alternative hypothesis (Jarosz & Wiley, 2014). The inverse ratio (BF_{10}) allows one to speak of the likelihood of the alternative hypothesis compared to the null (Jarosz & Wiley, 2014). We classified $BF_{10} > 5$ as convincing evidence in favor of the alternative hypothesis over the null hypothesis and $BF_{01} > 5$ as convincing evidence in favor of the null hypothesis over the alternative hypothesis. If Bayesian analyses showed $BF_{10} < 5$ and $BF_{01} < 5$, we did not draw definite conclusions on the directionality of the respective effect.

If effect sizes could not be computed because of missing means or *SD*s and could not be retrieved from the authors, the data analysis for the primary valence analyses of the respective outcome was conducted in two ways. First, we excluded these studies and conducted a complete case analysis only. Second, we recalculated the primary valence analysis with imputed averages of respective missing measures. When calculating the averages for the respective measures, we carefully investigated which reporting standards were used most often (e.g., proportions, or percentages) and used a uniform format to avoid biased averaging. The moderator, sensitivity, and publication bias analyses were only conducted on the complete cases.

For the meta-analysis, effect sizes were aggregated based on a random-effects model, using the restricted maximum likelihood (REML) method to estimate the heterogeneity in effect sizes (Viechtbauer, 2005). We report the estimated average effect, the I^2, which represents the percentage of the true variance (i.e., total observed variance − sampling error variance). As Higgins and Thompson (2002) suggested, we interpreted heterogeneity by the I^2 measure, as it is less dependent on the study sample size or scaling parameters. Moderator analyses were performed when $I^2 > 25\%$. For the Bayesian analysis, we conducted a Bayesian model-averaged meta-analysis and used JASP default settings—that is, an inverse-gamma (1, 0.15) heterogeneity prior, and a Cauchy (0, 0.707) effect size prior. For all analyses, corresponding 95% CIs were reported and forest plots were provided. Lastly, for descriptive purposes, we calculated the average neutral DRM effect for the recognition task. This effect size was computed as the standardized mean difference between reported neutral nonstudied critical and neutral nonstudied unrelated words.

Moderator Analyses

To address whether moderators can explain variations in effect sizes, selected variables were examined using mixed-effects models with REML estimation for the amount of (residual) heterogeneity. As the moderator clinical group (healthy vs. individuals with internalizing disorder) had only a few studies per level for each outcome measure in our analysis, only descriptive information is presented.

Publication Bias

Publication bias analyses were conducted separately on the random-effects analyses for all three outcomes. Publication bias, that is, the tendency for significant effects to be published more often than nonsignificant effects, may lead to overestimating effect

Table 2
Study Quality Items and Open Science Characteristics for the Recall Outcome

	Stimulus and design quality			Open science practices and accuracy of results		
Study	Backward associative strength	Order of lists	Sample justification	Preregistration	Data sharing	StatCheck
Baugerud et al. (2016)	Not controlled	Random/counterbalanced	No	No	No mention	4
Beckwé and Deroost (2016)	Not controlled	Random/counterbalanced	No	No	No mention	NA
Brennen et al. (2007)	Not controlled	Fixed/other	No	No	No mention	3
Calado et al. (2018)	Controlled	Random/counterbalanced	Yes	No	Openly available	0
Dehon et al. (2010)	Controlled	Random/counterbalanced	No	No	No mention	NA
Dewhurst et al. (2012)	Controlled	Random/counterbalanced	No	No	No mention	NA
Diliberto-Macaluso et al. (2016)	Controlled	Random/counterbalanced	No	No	No mention	0
El Sharkawy et al. (2008)	Not controlled	Random/counterbalanced	No	No	No mention	3
Harper (2017)	Not controlled	Random/counterbalanced	No	No	No mention	0
Houben et al. (2020) Exp. 1	Not controlled	Random/counterbalanced	Yes	Yes	Openly available	1
Houben et al. (2020) Exp. 2	Not controlled	Random/counterbalanced	Yes	Yes	Openly available	1
Howe et al. (2010) Exp. 1	Controlled	Random/counterbalanced	No	No	No mention	NA
Howe et al. (2010) Exp. 2	Controlled	Random/counterbalanced	No	No	No mention	NA
Howe et al. (2010) Exp. 3	Controlled	Random/counterbalanced	No	No	No mention	NA
Howe et al. (2010) Exp. 4	Controlled	Random/counterbalanced	No	No	No mention	NA
Howe et al. (2010) Exp. 5	Controlled	Random/counterbalanced	No	No	No mention	NA
Irwanda and Maulina (2018)	Not controlled	Random/counterbalanced	No	No	No mention	0
Joormann et al. (2009)	Not controlled	Random/counterbalanced	No	No	No mention	0
Maulina et al. (2021) Exp. 1	Not controlled	Random/counterbalanced	Yes	No	Openly available	0
Maulina et al. (2021) Exp. 2	Not controlled	Random/counterbalanced	Yes	No	Openly available	0
Maulina et al. (2021) Exp. 3	Not controlled	Random/counterbalanced	Yes	No	Openly available	0
Maulina et al. (2021) Exp. 4	Not controlled	Random/counterbalanced	Yes	No	Openly available	0
McKeon et al. (2012)	Controlled	Random/counterbalanced	No	No	No mention	1
Meeks et al. (2019)	Controlled	Fixed/other	No	No	No mention	1
Monds et al. (2017)	Not controlled	Random/counterbalanced	No	No	No mention	0
Monds et al. (2013)	Not controlled	Random/counterbalanced	No	No	No mention	NA
Nie et al. (2022)	Not controlled	Random/counterbalanced	Yes	No	Available upon request	2
Otgaar et al. (2020)	Not controlled	Random/counterbalanced	Yes	No	Openly available	0
Otgaar, Peters, and Howe (2012)	Controlled	Random/counterbalanced	No	No	No mention	1
Palmer and Dodson (2009) Exp.1	Controlled	Random/counterbalanced	No	No	No mention	NA
Palmer and Dodson (2009) Exp. 2	Controlled	Random/counterbalanced	No	No	No mention	NA
Ruci et al. (2009)	Not controlled	Random/counterbalanced	No	No	No mention	NA
Smeets et al. (2008)	Controlled	Random/counterbalanced	No	No	No mention	NA
Thijssen et al. (2013)	Not controlled	Random/counterbalanced	No	No	No mention	NA
Vannucci et al. (2012)	Not controlled	Fixed/other	No	No	No mention	NA
Velsor (2016)	Not controlled	Fixed/other	No	No	No mention	0
Welliver (2015)	Not controlled	Fixed/other	No	No	No mention	1
Zhang et al. (2018)	Controlled	Random/counterbalanced	Yes	Yes	No mention	1
Total of desired outcome	16	33	10	3	8	12

Note. Exp. = experiment; NA = not available.

sizes in meta-analyses (Thornton & Lee, 2000). As we had more than 10 eligible effect sizes for all outcomes, we constructed a funnel plot (i.e., the effect size of each study on the *x*-axis and its precision on the *y*-axis), whereby asymmetry, especially at the bottom of the plot, may indicate possible publication bias. Additionally, we conducted a rank correlation test (Begg & Mazumdar, 1994) and Egger's regression test (Egger et al., 1997). As we did not find evidence for potential publication bias for any outcome (i.e., the rank correlation and the regression test did not display significant results), we did not apply selection models (Vevea & Woods, 2005) or the procedure outlined by Carter et al. (2019; see preregistration for more information). Lastly, we conducted moderator analyses to further explore hints of possible publication bias. Extrinsic characteristics of studies (e.g., publication year and status) should theoretically be unrelated to study results, but this may not always be the case. For example, publication year may be negatively related to effect size, also called the "decline effect" (i.e., effect sizes tend to decline over time; Schooler, 2011). We, therefore, explored publication year (continuous variable) and publication type (0 = not published vs. 1 = published) as potential moderators. We used mixed-effects models with REML estimation for the amount of (residual) heterogeneity. As publication status had only a few studies per level for each outcome measure in our analysis, only descriptive information is presented.

Sensitivity Analyses

To test the robustness of our results, sensitivity analyses were conducted on the random-effects analyses (i.e., model excluding moderators) for all three effect size estimates separately. First, we examined whether the conclusions changed when varying the imputed average correlation estimates as explained in Effect Size Calculation. For all effect size estimates, we reinserted a correlation one *SD* above and one *SD* below the average imputed correlation for the respective correlation measure. For the sensitivity analysis on both the recall and recognition outcome (without intrusions in the computation), two correlation sensitivity analyses were conducted. For the analysis of

Table 3
Study Quality Items and Open Science Characteristics for the Recognition Outcome

Study	Stimulus and design quality			Open science practices and accuracy of results		
	Backward associative strength	Order of lists	Sample justification	Preregistration	Data sharing	StatCheck
Beckwé and Deroost (2016)	No	Random/counterbalanced	No	No	No mention	NA
Bland et al. (2016)	Yes	Random/counterbalanced	No	No	No mention	4
Brainerd et al. (2008)	Yes	Fixed/other	No	No	No mention	NA
Brueckner and Moritz (2009)	Yes	Random/counterbalanced	No	No	No mention	0
Budson et al. (2006)	No	Random/counterbalanced	No	No	No mention	0
Calado et al. (2018)	Yes	Random/counterbalanced	Yes	No	Openly available	0
Ching et al. (2015)	Yes	Random/counterbalanced	No	No	No mention	NA
Dehon et al. (2010)	Yes	Random/counterbalanced	No	No	No mention	NA
Deng and Lu (2022)	Yes	Fixed/other	No	No	No mention	6
Dewhurst et al. (2018)	Yes	Random/counterbalanced	No	No	No mention	NA
Duesenberg et al. (2016)	No	Fixed/other	No	No	No mention	NA
El Sharkawy et al. (2008)	No	Random/counterbalanced	No	No	No mention	3
Griffin and Schnyer (2020)	No	Random/counterbalanced	Yes	No	Openly available	2
Harper (2017)	No	Random/counterbalanced	No	No	No mention	0
Hauschildt et al. (2012)	NA	Random/counterbalanced	No	No	No mention	NA
Hellenthal et al. (2019) Exp. 1	Yes	Random/counterbalanced	Yes	No	Available upon request	2
Hellenthal et al. (2019) Exp. 2	Yes	Random/counterbalanced	Yes	No	Available upon request	2
Houben et al. (2020) Exp. 1	No	Random/counterbalanced	Yes	Yes	Openly available	1
Houben et al. (2020) Exp. 2	No	Random/counterbalanced	Yes	Yes	Openly available	1
Howe et al. (2010) Exp. 1	Yes	Random/counterbalanced	No	No	No mention	NA
Howe et al. (2010) Exp. 2	Yes	Random/counterbalanced	No	No	No mention	NA
Howe and Malone (2011)	Yes	Random/counterbalanced	No	No	No mention	NA
Irwanda and Maulina (2018)	No	Random/counterbalanced	No	No	No mention	0
Jelinek et al. (2009)	NA	Fixed/other	No	No	No mention	NA
Knott et al. (2018)	Yes	Random/counterbalanced	Yes	No	No mention	NA
Knott and Shah (2019)	Yes	Random/counterbalanced	Yes	No	No mention	1
Knott and Thorley (2014)	No	Random/counterbalanced	No	No	No mention	0
Maulina et al. (2021) Exp. 1	No	Random/counterbalanced	Yes	No	Openly available	0
Maulina et al. (2021) Exp. 2	No	Random/counterbalanced	Yes	No	Openly available	0
Maulina et al. (2021) Exp. 3	No	Random/counterbalanced	Yes	No	Openly available	0
Maulina et al. (2021) Exp. 4	No	Random/counterbalanced	Yes	No	Openly available	0
Meeks et al. (2019)	Yes	Fixed/other	No	No	No mention	1
Meusel et al. (2012)	No	Random/counterbalanced	No	No	No mention	0
Miano et al. (2022)	Yes	Random/counterbalanced	Yes	No	No mention	4
Monds et al. (2017)	No	Random/counterbalanced	No	No	No mention	0
Monds et al. (2013)	No	Random/counterbalanced	No	No	No mention	NA
Newbury (2019) Exp. 3	Yes	Fixed/other	Yes	No	No mention	28
Newbury (2019) Exp. 5	Yes	Random/counterbalanced	No	No	No mention	28
Otgaar, Alberts, and Cuppens (2012)	Yes	Random/counterbalanced	No	No	No mention	1
Otgaar et al. (2016) Exp. 1	Yes	Random/counterbalanced	Yes	No	No mention	NA
Otgaar et al. (2016) Exp. 2	Yes	Random/counterbalanced	Yes	No	No mention	NA
Otgaar et al. (2020)	No	Random/counterbalanced	Yes	No	Openly available	0
Otgaar, Howe, and Muris (2017)	Yes	Random/counterbalanced	No	No	No mention	0
Otgaar et al. (2013)	Yes	Random/counterbalanced	No	No	No mention	1
Otgaar, Moldoveanu, et al. (2017) Exp. 1	No	Random/counterbalanced	No	No	No mention	0
Otgaar, Moldoveanu, et al. (2017) Exp. 2	No	Random/counterbalanced	No	No	No mention	0
Quas et al. (2016)	Yes	Random/counterbalanced	No	No	No mention	2
Riesthuis et al. (2022)	Yes	Fixed/other	Yes	Yes	Openly available	0
Rodríguez-Ferreiro et al. (2019)	Yes	Random/counterbalanced	No	No	Available upon request	1
Ruci et al. (2009)	No	Random/counterbalanced	No	No	No mention	NA
Shah and Knott (2018)	Yes	Random/counterbalanced	Yes	No	No mention	0
Stea et al. (2013)	No	Random/counterbalanced	No	No	No mention	0
Velsor (2016)	No	Fixed/other	No	No	No mention	0
Yablonski (2016)	No	Fixed/other	No	No	No mention	0
Yüvrük and Kapucu (2022)	Yes	Random/counterbalanced	Yes	No	Openly available	0
Zhang et al. (2019)	Yes	Random/counterbalanced	Yes	No	Openly available	0
Zhang et al. (2021)	Yes	Random/counterbalanced	Yes	No	Openly available	NA
Zhang et al. (2017) Exp. 2	Yes	Random/counterbalanced	No	No	No mention	0
Zhong et al. (2018)	No	Random/counterbalanced	No	No	No mention	2
Total of desired outcome	32 (17)	50 (21)	22 (8)	3 (0)	13 (3)	24 (8)

Note. Studies in gray are the studies included for the recognition effect size estimate controlling for potential response bias. Exp. = experiment; NA = not available.

the recognition outcome controlling for potential response bias, a total of 26 additional analyses per dependent variable were conducted (three levels for three estimated correlations: $3 \times 3 \times 3 = 27$ excluding the one analysis in which all three correlations estimates are their averages, i.e., the original random-effects analysis).

Second, we checked our study sample for potential outliers (i.e., the study's 95% CI does not overlap with the 95% CI of the pooled effect). We excluded all outliers in a sensitivity analysis on the random-effects model and presented the average effect size without outliers.

Third, in line with Stanley et al. (2010), we also examined whether the analysis of the top 10% most precise studies revealed different results. This methodology is premised on the idea that the most precise studies (i.e., the precision index is the inverse of the estimated *SE*) are most informative. This technique reduced bias greatly in simulation studies and was often more efficient than conventional summary statistics (Stanley et al., 2010).

Lastly, preregistration can remedy publication bias and questionable research practices. In preregistered studies, all operationalizations are a priori defined which reduces the use of questionable research practices. In addition, manuscripts are likely accepted for publication even when effects of interest are not statistically significant. Therefore, we conducted a random-effects analysis on the subset of preregistered studies to examine whether conclusions differed. The top 10% and the preregistered studies sensitivity analyses were conducted if we had a minimum of three eligible effect sizes per outcome. Results were interpreted with caution when less than five eligible effect sizes per outcome were available.

Exploratory Analysis

True Memory. Next to making predictions about false memory formation, distinct theoretical frameworks also imply differential effects of valence (neutral vs. negative) on true memories. Therefore, we separately examined whether hit rates differed between neutral and negative valence for recall and recognition outcomes. The method employed for calculating effect sizes was akin to the primary computations used for determining false memory effects, with the distinction that rates of true memory were substituted instead of rates of critical false memory. Since this analysis was exploratory, we controlled for the within-subject correlation in the effect size calculation by utilizing the retrieved correlation from the replication study as a proxy (and including sensitivity analyses).

Controlling for Associative Strength. Prior literature suggests that the increased semantic density of negative material could underlie its increased false memory reporting rather than pure valence effects (Howe et al., 2010). To explore this idea further, we reran the primary analysis for each outcome separately on the subgroup of studies that explicitly mentioned having controlled for backward associative strength. Hereby, we investigated whether the effects of negative valence (vs. neutral) on false memory formation may differ when the associative strength of negative valence is controlled.

Results

Deviations From the Preregistration[2]

As stated above (see Exclusion Criteria section), we excluded studies that used mood manipulations without including a mood control group. Furthermore, we excluded the recall data reported by Zhang (2017, Experiment 3) because the retrieved effect size was an extreme outlier (i.e., the study's 95% CI did not overlap with the 95% CI of any other study). Thus, analyses of the recall outcome were based on a final 49 effect sizes (of 38 unique reports). The primary analysis including this extreme outlier can be found in Appendix B. Lastly, as open science practices only became normative in recent years, we decided not to interpret study quality in terms of their preregistration or data sharing status but solely reflect on these study characteristics.

Original Neutral DRM Effect

We computed the average standardized paired difference for the comparison between neutral nonstudied critical/semantically related words and nonstudied unrelated words in the recognition test ($k = 32$, $n = 1{,}158$). Both the frequentist and the model-averaged Bayesian analysis showed conclusive evidence for the original DRM effect ($d = 1.32$, 95% CI [0.93, 1.71], $p < .001$, $BF_{10} = 95{,}292.19$). Thus, across studies, individuals reported more false memories for neutral nonstudied semantically related items than neutral nonstudied unrelated items.

Study Quality and Open Science Characteristics

Overview of the study quality and open science characteristics can be found in Table 2 for the recall and Table 3 for the recognition outcome. Most studies reported having employed random or counterbalanced presentation orders in the learning phase to reduce bias (33 out of 38 studies for recall; 50 out of 59 studies for recognition). However, only about half of the studies reported to have mean backward associative strength (i.e., the average tendency for list words to evoke critical distractors) controlled for their DRM lists (16 out of 38 for recall; 32 out of 59 for recognition), and even fewer studies reported justifications for their chosen sample size (10 out of 38 studies for recall; 22 out of 59 studies for recognition). For the recall and recognition outcomes altogether, six included studies were preregistered and 21 studies shared their data openly. Lastly, *StatCheck* reported no inconsistencies in 12 cases for the recall outcome (for 15 out of the 38 papers, an error appeared) and in 24 cases for the recognition outcome (for 17 out of the 59 papers, an error appeared).

Recall Test

Overall Effect Size. Standardized paired differences were computed for 49 effect sizes ($n = 2{,}209$). The effect sizes of the single

[2] Some additional extraction choices will be outlined in the following. For the study by Diliberto-Macaluso et al. (2016), two between-subject conditions were extracted as both conditions closely resembled the DRM paradigm and only differed in the learning phase modality (i.e., visual vs. auditory). For the study by Griffin and Schnyer (2020), we could not reliably match statistics between raw data provided by the authors and those reported in the paper, which is why only the primary statistics (i.e., neutral and negative critical lures) were extracted from the published article. For the study by Howe et al. (2010, Experiments 2 and 4), statistics were only reported for two groups together. As these groups only differed in mean age but not age group (i.e., children), combined data was extracted (i.e., raw data was unavailable). For the study by Otgaar et al. (2020), the learning phase implemented in the task was slightly adapted; we chose the source monitoring procedure as it most closely resembled the original DRM procedure. For the study by Quas et al. (2016), the sample size of the extracted conditions was not reported. Upon contacting the authors, we were told that the specific number of participants was unavailable but that they had a minimum of 40 individuals per group. Therefore, $n = 40$ was inserted here.

studies and the average effect size across all studies are provided in Figure 2 and Table 4. Notably, in 17 cases, study estimates showed significant effects in the positive direction, whereas only six study estimates showed significant effects in the negative direction. The frequentist two-tailed analysis revealed a significant effect size, $d = 0.18$, 95% CI [0.05, 0.32], $p = .008$. In contrast to the hypothesized directionality of the effect, this result indicated that negative valence, compared to neutral valence, decreased false memories for critical items. However, the effect was small and the lower bound of the CI was only slightly above zero. Furthermore, the Bayesian model-averaged analysis showed inconclusive evidence ($BF_{10} = 2.34$; i.e., did not support any hypothesis over the other). In order to increase the study sample size, we reran the primary analysis with imputed means and *SD*s of missing values based on their respective averages (based on proportion measure; $k = 55$). This changed the results only slightly and the BF conclusively favored the alternative over the null hypothesis, $d = 0.20$, 95% CI [0.08, 0.33], $p = .001$, $BF_{10} = 9.89$.

The I^2 measure indicated that 88.88% (95% CI [84.52, 93.59]) of the observed variance between effect sizes was caused by systematic differences between studies, justifying our preregistered moderator analyses.

Moderator Analysis. The average effect size was 0.21 for healthy ($k = 47$) and −0.26 for clinical groups ($k = 2$) with internalizing disorders. Moreover, we found a significant effect in the metaregression analysis for the mean arousal of negative critical items ($Z = 3.18$, $p = .001$). Contrary to the expected relationship, this result revealed that higher mean arousal of negative critical items was associated with higher (more positive) effect size estimates, indicating more false memories in the neutral than the negative condition. However, in light of the inconclusive evidence in the primary analysis of valence, this finding should be treated with caution.

Publication Bias. Overall, there was no evidence of potential publication bias. The funnel plot (see Figure 3) was quite symmetric and both the rank correlation (Kendall's $\tau = <.01$, $p = .993$), and Egger's regression test for funnel plot asymmetry ($Z = 1.52$, $p = .128$) were nonsignificant.

When further investigating potential publication bias, we could not analyze publication status as a moderator because only five of the 49 computed effect sizes represented nonpublished work ($d_{unpublished} = -0.20$, $d_{published} = 0.24$). Additionally, in the metaregression analysis including publication year, we did not find significant moderation effects ($Z = -1.49$, $p = .137$). This analysis did not include publication dates for theses or dissertations that were not published in peer-reviewed journals.

Sensitivity Analyses. First, we reexamined the primary analysis on the effect of valence with imputed correlation estimates. In both cases (imputed correlation measures on *SD* below the average, $d = 0.17$, 95% CI [0.04, 0.29], $p = .008$, $BF_{10} = 2.13$; above the average, $d = 0.20$, 95% CI [0.05, 0.36], $p = .008$; $BF_{10} = 2.42$), the results did not change. Second, we reanalyzed the primary analysis with outliers removed (15 effect sizes removed). Excluding these studies resulted in a significant and conclusive effect, $d = 0.17$, 95% CI [0.10, 0.24], $p < .001$, $BF_{10} = 218.02$, providing strong evidence for the conclusion that neutral, compared to negative valence, increased false memories for critical items. Third, as the top 10% analysis consisted only of four effect sizes, the results should be interpreted with caution. In line with the primary analysis results, we found a significant but inconclusive average effect size for the recall test, in the opposite direction than initially expected ($d = 0.25$, 95% CI [0.09, 0.40], $p = .002$, $BF_{10} = 3.56$). Finally, we reanalyzed the recall test results with a subsample of preregistered studies ($k = 5$). Here, we found an overall negative average effect size, $d = -0.17$, 95% CI [−0.53, 0.20], $p = .377$, $BF_{01} = 3.53$, but the results were nonsignificant and showed inconclusive evidence.

Exploratory Analysis.

True Memory. Standardized paired differences were computed for 41 effect sizes in recall ($n = 1{,}753$). Negative, compared to neutral valence, showed decreased true memories for studied items, $d = 0.47$, 95% CI [0.30, 0.63], $p < .001$, $BF_{10} = 4{,}106.70$. In the two sensitivity analyses with differing imputed within-subject correlation estimates (correlation estimates for the neutral and low-arousal negative comparison, and the neutral and high-arousal negative comparison), the conclusions remained the same ($d = 0.41$, 95% CI [0.27, 0.56], $p < .001$, $BF_{10} = 4{,}377.28$; $d = 0.44$, 95% CI [0.29, 0.60], $p < .001$, $BF_{10} = 4{,}221.74$, respectively).

Controlling for Associative Strength. When only considering studies that explicitly mentioned having matched stimuli on mean backward associative strength, 23 effect sizes were included ($n = 900$). In line with the primary analysis, the frequentist analysis revealed a significant effect size, $d = 0.26$, 95% CI [0.03, 0.49], $p = .030$. Hence, similarly to the primary confirmatory outcome, negative valence (vs. neutral) decreased false memories for critical items, although the Bayesian evidence remained inconclusive ($BF_{10} = 1.21$).

Recognition Test

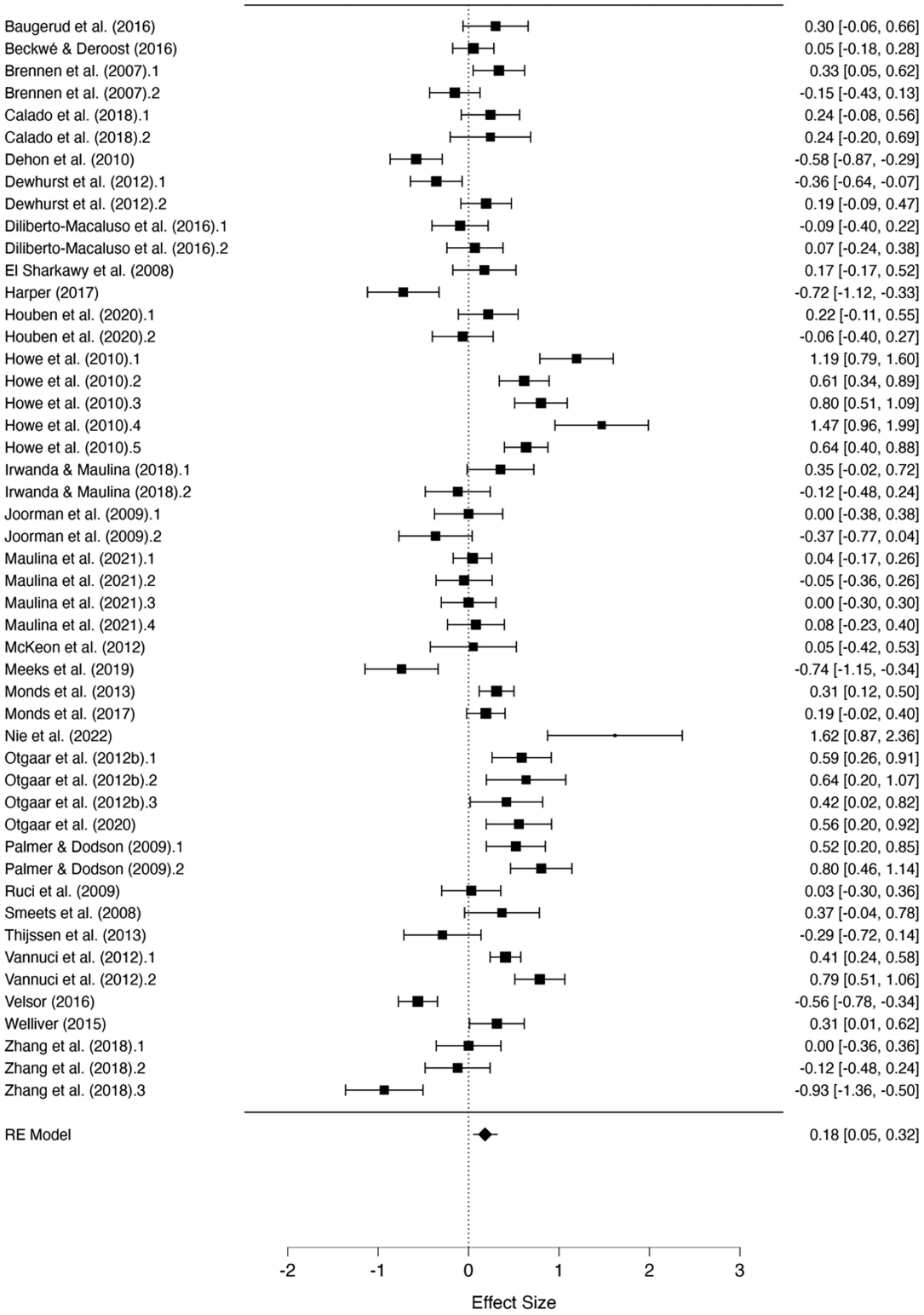
Overall Effect Size. We computed average standardized paired differences for 75 estimates ($n = 3{,}008$). The effect sizes of the single studies and the average effect size across all studies are provided in Figure 4 and Table 5. Notably, in 30 cases study estimates showed significant effects in the negative direction, whereas only two study estimates showed significant effects in the positive direction. The frequentist analysis revealed a significant negative effect size, $d = -0.23$, 95% CI [−0.32, −0.14], $p < .001$. Furthermore, the Bayesian analysis showed strong evidence in favor of the alternative over the null hypothesis ($BF_{10} = 5{,}514.42$). Thus, as hypothesized, we found that negative content induced more false memories for nonstudied critical items compared to neutral material. When imputing means and *SD*s based on their respective averages (based on proportion measure; $k = 89$), we still found significant and conclusive evidence in the same directionality, $d = -0.21$, 95% CI [−0.28, −0.14], $p < .001$, $BF_{10} = 26{,}146.79$. Lastly, I^2 indicated that 80.99% (95% CI [74.96, 87.68] of the observed variance between effect sizes was caused by systematic differences between studies, justifying the following metaregression analyses.

Moderator Analysis. The average effect size was −0.25 for healthy groups ($k = 70$) and 0.08 for clinical groups with internalizing disorders ($k = 4$). For the metaregression analysis including mean arousal level of negative critical items, we found nonsignificant effects ($Z = 0.68$, $p = .495$). Thus, higher mean arousal of negative critical items did not significantly increase false memories for negative critical compared to neutral critical items.

Publication Bias. Similar to the recall outcome, we did not find statistical evidence for potential publication bias (see Figure 5 for a funnel plot). Both the rank correlation (Kendall's $\tau = -.10$, $p = .204$) and the Egger's regression test for funnel plot asymmetry indicated nonsignificant deviations from symmetry ($Z = -1.89$, $p = .059$).

Figure 2

Forest Plot for the Recall Test Outcome



Note. RE = random-effects.

Table 4
Study Characteristics and Effect Sizes With Sample Sizes and Confidence Intervals per Included Effect for the Recall Outcome

Study	Participant age group	Proportion of females in sample (%)	Participant health	Mean arousal of negative critical lures	Arousal matched	*n*	*d*	95% CI *LL*	95% CI *UL*
Baugerud et al. (2016)	<18	38	Healthy	5.72	No	31	0.30	−0.06	0.66
Beckwé and Deroost (2016)	>18	84	Healthy			74	0.05	−0.18	0.28
Brennen et al. (2007) (1)	>18	50	Healthy	5.04		50	0.33	0.05	0.62
Brennen et al. (2007) (2)	>18	50	Clinical	5.04		50	−0.15	−0.43	0.13
Calado et al. (2018) (1)	<18 (adolescents)	34	Healthy			38	0.24	−0.08	0.56
Calado et al. (2018) (2)	<18 (children)	60	Healthy			20	0.24	−0.20	0.69
Dehon et al. (2010)	>18	56	Healthy	4.90	Yes	54	−0.58	−0.87	−0.29
Dewhurst et al. (2012) (1)	>18	100	Healthy	5.50	No	50	−0.36	−0.64	−0.07
Dewhurst et al. (2012) (2)	>18	0	Healthy	5.50	No	50	0.19	−0.09	0.47
Diliberto-Macaluso et al. (2016) (1)	>18	54	Healthy		Yes	40	−0.09	−0.40	0.22
Diliberto-Macaluso et al. (2016) (2)	>18	54	Healthy		Yes	40	0.07	−0.24	0.38
El Sharkawy et al. (2008)	>18	50	Healthy			32	0.17	−0.17	0.52
Harper (2017)	>18	74	Healthy			31	−0.72	−1.12	−0.33
Houben et al. (2020) Exp. 1	>18	92	Healthy	5.39		36	0.22	−0.11	0.55
Houben et al. (2020) Exp. 2	>18	87	Healthy	5.39		34	−0.06	−0.40	0.27
Howe et al. (2010) Exp. 1	>18	50	Healthy	6.15		40	1.19	0.79	1.60
Howe et al. (2010) Exp. 2	<18	50	Healthy	5.56		60	0.61	0.34	0.89
Howe et al. (2010) Exp. 3	>18	50	Healthy	5.42		60	0.80	0.51	1.09
Howe et al. (2010) Exp. 4	<18	47	Healthy	5.56		30	1.47	0.95	1.98
Howe et al. (2010) Exp. 5	<18	48	Healthy	5.71		80	0.64	0.39	0.88
Irwanda and Maulina (2018) (1)	>18	0	Healthy	5.52	No	30	0.35	−0.02	0.72
Irwanda and Maulina (2018) (2)	>18	100	Healthy	5.52	No	30	−0.12	−0.48	0.24
Joormann et al. (2009) (1)	>18	70	Healthy		Yes	27	0	−0.38	0.38
Joormann et al. (2009) (2)	>18	56	Clinical		Yes	25	−0.36	−0.77	0.04
Maulina et al. (2021) Exp. 1		52	Healthy	5.52	No	84	0.04	−0.17	0.26
Maulina et al. (2021) Exp. 2	>18	50	Healthy	5.52	No	40	−0.05	−0.36	0.26
Maulina et al. (2021) Exp. 3		52	Healthy	5.52	No	42	0	−0.30	0.30
Maulina et al. (2021) Exp. 4		77	Healthy	5.52	No	39	0.08	−0.23	0.40
McKeon et al. (2012)	>18	94	Healthy			17	0.05	−0.42	0.53
Meeks et al. (2019)	>18	78	Healthy	5.60	No	30	−0.74	−1.15	−0.34
Monds et al. (2017)	>18	75	Healthy	5.59	Yes	87	0.19	−0.02	0.40
Monds et al. (2013)	>18	69	Healthy	5.59	Yes	109	0.31	0.12	0.50
Nie et al. (2022)	>18	48	Healthy	5.80	No	16	1.62	0.87	2.36
Otgaar et al. (2020)	>18	69	Healthy	5.25		34	0.56	0.20	0.92
Otgaar, Peters, and Howe (2012) (1)	<18 (younger)	55	Healthy	5.71	Yes	42	0.59	0.26	0.91
Otgaar, Peters, and Howe (2012) (2)	<18 (older)	55	Healthy	5.71	Yes	24	0.64	0.20	1.07
Otgaar, Peters, and Howe (2012) (3)	>18	55	Healthy	5.71	Yes	26	0.42	0.02	0.82
Palmer and Dodson (2009) Exp. 1	>18	68	Healthy	5.52		41	0.52	0.20	0.85
Palmer and Dodson (2009) Exp. 2	>18	43	Healthy	5.52		44	0.80	0.46	1.14
Ruci et al. (2009)	<18	65	Healthy	4.86		36	0.03	−0.30	0.36
Smeets et al. (2008)	>18	93	Healthy	5.39		24	0.37	−0.04	0.78
Thijssen et al. (2013)	<18	59	Healthy	5.71	No	22	−0.29	−0.72	0.14
Vannucci et al. (2012) (1)	<18	100	Healthy	5.30		145	0.41	0.24	0.58
Vannucci et al. (2012) (2)	<18	0	Healthy	5.30		66	0.79	0.51	1.06
Velsor (2016)	>18	75	Healthy			95	−0.56	−0.78	−0.34
Welliver (2015)	>18	77	Healthy	4.94	Yes	44	0.31	0.01	0.61
Zhang et al. (2018) (1)	<18 (younger)	50	Healthy	4.93	Yes	30	0	−0.36	0.36
Zhang et al. (2018) (2)	<18 (older)	50	Healthy	4.93	Yes	30	−0.12	−0.48	0.24
Zhang et al. (2018) (3)	>18	50	Healthy	4.93	Yes	30	−0.93	−1.36	−0.50

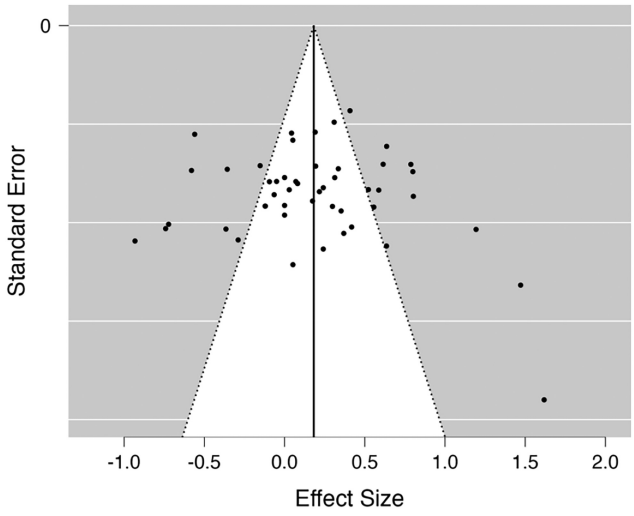
Note. Exp. = experiment; CI = confidence interval; *LL* = lower limit; *UL* = upper limit.

We further investigated the potential effect of publication bias by examining the effects of publication year and status. We could not conduct a metaregression analysis on publication status as only five included effect sizes were retrieved from nonpublished literature (published $d = -0.24$, nonpublished $d = -0.21$). Furthermore, the metaregression analysis including publication year showed no predictive effect ($Z = -0.18$, $p = .854$). Again, dates for unpublished theses and dissertations were not included.

Sensitivity Analyses. First, we reanalyzed the primary effect of valence with imputed correlation estimates. In both cases (imputed correlation measures on *SD* below the average, $d = -0.21$, 95% CI [−0.29, −0.13], $p < .001$, $BF_{10} = 4{,}949.60$; above the average, $d = -0.26$, 95% CI [−0.35, −0.16], $p < .001$, $BF_{10} = 3{,}804.35$), the results remained the same. Second, we removed outliers (15 effect sizes in total removed). Excluding these studies still led to a significant negative effect, $d = -0.20$, 95% CI [−0.27, −0.14], $p < .001$, $BF_{10} = 94{,}067.67$, providing strong evidence for the conclusion

Figure 3

Funnel Plot for the Recall Test Outcome



Note. The effect size of each study is plotted on the *x*-axis against its precision (i.e., *SE*) on the *y*-axis. Asymmetry on either side may indicate a bias toward imprecise but large effect size in one directionality, and hereby, potential publication bias.

that negative valence was associated with higher false memory production in comparison to neutral valence. Third, the top 10% analysis was conducted on seven effect size estimates. In this analysis, we also found a negative average effect size of $d = -0.16$ (95% CI [−0.42, 0.10], $p = .233$, $BF_{01} = 3.22$), but this effect was nonsignificant and inconclusive. Finally, the analysis on the preregistered subsample showed a similar negative average effect size, but the evidence was nonsignificant and inconclusive ($k = 3$, $d = -0.22$, 95% CI [−0.53, 0.08], $p = .155$, $BF_{01} = 1.69$).

Exploratory Analysis.

True Memory. On 69 computed effect sizes ($n = 2{,}755$), the frequentist two-tailed analysis revealed a nonsignificant effect size, $d = -0.05$, 95% CI [−0.15, 0.04], $p = .290$. Hence, there was no statistical difference in true recognition memory between valence conditions. This was further underlined by the conclusive Bayesian evidence for the null over the alternative hypothesis ($BF_{01} = 10.20$). Equally, in the two sensitivity analyses with differing imputed within-subject correlation estimates, the evidence continued to favor the null over the alternative hypothesis ($d = -0.05$, 95% CI [−0.13, 0.04], $p = .306$, $BF_{01} = 12.05$; $d = -0.05$, 95% CI [−0.14, 0.04], $p = .296$, $BF_{01} = 10.20$, respectively).

Controlling for Associative Strength. Forty-three effect sizes ($n = 1{,}680$) were included for the recognition outcome when only considering studies explicitly mentioning having matched stimuli on mean backward associative strength. In line with the primary result, we found significant and conclusive evidence in favor of the alternative over the null hypothesis that negative valence (vs. neutral) showed increased false memories in the recognition outcome, $d = -0.31$, 95% CI [−0.43, −0.20], $p < .001$, $BF_{10} = 4{,}353.61$.

Recognition Test; Controlling for Potential Response Bias

Overall Effect Size. We computed the average standardized paired difference for 31 effect sizes ($n = 1{,}128$). The effect sizes of the single studies and the average effect size across all studies are provided in Figure 6 and Table 5. As stated above, for this estimate, we calculated the difference between nonstudied critical and nonstudied unrelated items within valence conditions, contrasted between valence conditions. As many studies failed to report nonstudied unrelated responses per valence condition, the recognition outcome controlling for potential response bias resulted in fewer effect sizes compared to the recognition outcome not including nonstudied unrelated item responses. The frequentist analysis revealed a nonsignificant average effect size close to zero, $d = -0.05$, 95% CI [−0.19, 0.08], $p = .431$. Furthermore, the Bayesian model-averaged analysis showed evidence in favor of the null over the alternative hypothesis ($BF_{01} = 9.52$), indicating that, when controlling for potential response bias, false memory production did not statistically differ per valence. We also imputed means and *SD*s of missing values (based on proportion measures, $k = 89$) to reinvestigate the primary analysis when statistical power is increased. This changed the results such that it revealed a significant and conclusive negative effect size ($d = -0.11$, 95% CI [−0.17, −0.05], $p < .001$, $BF_{10} = 13.14$), indicating that negative content (vs. neutral) induces more false memories for critical content, even when controlling for a potential response tendency toward "old" responses. Lastly, the I^2 measure indicated that 78.97% (95% CI [67.52, 89.70] of the observed variance between effect sizes was caused by systematic differences between studies, justifying the need to proceed to the preregistered metaregression analyses.

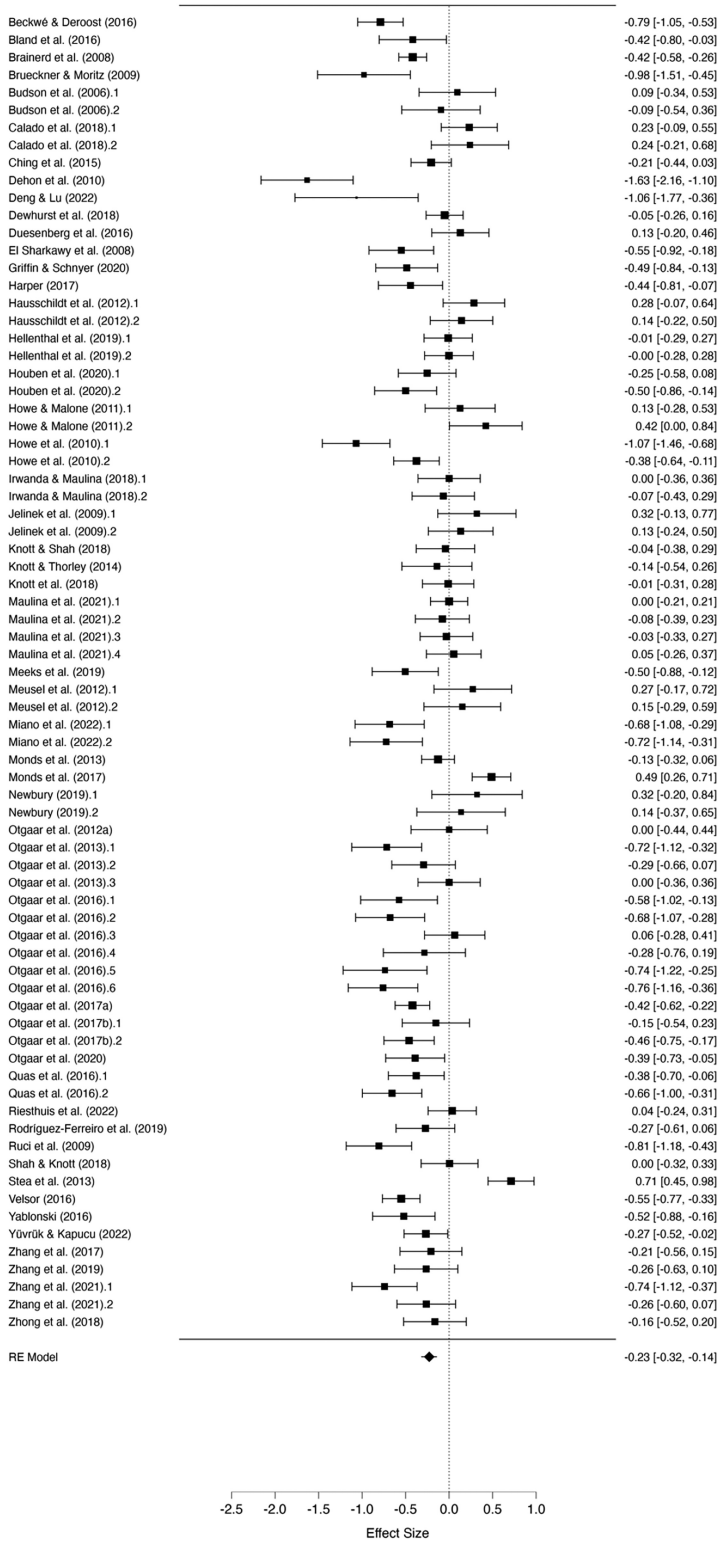
Moderator Analysis. For the metaregression analysis including the effect of mean arousal levels of negative critical items, we found a significant result ($Z = 2.32$, $p = .020$), indicating that more arousing negative material was associated with smaller differences in false memories between negative and neutral items. Nevertheless, in light of the null effect in the primary analysis of valence, this finding should be treated with caution. Furthermore, the average effect size was −0.10 for healthy ($k = 28$), and 0.33 for clinical groups with internalizing disorders ($k = 3$).

Publication Bias. No evidence for publication bias was found for the recognition outcome, also when controlling for potential response bias (see Figure 7 for the funnel plot). Both the rank correlation (Kendall's $\tau = .08$, $p = .544$) and Egger's regression test for funnel plot asymmetry indicated nonsignificant deviations from symmetry ($Z = -0.47$, $p = .635$). Following this, we examined additional moderators to test for publication bias. We could not conduct a metaregression analysis including publication status as only two included effect sizes were retrieved from nonpublished literature ($d_{unpublished} = -0.03$, $d_{published} = -0.06$). For the metaregression analysis (i.e., not including unpublished work), we found no significant predictive effect of publication year ($Z = -0.65$, $p = .516$).

Sensitivity Analyses. First, 26 sensitivity analyses with imputed correlation estimates were conducted for the primary analysis on the effect of valence (see Appendix C for an overview). The conclusions of these analyses were unaffected by the variations in the imputed correlation estimates. Overall, Cohen's *d* values ranged from −0.05 to −0.06, *p* values from .368 to .520, and BF_{01} values from 9.01 to 10.00. Second, we reanalyzed the primary effect of valence when six outliers were removed. Excluding these studies persisted in showing a nonsignificant effect and conclusive evidence in favor of the null hypothesis over the alternative hypothesis, $d = -0.06$, 95% CI [−0.12, 0.01], $p = .085$, $BF_{01} = 6.49$. Third, the top 10% analysis on three effect size estimates still showed a significant effect, but the Bayesian analysis failed to provide conclusive

Figure 4

Forest Plot for the Recognition Test Outcome



Note. RE = random-effects.

Table 5

Study Characteristics and Effect Sizes With Sample Sizes and Confidence Intervals per Included Effect for the Recognition Outcome

Study	Participant age group	Proportion of females in sample (%)	Participant health	Mean arousal of negative critical lures	Arousal matched	n	d_1	95% CI		d_2	95% CI	
								LL	*UL*		*LL*	*UL*
Beckwé and Deroost (2016)	>18	84	Healthy			74	−0.79	−1.05	−0.53			
Bland et al. (2016)	>18	70	Healthy	5.99	No	28	−0.42	−0.80	−0.03	0.17	−0.20	0.54
Brainerd et al. (2008)	>18	75	Healthy		Yes	163	−0.42	−0.58	−0.26	−0.11	−0.26	0.05
Brueckner and Moritz (2009)	>18	60	Healthy	4.09		20	−0.98	−1.51	−0.44			
Budson et al. (2006) (1)	>18 (younger)		Healthy	5.59		20	0.09	−0.34	0.53	0.55	0.08	1.02
Budson et al. (2006) (2)	>18 (older)	63	Healthy	5.59		19	−0.09	−0.54	0.36	0.12	−0.33	0.57
Calado et al. (2018) (1)	<18 (adolescents)	34	Healthy			38	0.23	−0.09	0.55			
Calado et al. (2018) (2)	<18 (children)	60	Healthy			20	0.24	−0.21	0.68			
Ching et al. (2015)	>18	66	Healthy	4.63	Yes	73	−0.21	−0.44	0.03	−0.20	−0.43	0.03
Dehon et al. (2010)	>18	44	Healthy	4.90	Yes	32	−1.63	−2.16	−1.10	−1.58	−2.10	−1.06
Deng and Lu (2022)	>18	58	Healthy	6.02	No	12	−1.06	−1.77	−0.35	−0.12	−0.69	0.44
Dewhurst et al. (2018)	>18	79	Healthy	5.50	No	85	−0.05	−0.26	0.16			
Duesenberg et al. (2016)	>18	50	Healthy	4.09	No	36	0.13	−0.20	0.46			
El Sharkawy et al. (2008)	>18	50	Healthy			32	−0.55	−0.92	−0.18	−0.35	−0.71	0.01
Griffin and Schnyer (2020)	>18	57		5.40	No	34	−0.49	−0.84	−0.13			
Harper (2017)	>18	74	Healthy			31	−0.44	−0.81	−0.07			
Hauschildt et al. (2012) (1)	>18	72	Clinical			32	0.28	−0.07	0.64	0.22	−0.13	0.58
Hauschildt et al. (2012) (2)	>18	70	Healthy			30	0.14	−0.22	0.50	0.13	−0.23	0.49
Hellenthal et al. (2019) Exp. 1	>18	64	Healthy	6.06	Yes	50	−0.01	−0.29	0.27			
Hellenthal et al. (2019) Exp. 2	>18	73	Healthy	6.06	Yes	49	<−0.01	−0.28	0.28	−0.01	−0.29	0.27
Houben et al. (2020) Exp. 1	>18	92	Healthy	5.39		36	−0.25	−0.58	0.08			
Houben et al. (2020) Exp. 2	>18	87	Healthy	5.39		34	−0.50	−0.86	−0.14			
Howe et al. (2010) Exp. 1	>18	50	Healthy	6.15		40	−1.07	−1.45	−0.68			
Howe et al. (2010) Exp. 2	<18	50	Healthy	5.56		60	−0.37	−0.64	−0.11			
Howe and Malone (2011) (1)	>18	58	Healthy	5.80	No	24	0.13	−0.28	0.53	0.11	−0.29	0.51
Howe and Malone (2011) (2)	>18	60	Clinical	5.80	No	24	0.42	0.00	0.84	0.72	0.27	1.17
Irwanda and Maulina (2018) (1)	>18	0	Healthy	5.52	No	30	0	−0.36	0.36			
Irwanda and Maulina (2018) (2)	>18	100	Healthy	5.52	No	30	−0.07	−0.43	0.29			
Jelinek et al. (2009) (1)	>18	70	Clinical		No	20	0.32	−0.13	0.77	0.05	−0.39	0.49
Jelinek et al. (2009) (2)	>18	61	Healthy		No	28	0.13	−0.24	0.50	0.03	−0.34	0.40
Knott et al. (2018)	>18	77	Healthy	5.20	No	44	−0.01	−0.31	0.28	−0.01	−0.31	0.28
Knott and Shah (2019)	>18	69	Healthy	5.14	No	34	−0.04	−0.38	0.29	−0.05	−0.38	0.29
Knott and Thorley (2014)	>18	75	Healthy	5.43	No	24	−0.14	−0.54	0.26	0.07	−0.33	0.48
Maulina et al. (2021) Exp. 1		52	Healthy	5.52	No	84	0	−0.21	0.21			
Maulina et al. (2021) Exp. 2	>18	50	Healthy	5.52	No	40	−0.08	−0.39	0.23			
Maulina et al. (2021) Exp. 3		52	Healthy	5.52	No	42	−0.03	−0.33	0.27			
Maulina et al. (2021) Exp. 4		77	Healthy	5.52	No	39	0.05	−0.26	0.37			
Meeks et al. (2019)	>18	78	Healthy	5.60	No	30	−0.50	−0.88	−0.12			
Meusel et al. (2012) (1)	>18 (younger)	65	Healthy		No	20	0.27	−0.17	0.72	0.11	−0.33	0.55
Meusel et al. (2012) (2)	>18 (older)	60	Healthy		No	20	0.15	−0.29	0.59	−0.03	−0.47	0.41
Miano et al. (2022) (1)	>18	83	Healthy	4.09		30	−0.68	−1.08	−0.29			
Miano et al. (2022) (2)	>18	75	Clinical	4.09		28	−0.72	−1.14	−0.31			
Monds et al. (2017)	>18	75	Healthy	5.59	Yes	87	0.49	0.26	0.71			
Monds et al. (2013)	>18	69	Healthy	5.59	Yes	109	−0.13	−0.32	0.06			
Newbury (2019) Exp. 3	>18	78	Healthy	4.23	Yes	15	0.32	−0.20	0.84	−0.10	−0.60	0.41
Newbury (2019) Exp. 5	>18	87	Healthy	4.23	Yes	15	0.14	−0.37	0.65	0.03	−0.47	0.54
Otgaar, Alberts, and Cuppens (2012)	>18	70	Healthy	5.39		20	0	−0.44	0.44			
Otgaar et al. (2016) Exp. 1 (1)	<18 (younger)	48	Healthy	5.71		23	−0.57	−1.02	−0.13			
Otgaar et al. (2016) Exp. 1 (2)	<18 (older)	43	Healthy	5.71		30	−0.68	−1.07	−0.28			
Otgaar et al. (2016) Exp. 1 (3)	>18	72	Healthy	5.71		32	0.06	−0.28	0.41			
Otgaar et al. (2016) Exp. 2 (1)	<18 (younger)		Healthy	5.71		18	−0.28	−0.75	0.19			
Otgaar et al. (2016) Exp. 2 (2)	<18		Healthy	5.71		21	−0.74	−1.22	−0.25			
Otgaar et al. (2016) Exp. 2 (3)	<18 (older)		Healthy	5.71		31	−0.76	−1.16	−0.36			
Otgaar et al. (2020)	>18	69	Healthy	5.25		36	−0.39	−0.73	−0.05			
Otgaar, Howe, and Muris (2017)	<18	55	Healthy	5.71		106	−0.42	−0.62	−0.22			
Otgaar et al. (2013) (1)	<18 (younger)	60	Healthy	5.71		30	−0.72	−1.12	−0.32			
Otgaar et al. (2013) (2)	<18 (older)	60	Healthy	5.71		30	−0.29	−0.66	0.07			
Otgaar et al. (2013) (3)	>18	66	Healthy	5.71		30	0	−0.36	0.36			
Otgaar, Moldoveanu, et al. (2017) Exp. 1	>18	70	Healthy			26	−0.15	−0.54	0.23			
Otgaar, Moldoveanu, et al. (2017) Exp. 2	>18	63	Healthy			51	−0.46	−0.75	−0.17			
Quas et al. (2016) (1)	<18 (younger)	52	Healthy		No	40	−0.38	−0.70	−0.06			
Quas et al. (2016) (2)	<18 (older)	52	Healthy		No	40	−0.65	−1.00	−0.31			
Riesthuis et al. (2022)	>18	90	Healthy	5.71	No	50	0.04	−0.24	0.31			
Rodríguez-Ferreiro et al. (2019)	>18	49	Healthy	4.76	Yes	35	−0.27	−0.61	0.06	−0.29	−0.63	0.05
Ruci et al. (2009)	<18	65	Healthy	4.86		36	−0.81	−1.18	−0.43	−0.78	−1.15	−0.41
Shah and Knott (2018)	>18	58	Healthy	5.03	No	36	<0.01	−0.32	0.33	0.03	−0.30	0.35

(*table continues*)

Table 5 (*continued*)

Study	Participant age group	Proportion of females in sample (%)	Participant health	Mean arousal of negative critical lures	Arousal matched	n	d_1	95% CI		d_2	95% CI	
								LL	*UL*		*LL*	*UL*
Stea et al. (2013)	>18	78	Healthy			69	0.71	0.45	0.98	0.69	0.42	0.95
Velsor (2016)	>18	75	Healthy			95	−0.55	−0.77	−0.33			
Yablonski (2016)	>18	56	Healthy	5.42		34	−0.52	−0.88	−0.16			
Yüvrük and Kapucu (2022)	>18	87	Healthy	5.48	No	63	−0.27	−0.52	−0.02	−0.04	−0.28	0.21
Zhang et al. (2019)	>18	50	Healthy	4.73	Yes	30	−0.26	−0.63	0.10	−0.17	−0.53	0.19
Zhang et al. (2021) (1)	>18 (younger)	76	Healthy	4.93	Yes	35	−0.74	−1.12	−0.37	−0.61	−0.97	−0.24
Zhang et al. (2021) (2)	>18 (older)	63	Healthy	4.93	Yes	35	−0.26	−0.60	0.07	−0.19	−0.52	0.15
Zhang et al. (2017) Exp. 2	>18	71	Healthy	4.93	Yes	31	−0.21	−0.56	0.15	−0.25	−0.61	0.11
Zhong et al. (2018)	>18	57	Healthy			30	−0.16	−0.52	0.20			

Note. Studies in gray are the studies included for the recognition effect size estimate controlling for potential response bias. Exp. = experiment; CI = confidence interval; *LL* = lower limit; *UL* = upper limit.

evidence, $d = -0.11$, 95% CI $[-0.23, < -0.01]$, $p = .048$, $BF_{01} = 2.67$. Lastly, the sensitivity analysis on the preregistered subsample of studies was not conducted as none of the retrieved estimates were preregistered.

Exploratory Analyses.

Controlling for Associative Strength. In line with the primary analysis, the frequentist analysis revealed a nonsignificant average effect size, $k = 19$, $n = 778$, $d = -0.13$, 95% CI $[-0.29, 0.03]$, $p = .120$. Furthermore, although inconclusive, the Bayesian model-averaged analysis showed evidence more in favor of the null over the alternative hypothesis ($BF_{01} = 2.92$), indicating that, when controlling for potential response bias but also associative strength, false memory production did not statistically differ per valence.

Discussion

In the current meta-analysis, we investigated the generalizability and reliability of the hypothesis that negative, compared to neutral valence, increases false memory formation in the DRM. For this, we examined three effect size measures—recall, recognition, and recognition controlled for response bias. Critically, we found strong evidence ($BF_{10} = 95{,}292.19$) for the original DRM effect ($d = 1.32$), based on neutral words.

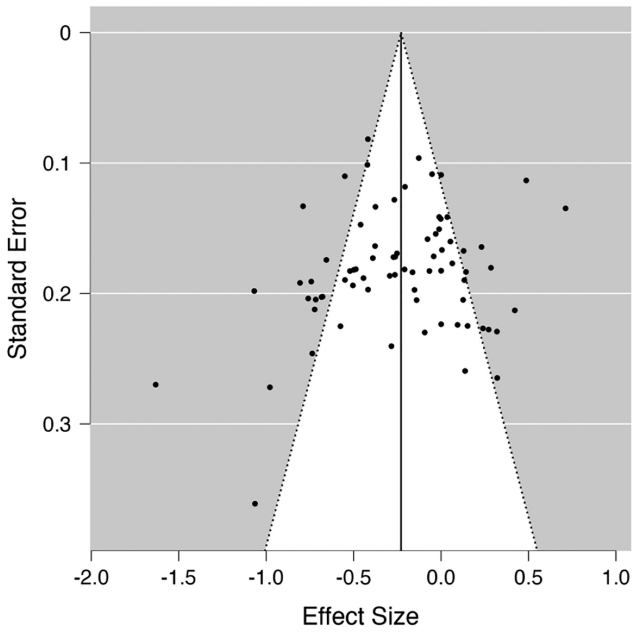
Importantly, there was no evidence of publication bias in our meta-analysis.[3] This could be attributed to the strong and robust DRM effect in its original procedure. Consequently, studies finding the general DRM effect might be published irrespective of the directionality of their valence results.

Overall, for the recall outcome, the evidence did not support our hypothesis regarding negative valence. On the contrary, the frequentist analysis showed evidence in the opposite direction: negative valence was associated with fewer false memories than neutral valence. When testing the robustness of this finding, there were more indications that neutral rather than negative valence may be associated with increased false memories in the DRM task. Although this does not align with our primary hypothesis, alternative accounts can explain the observed recall findings. As stated previously, it is argued that negative information (vs. neutral) may, in fact, be more distinct, resulting in increased item-specific processing and retrieval monitoring in the DRM (Howe et al., 2010; Palmer & Dodson, 2009). However, this finding should still be treated cautiously due to the inconclusive Bayesian result.

In contrast, we found robust results for the recognition test aligned with our valence hypothesis, indicating that negative valence (vs. neutral) was associated with higher false memory formation. Further sensitivity analysis results still strongly favored the expected relationship. However, conclusions differed when controlling for potential response bias in recognition. Although the average effect remained negative, we found conclusive evidence for the null hypothesis, indicating no statistical differences in false memory formation between negative and neutral valence.

Several conclusions can be drawn from the current results. First, while we found a reasonably robust valence effect for the recognition outcome, we did not find the hypothesized effect when controlling

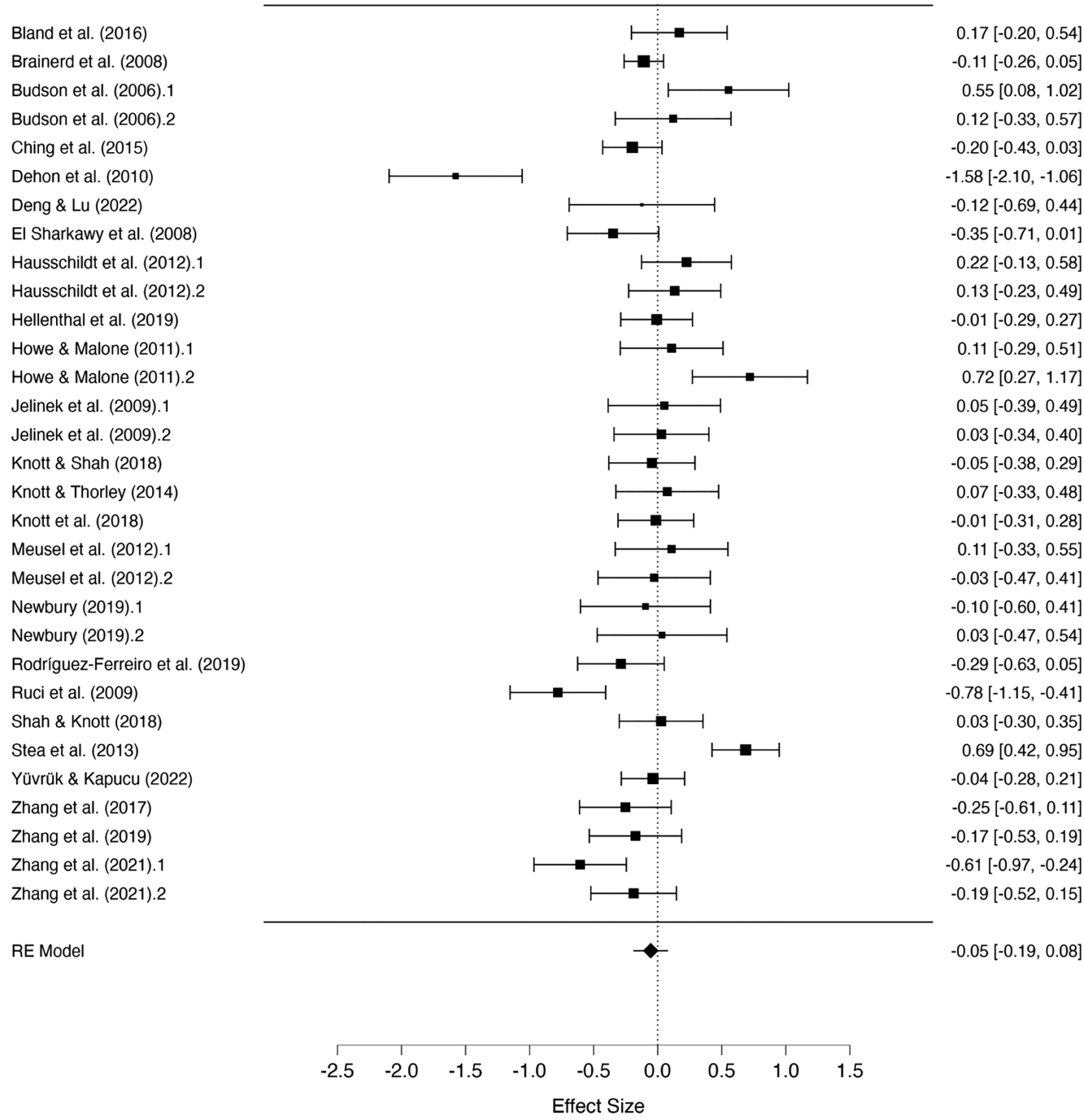
Figure 5
Funnel Plot for the Recognition Test Outcome



[3] Including Zhang (2017) in the recall outcome would make the Egger's test significant. However, this result should be interpreted with caution because the large heterogeneity, further increased by including extreme outliers, may erroneously point toward publication bias (Ioannidis & Trikalinos, 2007).

Figure 6

Forest Plot for the Recognition Test Outcome Controlling for Potential Response Bias

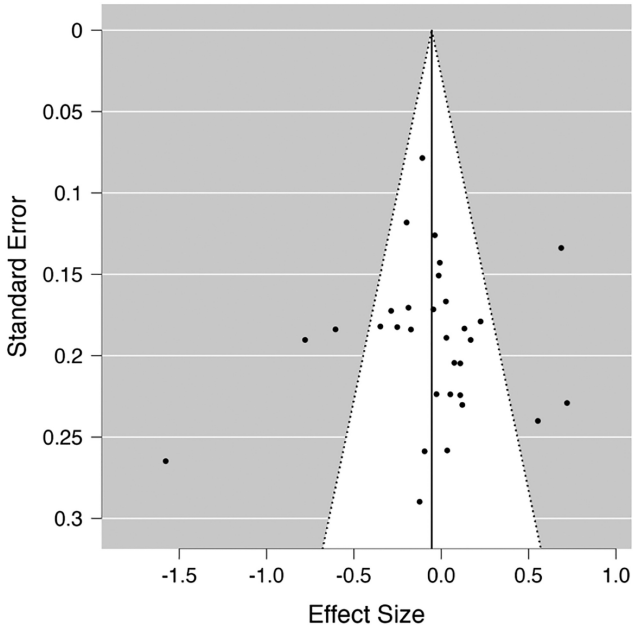


Note. RE = random-effects.

for potential response bias. On the one hand, the number of available effect sizes for the recognition outcome controlling for potential response bias was less than half of the total recognition estimates. This resulted from the fact that nonstudied unrelated items are often not reported per valence condition. Hence, the null effect for the recognition outcome when controlling for response bias may be due to decreased statistical power, rather than an indication of response bias. This explanation is underlined by our sensitivity analysis, showing that adding imputed missing values lead to a conclusive overall effect in the hypothesized direction. On the other hand, we cannot reject the possibility that negative valence does not increase false recognition but may instead increase response bias.

Figure 7

Funnel Plot for the Recognition Test Outcome Controlling for Potential Response Bias



Thus, individuals may recognize more words as "old" when the word is negative, irrespective of whether the item was related to lists or not. This explanation is in line with prior research stating that negative valence increases response bias to negative items, rather than false memories (e.g., Budson et al., 2006; Yüvrük & Kapucu, 2022).

Second, arousal showed significant moderating effects contrary to the expected direction when the primary recall analysis favored the null hypothesis (i.e., elevated arousal in negative lists further decreased false memories in negative valence, vs. neutral). We argue that our operationalization of assessing arousal differences between valence conditions may have been inadequate. Negative critical lures or arousal ratings were not always reported and the resulting reduced power makes it difficult to draw reliable conclusions. Furthermore, using only the arousal level of negative critical lures as a moderator may not be appropriate regarding arousal differences between valence conditions because arousal of negative material might not fully correlate with arousal differences between negative and neutral valence. To tackle this limitation, we conducted an additional exploratory moderation analysis in which we compared findings of studies that explicitly reported having matched arousal between neutral and negative valence versus those that did not (see Appendix D). Matched arousal did not moderate the recall and recognition findings. However, the moderation was significant for the recognition outcome controlling for potential response bias such that studies that reported matched arousal levels were associated with more false memories for negative valence (vs. neutral). Notably, the latter result should again be treated with caution as the primary valence analysis revealed null results.

Lastly, as prior literature suggests that valence effects on false memory formation may be attributed to emotional material being generally more densely integrated and richly interconnected (Howe et al., 2010), we explored whether the observed results differ when only considering studies that explicitly mentioned having controlled for associative strength. Importantly, the primary results remained robust for all three outcomes. Hence, in our meta-analysis, the effect of valence on false memory formation in the DRM did not purely depend on associative strength differences between negative and neutral valence.

Two main limitations should be considered. First, the number of available studies for our moderators was limited, implying low power to detect a given moderating effect (Schmidt & Hunter, 1996). Second, the quality of a meta-analysis is determined mainly by the quality of the studies it includes. Overall, many studies did not report the desired design and sample quality.[4] In sum, our results indicate that negative valence may not increase false memories systematically but may instead depend on how memory is tested, that is, whether recall or recognition is used. Moreover, the effect of valence on recognition may be associated with a shift in the response criterion for negatively valanced items. We further expand upon these concluding points in the General Discussion section.

Study 2: Preregistered Replication

For the current replication effort, rather than replicating a specific published protocol, we conducted a paradigmatic replication (inspired by, but not identical, to the replication format by Vohs et al., 2021). A direct replication aims to repeat a specific published protocol deemed suitable to produce the effect of interest (Simons, 2014). However, a downside of direct replications is that the original protocol may not incorporate subsequent knowledge on the effect of interest. Therefore, we chose to conduct a paradigmatic replication of a DRM protocol developed based on the literature review in preparation for the meta-analysis. We chose procedural and design components based on their merits in prior literature. Specifically, we used validated stimuli sets balanced in arousal and backward associative strength. Furthermore, the materials were counterbalanced with randomized order. In addition, we employed a large sample of participants, relied on Bayesian statistics, and applied open science practices.

Briefly, 300 participants ($n_{final} = 278$) completed the DRM procedure, including recall and recognition tests in an online setting. The main goal was to examine whether negative (vs. neutral valence) induced more false memories for critical items in the recall and recognition test. Additionally, we investigated whether there was evidence of differential discriminability or response bias tendencies for valence conditions. Lastly, we tested whether arousal was associated with false memory formation (please see preregistration for a more detailed list of preregistered hypotheses). Ethics approval was obtained by the Ethics Review Board of the University of Amsterdam (2021-CP-14007). In the following, we report how we determined our sample size, all data exclusions, all manipulations, and all measures used in the study.

Transparency and Openness

We adhered to the JARS guidelines for article reporting standards. All data, analysis code, and research materials are available

[4] To note—mean backward associative strength or learning phase order was only coded by what was explicitly reported. Some studies may have controlled for respective factors but did not report it in their article.

at https://osf.io/9hxrg/. Data were analyzed using RStudio (Version 1.2.5019) and JASP (Version 0.14.1). The study was preregistered; the preregistration can be found at https://osf.io/wqfkh.

Sample

As the preregistered Bayesian stopping rule (i.e., conclusive evidence for the primary analysis in the recall and recognition test) was not met for the first two batches of $n = 100$, we collected data of $n = 300$ participants (prespecified upper limit). Participants completed one online session on Prolific.co (30 min) and were compensated monetarily (€4.50). The inclusion criteria for study participation were: first language being English, age between 18 and 45 (i.e., to minimize developmental or aging differences; Abichou et al., 2020; Otgaar et al., 2013), use of windows computer (i.e., as the graphics work most consistently with this operating system) with working audio and willingness to download the *Inquisit* player software. The inclusion criteria were built-in prescreening measures of Prolific and we filtered participants accordingly. Furthermore, there were three exclusion criteria. First, we excluded participants who did not complete the session's necessary parts: learning phase, recall, and recognition tests. Second, we excluded participants who failed the attention check during the learning phase. Third, participants were excluded if they gave "old" responses to more than 95% or less than 5% of the trials in the recognition test as this may indicate decreased attention by just "clicking through" (i.e., actual percentage of "old" items = 37.5%).

Design

The study implemented a within-subject quasi-experimental design (i.e., word lists are not randomly allocated to valence conditions but are inherently neutral or negative) with list valence (levels: negative vs. neutral), and negative list arousal (levels: low vs. high) as the independent variables (within-subject), resulting in a total of three conditions (low-arousal negative, high-arousal negative and low-arousal neutral). The three dependent variables were: rates of reported studied words, nonstudied critical words and nonstudied unrelated words and were calculated and analyzed for the recall and recognition test separately.

Materials

Stimuli

In total, 16 word lists were selected from existing lists (using 12 words in each list), controlling for the need for nonoverlapping words and satisfying desired valence, arousal, and mean backward associative strength ratings. All neutral lists (i.e., critical lures: mountain, stove, car, cold, lion, command, king, bitter) were adopted from Roediger et al. (2001). The negative list with the critical lure "cry" was extracted from Shah and Knott (2018), and all other negative lists (i.e., critical lures: sick, trash, dead, thief, spider, anger, hurt) were extracted from Chang et al. (2021).

To avoid stimulus-specific effects (Wells & Windschitl, 1999), there were two sets of eight lists randomized across participants by the program *Inquisit*. We used the data of Warriner et al. (2013) to examine valence and arousal measures for each word. Please see the preregistration for the two sets of stimuli, including valence, arousal, and mean backward associative strength ratings. There were no outliers (defined as more than 1.5 times the interquartile range above the third quartile or below the first quartile) in the mean backward associative strength of lists, and no evidence for differential mean backward associative strength ratings between neutral and negative valence conditions, $t(9.05) = 0.29$, $p = .781$, $BF_{01} = 2.27$. Furthermore, across all 16 lists, mean valence ratings differed between neutral and negative lists, $t(13.14) = 12.21$, $p < .001$, $BF_{10} = 1{,}638{,}877$, and mean arousal ratings differed between low- and high-arousal lists, $t(4.42) = -5.11$, $p = .003$, $BF_{10} = 546.97$.[5] More specifically, across the eight negative lists, mean arousal ratings differed between low- and high-arousal negative lists, $t(5.96) = -3.56$, $p = .006$, $BF_{10} = 9.04$. Furthermore, arousal did not statistically differ between low-arousal negative and neutral lists, $t(5.01) = 1.12$, $p = .312$, $BF_{01} = 1.33$. Notably, some analyses did not reach conclusive evidence; however, we argue that this is attributable to the low number of comparisons (i.e., low power).

Signal Detection Measures

For the signal detection analyses on the recognition test, we calculated a discriminability index per valence condition (i.e., how well participants can accurately discriminate between studied and nonstudied items) and a response bias index per valence condition (i.e., whether participants are favoring one response type, for example, "old," over the other when comparing studied and nonstudied items). For all computations, raw scores were corrected following the procedure recommended by Snodgrass and Corwin (1988). The formulas for all response bias and discriminability computations can be found in Appendix E.

The calculation was twofold for the response bias index: measuring response bias including studied and nonstudied critical words per valence, and studied and nonstudied unrelated words per valence. The neutral point ($c = 0$) indicates that no response was favored, while negative values correspond to a liberal bias (responding "old" frequently), and positive values to a conservative bias (responding "old" infrequently; Stanislaw & Todorov, 1999).

For the discriminability index, we calculated two difference scores between correct "old" responses to studied words (i.e., hit rate) and incorrect "old" responses to nonstudied critical words (i.e., false alarm rate) per valence condition, as well as between correct "old" responses to studied words (i.e., hit rate) and incorrect "old" responses to nonstudied unrelated words (i.e., false alarm rate) per valence condition. In both cases, d' values close to zero indicated that the participant could not discriminate between studied words and nonstudied words. Higher values indicated that participants tend to (correctly) accept studied and (correctly) reject nonstudied words.

Demographics and Exploratory Measures

For the demographics, the internal information of Prolific was extracted and coupled to our data with participants' anonymous Prolific IDs. We extracted information on participant's gender (female, male), age (continuous variable), current psychological health (yes, no,

[5] Bayes factors differed slightly from the reported BFs in the preregistration due to a minor coding error, but the conclusions were unaffected by this correction.

prefer not to answer), nationality (dropdown list), country of birth (dropdown list), the highest level of obtained education (less than a high school degree, high school degree or equivalent, bachelor's degree, master's degree, doctorate), and student status (yes, no). For exploratory purposes, additional measures (i.e., Fantasy Proneness by Merckelbach et al., 1998; Dissociative Experiences Scale by Bernstein & Putnam, 1986) were assessed.

Procedure

The study was conducted online in English with native English-speaking participants. For an overview of the procedure, please see Figure 8. The session commenced once participants read the information letter and consented to participation. Afterward, participants were randomly assigned to one of the two randomized word list sets. They were told that they would be participating in a simple memory test. The task entailed five phases (i.e., learning phase, filler task, recall test, recognition test, end-of-study questionnaires including exploratory measures and debriefing).

In the learning phase, eight lists (four neutral and four negative) of 12 words (black font on white background) were presented on the screen at a 2-s rate, with a 12-s break between lists. Words within a list were presented in decreasing order of mean backward associative strength, and the lists were presented in a quasirandom intermixed order for valence (i.e., randomized with the exception that lists of the same valence were never presented consecutively three times or more). Participants were instructed to learn the words to the best of their ability for a later memory test. At the end of the learning phase, participants were asked to press the spacebar upon instruction within 6 s (i.e., attention check). After the learning phase, participants completed math exercises. We used simple addition and multiplication problems with two numbers ranging from 1 to 14. The task moved to the next phase automatically after 5 min.

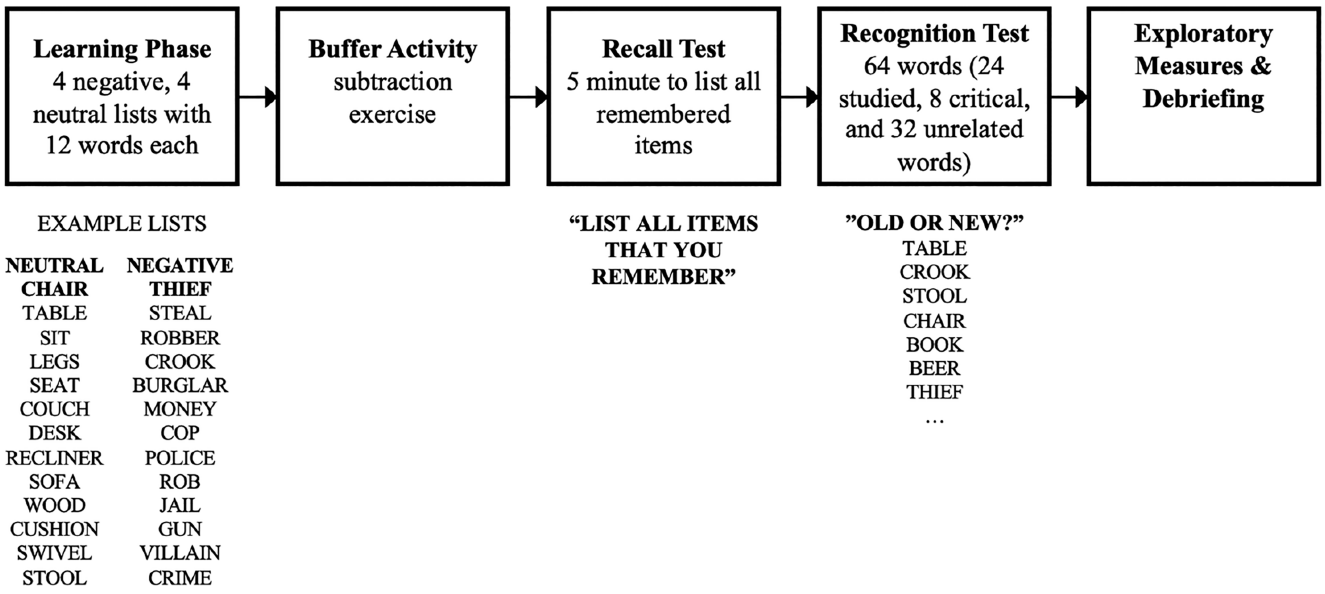
In the recall test, participants had 5 min to write down, into a textbox, as many words they remembered from the learning phase. *Inquisit* coded entries as separate words if the following separators were used: \ | /; :, . or the shift function. Written responses were reviewed manually, and the following rules were added to decrease the ambiguity of written textbox responses. Adding onto Nelson et al.'s (2004) prespecified rules on how to code written one-word responses, we coded words as respective studied or critical words in the following cases: minor spelling errors, U.K./U.S.-related differential spelling, plurals of words, different grammatical tenses, slight deviation of a word with the same word stem (with a maximum difference of two letters in total; different letters need to be consecutive). For nonstudied unrelated words, all words were coded as a recalled nonstudied unrelated word; hereby, only excluding nonwords. Directly after the recall test, participants completed a self-paced recognition test. In total, 64 words appeared on the screen one by one. Participants had to press "F" on their keyboard if the word was "old" (encountered in the learning phase) and "J" if it was "new" (not seen during the learning phase). The recognition test consisted of 24 studied words (12 neutral and 12 negative, of the first, eighth, and 10th position of each list; as suggested by Pardilla-Delgado & Payne, 2017), their corresponding critical lures (four neutral, four negative), 24 unrelated new words (12 neutral and 12 negative, out of the other counterbalancing set) and their corresponding eight unrelated lures (four neutral, four negative). At the end of the study, participants completed additional questionnaires and were then thanked for their participation, debriefed (with a downloadable debriefing document), and compensated.

Statistical Analysis

We used a Bayesian approach for all confirmatory analyses with a standard Jeffreys-Zellner-Siow prior with scaling factor $r = .707$. Additionally, a BF robustness check was conducted for the primary analyses to show the BF for a wide range of prior distributions, allowing to examine the extent to which conclusions depend on the prior specification. For all analyses, we classified $BF_{10} > 5$ as

Figure 8

Overview of the Planned Procedure



convincing evidence in favor of the alternative hypothesis over the null hypothesis and $BF_{01} > 5$ as convincing evidence in favor of the null hypothesis over the alternative hypothesis. We want to note that if a Bayesian analysis shows a $BF_{10} < 5$ and $BF_{01} < 5$, we did not draw definite conclusions on the directionality of the respective effect.

Manipulation Check

As a positive control, we tested the DRM effect in the recognition test for neutral material, which has already been robust in online studies (Zwaan et al., 2018). We conducted a Bayesian one-tailed paired-samples *t* test to compare the null hypothesis (i.e., no difference in "old" responses) to a one-sided alternative (i.e., more "old" responses to neutral nonstudied critical in comparison to neutral nonstudied unrelated words). We investigated extreme outliers (i.e., difference score that is larger or smaller than the third and first quartiles, respectively, by three times of the interquartile range). Due to using a large sample size in the study, we refrained from testing normality. Furthermore, we examined potential ceiling (correct response > 95%) or flooring effects (correct response < 5%) in both outcomes.

Confirmatory Analysis

We reported Cohen's *d* effect sizes with their respective 95% CIs. For the primary analyses of both the recall and recognition outcomes, we conducted a Bayesian one-tailed paired-samples *t* test to compare the null hypothesis (i.e., no difference in recalled/recognized nonstudied critical words between neutral and negative valence) to a one-sided alternative (i.e., more recalled/recognized nonstudied critical words in negative compared to the neutral condition). For all secondary confirmatory analyses, we conducted separate Bayesian paired-samples *t* tests to compare the null hypothesis (i.e., no difference between contrasts in secondary measure analyses) to a one-sided alternative (see preregistration for more detailed information).

Exploratory Analysis

True Memory. We reran the primary analyses on true memory rates to investigate the effect of negative valence (vs. neutral) on true memory in the DRM paradigm. More specifically, we investigated whether hit rates differed between neutral and negative valence in two separate Bayesian two-tailed paired-samples *t* tests to compare the null hypothesis (i.e., no difference in recalled/recognized studied words) to a two-sided alternative (i.e., difference in recalled/recognized studied words between neutral and negative valence). As this analysis was exploratory, we made no predictions regarding the finding's directionality.

Results

Sample

Of the 300 participants, data of one participant were unavailable due to a system glitch, two participants were excluded because they gave "old" responses to more than 95% or less than 5% of the trials in the recognition test, and 19 participants were excluded because they failed the attention check. Thus, the final sample consisted of 278 participants (92.67% of the overall sample). Overall, 196 females participated (70.50% of the final sample; 81 males, one missing response). The mean age of the final sample was 30.59 ($SD = 7.37$, range = 18–45, 82 current students), and most participants reported to have a minimum obtained education level of a high school degree or equivalent ($n = 123$, n bachelor's degree = 114, n master's degree = 34, n less than a high school degree = 4, n doctorate = 3). The majority of the sample ($n = 215$, 77.34%) did not report a current diagnosed mental disorder. Furthermore, the majority of the sample was born in the United Kingdom ($n = 133$, 47.84%) and/or of U.K. nationality ($n = 138$, 49.64%).

Manipulation Check

For the positive control analysis, the BF indicated strong evidence in favor of the alternative hypothesis over the null hypothesis ($BF_{10} = 1.70e^{+86}$). Thus, participants reported more neutral nonstudied critical lures ($M = 0.65$, $SD = 0.24$) in comparison to neutral nonstudied intrusions ($M = 0.15$, $SD = 0.17$, Cohen's $d = 2.33$, 95% CI [2.04, 2.62], no extreme outliers). Furthermore, we also did not find a floor or ceiling effect for both the recall ($M = 0.21$, $SD = 0.12$) and the recognition test outcome ($M = 0.70$, $SD = 0.10$).

Confirmatory Analysis

Primary Focus. For the recall outcome, we conducted a Bayesian one-tailed paired-samples *t* test to compare the null hypothesis (i.e., no difference in recalled nonstudied critical words between neutral and negative valence) to a one-sided alternative (i.e., more recalled nonstudied critical words in the negative valence compared to the neutral condition). On average, participants reported 0.82 out of four neutral critical words ($SD = 0.79$, proportion $M = 0.20$, $SD = 0.20$) and 0.84 negative critical words ($SD = 0.85$, proportion $M = 0.21$, $SD = 0.21$; no extreme outliers). The resulting BF_{01} was 10.83, showing strong evidence in favor of the null over the alternative hypothesis, indicating that there was no statistically differential false memory report in the recall test between valence conditions (Cohen's $d = 0.03$, 95% CI [−0.11, 0.17]). The BF robustness check showed that the directionality of the result did not change when adopting a range of prior distributions (i.e., Cauchy prior width ranging from 0 to 1.5); thus, the analysis did not depend on the prior specification. When only including low-arousal negative lists in the sensitivity analysis, the conclusion remained the same. There was no statistical false memory difference in recall when comparing neutral and low-arousal negative lists ($BF_{01} = 20.44$, Cohen's $d = -0.03$, 95% CI [−0.18, 0.11]).

In the recognition test, participants reported on average 2.59 out of four neutral critical words ($SD = 0.98$, proportion $M = 0.65$, $SD = 0.24$) and 2.94 negative critical lures ($SD = 1.00$, proportion $M = 0.73$, $SD = 0.25$) and there were no extreme outliers. In the Bayesian one-tailed paired-samples *t* test, we found strong support for the alternative over the null hypothesis ($BF_{10} = 28{,}959.52$, Cohen's $d = 0.35$, 95% CI [0.21, 0.48]). As hypothesized, participants reported more false memories for negative compared to neutral material. The BF robustness check showed that the conclusion remained the same when adjusting the prior widths. In the sensitivity analysis comparing neutral to low-arousal negative valence, the evidence still favored the alternative over the null hypothesis ($BF_{10} = 279.51$, Cohen's $d = 0.24$, 95% CI [0.14, 0.44]), indicating that negative valence (vs. neutral) produced more false recognition even when only low-arousal negative material was considered.

Secondary Focus.

Response Bias.[6] To investigate whether participants showed a response bias tendency to negative material in the recognition test, we conducted two Bayesian one-tailed paired-samples *t* tests to test whether participants showed lower response bias scores in the negative compared to the neutral condition. The first analysis, which included nonstudied critical items, revealed average *c′* scores of −0.45 (*SD* = 0.44) for neutral and −0.52 (*SD* = 0.47) for negative valence. Thus, individuals tended toward a liberal response bias in both conditions, but the evidence for a valence-related difference was inconclusive ($BF_{10} = 2.72$, Cohen's $d = 0.16$, 95% CI [0.03, 0.29]). In contrast, the second analysis, which included nonstudied unrelated items, revealed average *c′* scores of 0.30 (*SD* = 0.42) for neutral and 0.13 (*SD* = 0.44) for negative valence, indicating a conservative response tendency for both valence conditions. However, in this case, a BF_{10} of 100,753,218 (Cohen's $d = 0.39$, 95% CI [0.27, 0.50]) indicated strong conclusive evidence for the alternative hypothesis that negative valence showed lower c' scores in comparison to neutral valence. This difference between the two analyses in the directionality of bias (i.e., liberal vs. conservative bias) is in line with the nature of the items. Critical items are specifically used to induce false memories; hence, more "old" responses. Therefore, an average conservative bias for the latter calculation in comparison to an average liberal response bias in the former calculation can be expected. Nevertheless, we found that participants less often accurately responded "new" to nonstudied unrelated items when they were negative compared to when they were neutral.

Discriminability. For both calculations of discriminability (including nonstudied critical items, and nonstudied unrelated items), separate Bayesian one-tailed paired-samples *t* tests were conducted to compare the null hypothesis (i.e., no difference in discriminability measures between valence conditions) to a one-sided alternative (i.e., lower discriminability in the negative compared to the neutral condition). Neutral valence showed an average discriminability index of 0.21 (*SD* = 0.71), whereas negative was −0.10 (*SD* = 0.64). Based on $BF_{10} = 11{,}459{,}728$ (Cohen's $d = 0.45$, 95% CI [0.30, 0.60]), there was strong conclusive evidence for the alternative over the null hypothesis. Thus, participants tended to more often (correctly) accept studied words and (correctly) reject nonstudied critical words for items of neutral valence (vs. negative). Moreover, the result based on nonstudied unrelated items showed the same directionality. Discriminability index was 1.70 (*SD* = 0.87) for neutral valence and 1.21 (*SD* = 0.68) for negative valence ($BF_{10} = 3.58e^{+20}$, Cohen's $d = 0.62$, 95% CI [0.50, 0.74]). Thus, participants tended to more often (correctly) accept studied words and (correctly) reject nonstudied unrelated words for items of neutral versus negative valence.

The Effect of Arousal. When investigating the effect of arousal, many extreme outliers appeared (126 in recall, 130 in recognition). However, as the outcome (i.e., number of critical lures reported) ranged only from 0 to 2, there was less variability in individual responses, which biased the outlier analysis (individuals at the end of the scale that is, reporting 0 or 2, were regarded as outliers). Thus, when outliers were removed, all means became identical. Therefore, we deviated from the preregistration and did not remove extreme outliers.

On average, participants reported 0.45 (*SD* = 0.59, proportion *M* = 0.22, *SD* = 0.30) high-arousal and 0.39 (*SD* = 0.56, proportion *M* = 0.20, *SD* = 0.28) low-arousal negative critical lures in the recall test. When testing the Bayesian one-tailed paired-samples *t* test (i.e., to test whether participants recalled more nonstudied critical words in the high-arousal negative compared to the low-arousal negative condition), we found inconclusive evidence. Although the BF indicated that the null hypothesis was more than four times more likely than the alternative hypothesis ($BF_{01} = 4.45$, Cohen's $d = 0.09$, 95% CI [−0.07, 0.25]), it did not reach our prespecified threshold for conclusive evidence.

In recognition, participants reported 1.47 (*SD* = 0.63, proportion *M* = 0.73, *SD* = 0.32) high- and 1.47 (*SD* = 0.66, proportion *M* = 0.73, *SD* = 0.33) low-arousal negative critical lures on average. When testing the Bayesian one-tailed paired-samples *t* test, there was strong evidence for the null over the alternative hypothesis ($BF_{01} = 14.88$, Cohen's $d = <0.01$, 95% CI [−0.15, 0.15]). Hence, there was no statistical difference in false memories between high-arousal and low-arousal negative words in the recognition test.

Exploratory Analysis

True Memory. For the recall outcome, participants reported 10.51 neutral (*SD* = 6.79, proportion *M* = 0.22, *SD* = 0.14) and 9.93 negative studied words (*SD* = 6.66, proportion *M* = 0.21, *SD* = 0.14) on average (i.e., no extreme outliers). Although inconclusive, the resulting BF_{01} was 4.63, showing evidence in favor of the null over the alternative hypothesis, indicating that there was no statistically differential true memory report in the recall test between valence conditions (Cohen's $d = 0.09$, 95% CI [−0.02, 0.20]). In the recognition test, participants reported on average 8.44 neutral (*SD* = 2.16, proportion *M* = 0.70, *SD* = 0.18) and 8.12 negative studied words (*SD* = 2.13, proportion *M* = 0.68, *SD* = 0.18), and there were no extreme outliers. Here, we did not find conclusive support for either hypothesis ($BF_{01} = 1.08$, Cohen's $d = 0.15$, 95% CI [0.02, 0.28]).

Discussion

The current replication ($n = 278$) investigated the effect of negative valence on false memory formation within the DRM paradigm. We conducted a highly powered, preregistered paradigmatic replication with the operationalization and procedural elements carefully chosen based on prior literature, to investigate the valence effect on recall and recognition outcomes separately. Similar to the meta-analysis, we found strong evidence for the original DRM effect with neutral stimuli ($BF_{10} = 1.70e^{+86}$; $d = 2.33$).

Overall, the recall outcome did not reveal statistically differential false memory formation between valence conditions and this effect was not influenced by arousal differences or response bias. For the recognition outcome, different results emerged. In line with the hypothesis, negatively valanced words induced more false recognition than neutral words (even when only examining low-arousal negative content). Moreover, participants displayed reduced discriminability for negative material—they were worse in discriminating negative studied from nonstudied items. However, we also found evidence that negative, compared to neutral content, induced a stronger tendency toward "old" responses in the recognition test, irrespective of false item type

[6] We had preregistered to draw inferences on response bias tendencies in the recall test by comparing true memory rates per valence conditions. As we now added true memory analyses exploratorily to the project, we instead present the recall true memory rate differences per valence conditions in the True Memory section.

(nonstudied critical and unrelated). Thus, participants exhibited a shift in response tendency for negative material—they generally reported remembering more negative items, accurately or inaccurately.

Several limitations need to be noted. First, even though the DRM task has been tested in prior online research (Zwaan et al., 2018), only the original DRM recognition effect was verified in this prior study. As an online setting might allow for less procedural control than a laboratory study, we implemented attention checks, time-out procedures, and prespecified timings on pages (e.g., a specific amount of time had to pass before participants could click "continue"). Nevertheless, we recognize that participants were not observed in real time. Therefore, the examined null effect in the recall test might result from the online nature of the test. Participants may have felt less pressured to "do their best" and may have generally reported fewer words, making it more difficult to find effects across valence conditions. However, the absence of a floor effect in the recall test is inconsistent with the above explanation. Furthermore, null effects within the recall outcome align with the mixed nature of recall results in prior literature (e.g., Joormann et al., 2009).

Second, we acknowledge that memory tests were not conducted in a counterbalanced order; the recall test always preceded the recognition test. DRM research has shown that performance on recognition can be affected by the responses on the previous recall test (Brainerd et al., 2008). However, the fixed order of memory tests in DRM studies is the general procedure and closely mimics real-life situations in which recognition-like tasks follow free elaboration, for example, in best-practice forensic interviewing protocols (Brainerd & Reyna, 2005).

To conclude, the findings of the current replication study are in line with the meta-analytic result and prior literature showing valence effects on false memory production in the DRM paradigm in the recognition test (e.g., Brainerd et al., 2008, El Sharkawy et al., 2008), and more mixed results on the effect of valence on false memories in the recall test (e.g., Joormann et al., 2009). Similar to the meta-analytic results, we found that the effect of valence on false memory formation within the DRM paradigm may not be due to a systematic semantic difference in emotional information, but may instead depend on the type of memory test. Moreover, the effect of valence in the recognition outcome may be associated with a more liberal (or less conservative) response tendency toward recognition in negative valence, irrespective of the false memory type (e.g., nonstudied critical or unrelated). We further expand upon these explanations in the General Discussion section.

General Discussion

In recent decades, many studies have investigated the effect of content valence on false memory formation in the DRM paradigm. However, clear empirical evidence on the directionality of the effect has been lacking so far, specifically for the recall test. To get a more precise estimate of the valence effect on false memory formation in the DRM task, using recall and recognition tests, we conducted both a meta-analysis and a paradigmatic replication. Our primary focus was the fuzzy-trace theory (Reyna & Brainerd, 1995), predicting increased false memories for negative versus neutral valence in both tests, as negative material is hypothesized to enhance gist processing more strongly. Notably, other prominent frameworks (i.e., associative–activation, activation–monitoring; Howe et al., 2009; Roediger & McDermott, 1995) also align with our predicted outcomes.

By conducting both a meta-analysis and replication on the topic, we investigated the findings' validity and reliability. Meta-analysis results are essential in statistically aggregating quantitative summaries of the literature. Still, preregistered replications can complement this knowledge by contributing effect estimates, while avoiding publication bias and questionable practices. Even though both approaches aim to estimate the true effect size of a phenomenon, recent work shows that they tend to be discrepant (Kvarven et al., 2020; Lewis et al., 2022; Nieuwenstein et al., 2015). According to such prior research, meta-analytic average effect size estimates tend to be three times larger than their replication counterparts. It is argued that this discrepancy does not discount one of the two approaches and can be explained, at least to some extent, by publication bias or questionable research practices in studies included in the meta-analysis. Additionally, studies differ along multiple relevant dimensions (e.g., method, sample, stimuli, analysis), further increasing heterogeneity in the meta-analysis.

We argue that both meta-analyses and replications have important strengths. However, their systematic comparison can further advance to understand the studied phenomenon. Therefore, in light of the extensive empirical and theoretical research on the effect of valence on false memory formation in the DRM paradigm, we conducted and presented both a meta-analysis and replication study together.

In contrast to the expected divergence of effect sizes, the results of our meta-analytic and replication efforts largely converged. In line with the majority of recognition studies (see Bookbinder & Brainerd, 2016), our meta-analysis and replication revealed a robust valence effect for the recognition outcome, such that negative valence increased false memories compared to neutral valence. Retrieved effect sizes were similar in the two approaches—the replication effect size was $d = 0.35$ ($d = 0.24$ when only including matched-arousal lists), and the meta-analytic average effect was $d = 0.23$. Additionally, arousal differences did not confound the valence effect in the recognition test. Yet, we found indications that the observed recognition effect may be instigated by a stronger liberal response bias tendency to negative versus neutral material. Specifically, for the meta-analysis, the effect of valence became nonsignificant for the recognition outcome when controlling for potential response bias by including neutral and negative nonstudied unrelated word responses. Comparably, in our replication, we found conclusive evidence that negative valence was associated with less conservative response bias compared to neutral valence when controlling for performance on neutral and negative nonstudied unrelated words.

For the recall test, we did not observe the expected valence effect in either approach. Interestingly, in our meta-analysis, false memories tended to even decrease for negative valence (vs. neutral), and this effect was pronounced for higher levels of arousal in negative stimuli. Although the Bayesian evidence was inconclusive, and we did not find the stated effects in the replication, we acknowledge that negative valence (vs. neutral) might reduce false memories in a recall test. This notion is underlined by theoretical perspectives suggesting that negative information (vs. neutral) may actually enhance, rather than diminish, verbatim processing and retrieval monitoring (Doss et al., 2020; Gallo et al., 2009; Howe et al., 2010; Palmer & Dodson, 2009). However, it contrasts our true memory finding in the meta-analytic recall outcome in which true memories were also decreased in response to negative valence (vs. neutral).

Notably, by examining true memories in response to valence manipulations exploratorily, we made further inferences about potential underlying theories. In short, while spreading activation-based single-process theories imply the same directionality of valence effects for both true and false memories, the fuzzy-trace theory (i.e., two processing systems)

would imply that true memories would remain unaffected or be suppressed when false memories differ. In both approaches, the more consistent finding indicated that true memory rates did not statistically differ even when false memory rates did. Hence, these results largely align with gist-based processing within the fuzzy-trace account—increased gist processing for negative content (vs. neutral) increases false memory but does not necessarily affect verbatim processing (i.e., accurate recollection rejection) and thus, true memory report. Importantly, while the true memory results may offer a glance at underlying mechanisms, a direct inference on underlying processes may not be possible due to other potential confounding factors across valence conditions (e.g., interitem connectivity, word frequency).

Overall, three overarching conclusions emerge when comparing the meta-analytic and replication findings: First, as indicated above, the two approaches converged; second, both approaches revealed that the effect of valence on false memory formation within the DRM paradigm depends on the type of memory test conducted; third, the effect of valence on false memory observed in the recognition test might be partly attributable to response bias.

Several factors may explain why our meta-analysis and replication results converged. First, we did not find evidence in our meta-analysis for publication bias. Again, this may be because the original (neutral) DRM effect is robust, and DRM studies are published irrespective of differences between valence conditions. Second, some researchers propose that differences in replication and meta-analysis effect sizes could arise due to biased replication target selection (Laws, 2016; Pittelkow et al., 2021). For example, researchers might choose to replicate a simpler and cheaper study version to save resources, or a surprising and more improbable finding to increase publication chances, making the replication selection prone to bias. When assuming additional heterogeneity, this increases the likelihood for replication studies to select operationalizations (i.e., based on sample or design choices) associated with skewed effect sizes; hereby creating a difference between the average effect size in meta-analyses and observed effect sizes in replication studies. In line with this, prior research has hypothesized an association between heterogeneity in the meta-analysis and divergence in meta-analytic and replication effect sizes (Kvarven et al., 2020). We propose that by selecting a generally robust paradigm (i.e., DRM), and carefully examining design and sample choices based on the merits of prior literature for the replication effort, we may have chosen an operationalization that yielded comparable results between meta-analytic and replication effect sizes.

The second conclusion refers to the difference between the results of the two memory tests. Prior literature indicates that recognition tests may be more sensitive to gist memory (i.e., based mainly on the principle of familiarity; Bookbinder & Brainerd, 2017; Gomes et al., 2013), whereas recall tests may more strongly depend on verbatim memory (i.e., based on recollective retrieval; Abadie et al., 2021; Brainerd et al., 2009). Thus, the two memory tests trigger different processing systems; familiarity-based retrieval depends more on areas surrounding the hippocampus, while recollection-based retrieval depends on hippocampal and prefrontal processing (Yonelinas, 2002). Given that the fuzzy-trace theory hypothesizes negative valence to increase false memories by enhancing gist processing (Brainerd et al., 2008; Reyna & Brainerd, 1995), increased false memories for negative valence may more robustly arise in recognition than recall tests. Furthermore, as negatively valanced content (vs. neutral) is hypothesized to be heightened in connectedness, spread activation, and interrelatedness (Anderson & Bower, 1973; Collins & Loftus, 1975; Howe et al., 2010), this increased activation may lead to more confusion about what was presented, further triggered by offering individuals answer options (i.e., recognition test) compared to a recall test in which individuals are not distracted by the presence of critical lures (Howe et al., 2010). In addition, it is important to consider the timing of memory tests when interpreting valence findings. Free recall can be administered either after each or all lists, while recognition is typically conducted only after all lists. Consequently, the delay difference between recall and recognition could pose a potential confound. More specifically, at immediate testing, the increased salience of list items might counteract the heightened associative activation or gist processing attributed to negative valence (vs. neutral). Hence, in immediate recall, monitoring effectiveness (i.e., activation-based theory) or recollection rejection (i.e., fuzzy-trace theory) might be elevated for negative valence (vs. neutral), explaining the observed null (or even reversed) valence findings. Although our meta-analysis did not explore this potential confound, our replication minimized this timing difference by conducting both tests after all lists. Nevertheless, distinct valence-related findings persisted.

Furthermore, while a general disparity between recall and recognition results was demonstrated, both our meta-analysis and replication showed that the reported valence effect for the recognition outcome might be confounded by a differential response tendency toward negative valence (vs. neutral). In line with our results, several studies found that increased false memories of negatively valanced material were associated with shifts in response bias (e.g., Brainerd et al., 2008; Budson et al., 2006; Dehon et al., 2010; El Sharkawy et al., 2008; Howe et al., 2010). Additionally, a recent study (Yüvrük & Kapucu, 2022) showed that the effect of valence was nonsignificant when recognition responses were controlled for response bias. Hence, stimulus valence may alter participants' response tendencies in discriminating old versus new items, rather than decreasing memory accuracy per se.

Yet, what could explain a potential response tendency shift? Various theoretical frameworks offer explanations for the systematic bias observed in negative valence. First, different monitoring styles may explain differences in response tendency. Based on the activation-monitoring theory, two monitoring processes underlie any decision process, that is, disqualifying and diagnostic monitoring (Gallo, 2010); the former relies on collateral information, while the latter uses expectations to make a decision. Negative valence (vs. neutral) may motivate diagnostic monitoring (i.e., creating a negative context and expectancy toward the negative), promoting liberal response bias (Yüvrük & Kapucu, 2022). The findings of Calvillo and Parong (2016) underline this notion by showing that a warning manipulation, which is supposed to facilitate more effective monitoring (Watson et al., 2005), reduced individuals' liberal response bias while memory sensitivity remained constant. Hence, differential monitoring processes may explain response bias observed in negative valence (vs. neutral).

Next to differential monitoring styles, the network of negative information (vs. neutral) is also assumed to be elevated in associative (i.e., higher semantic interrelatedness and denser organization; Otgaar et al., 2016; Talmi, 2013) and gist strength (i.e., higher categorical membership and thematic links processing; Talmi, 2013; White et al., 2014). Importantly, gist and associative strength are considered to be (at least partly) independent mechanisms (Brainerd et al., 2020) that distinctly contribute to false memory formation in the DRM paradigm (e.g., Huff et al., 2015; Oliveira et al., 2019). Even though we controlled for associative strength in both approaches (i.e., subgroup

analysis in the meta-analysis, and matched stimuli in replication), we note that we did not assess gist strength. Therefore, our findings may still be influenced by negative valence's increased gist strength (vs. neutral) despite being balanced in associative strength. Specifically, the deeper and more familiarity-based processing that is associated with increased gist may affect negative information more generally (Brainerd et al., 2008; Gomes et al., 2013). Hence, negative items (vs. neutral) may be more strongly endorsed in the recognition test (i.e., "old" responses) irrespective of their false memory type (related or unrelated; Howe et al., 2009), supporting liberal response bias. Crucially, it is even hypothesized that the effect of gist and associative strength in negative material (vs. neural) may go beyond the matching alone. It is argued that perceived gist and associative strength may be heightened even when actual gist and associative strength levels are balanced (White et al., 2014; Yüvrük & Kapucu, 2022), further promoting response bias.

Overall, heightened (perceived) gist processing of negative information (vs. neutral) appears to play a central role in explaining the effect of valence on false memory reports in the DRM paradigm. Accordingly, one may wonder whether the observed valence effect would also generalize to single-event situations when no list can produce or strengthen gist. Particularly, the DRM has often been criticized for lacking ecological validity (e.g., DePrince et al., 2004; Freyd & Gleaves, 1996; Pezdek & Lam, 2007) because learning and reporting lists of words is not a typical process involved in real-life false memories. In contrast, suggestion-induced memory paradigms, that is, the misinformation and implanted memory paradigms, are two experimental procedures investigating false memory formation for more complex events. In short, the misinformation paradigm studies whether participants' memories of an event can be distorted by presenting them with misleading postevent information (Loftus et al., 1978). In implanted memory paradigms, researchers examine the potential of making individuals believe that they experienced a specific past event that did not occur, by employing suggestive methods (e.g., guided imagery, leading questions, and misinformation; Loftus & Pickrell, 1995). Both paradigms examine rich false memories and are specifically relevant to the context of eyewitness testimony (see Loftus, 2005 for more information).

Earlier studies integrating emotion into the misinformation paradigm showed that negative events (vs. neutral) elicited more false memories for peripheral (Van Damme & Smets, 2014) and central details (Porter et al., 2003, 2010; see Porter et al., 2014 for an adapted affective priming paradigm). Furthermore, a recent review on the relationship between emotion and the misinformation effect concluded that negative valence might be associated with increased susceptibility to misinformation, but more strongly for details (Sharma et al., 2023). Similar evidence emerges in false memory implantation paradigms in which negative rich false memories are more often successfully implanted than neutral events (Otgaar et al., 2008; Porter et al., 2008). These results are commonly explained by the paradoxical negative emotion hypothesis—emotion may facilitate attention and memory in general, but at the cost of greater vulnerability to memory distortion (Porter et al., 2008).

From the above considerations, we propose the following future recommendations. First, we recognize that the emotional DRM literature would greatly benefit from more consistent use of open science practices, further enabling the investigation of remaining questions. Second, the underlying theories of the valence effect should be examined more systematically alongside the testing procedure. For example, it would be useful to more strongly implement techniques to measure the direct contribution of verbatim- and gist-based processing to true and false memory (e.g., as was done in Brainerd et al., 2008). Furthermore, measuring response bias tendencies should be more consistently applied to DRM studies. By this, underlying processes can be disentangled, and theoretical explanations for the valence effect in the DRM paradigm can be informed. Third, we believe that the ecological validity of DRM research studying the effect of emotion could be further increased by more strongly endorsing the application of content more related to the studied population (e.g., war-related items for a population with posttraumatic stress disorder; Brennen et al., 2007) or content that more realistically represents real-life experiences (e.g., pictures of scenes, Foley & Foy, 2008). Lastly, we realize that the current default experimental procedure of investigating the effect of valence in the DRM task is quasi-experimental, as stimuli are not randomly assigned to conditions but rather inherently negative or neutral. We argue that a conditioning or priming procedure could be applied to the DRM procedure to increase experimental control. For example, experimentally induced aversive conditioning could be implemented to taint some of the neutral content negative (e.g., by a point-losing game). This specific design, which has not been implemented so far, could advance the field by elucidating false memory formation in the DRM in a rigor manner while increasing ecological validity (i.e., the negative nature of the event taints the content negative). Interestingly, during the review process of this project, we heard from colleagues adopting similar conditioning procedures to the DRM paradigm involving valence manipulations, which should be publicly available soon (Jianqin Wang and colleagues).

Conclusion

Initially, false memories in the DRM were criticized for lacking ecological validity, motivating the incorporation of the emotional component within the paradigm (e.g., DePrince et al., 2004; Freyd & Gleaves, 1996; Pezdek & Lam, 2007). Moreover, mixed results were reported regarding the effect of valence on false memory formation within the DRM paradigm, specifically with respect to the recall outcome. Based on such mixed evidence, we conducted both a meta-analysis and a replication to study this effect. Overall, the results of the two efforts mostly converged. When examining the effect of emotion within the DRM by implementing negative versus neutral content, we did not find a robust valence effect in the recall outcome. Yet, in line with our expectations, negatively valanced material increased false memory reports when assessed by a recognition test. Notably, the valence effect on the recognition outcome was confounded by response bias to negative valence across false item types. Therefore, increased reports of false memories for negative content in the recognition test may not be (fully) due to a systematic semantic difference of emotional information but may be partly attributed to differential response bias tendencies. Still, our central finding remains that individuals report more intrusive errors for negative content (vs. neutral) when tested in recognition, irrespective of its underlying mechanism (i.e., shift in response tendency or memory distortion). Hence, in real-world cases, special care has to be taken for recognition-like memory tests as the negativity of the addressed event may further increase the suggestibility and report of false information.

Constraints on Generality

Our meta-analysis and replication results converge in showing a robust valence effect in the recognition test and more mixed evidence for the recall test, as underlined by prior literature. Notably, our replication was conducted online even though DRM studies are commonly conducted in the laboratory. This may have influenced participants' motivation and decreased overall report in the recall test. Furthermore, the replication study only included individuals between 18 and 45 to exclude developmental and aging effects, which is the common sample tested in this research field. Therefore, the results may not fully generalize to younger or older age groups. Moreover, we only included English-speaking individuals, excluding other language groups. We have no reason to assume that the results depend on other characteristics of the participants, materials, or context.

References

Abadie, M., Gavard, E., & Guillaume, F. (2021). Verbatim and gist memory in aging. *Psychology and Aging*, *36*(8), 891–901. https://doi.org/10.1037/pag0000635

Abichou, K., La Corte, V., Nicolas, S., & Piolino, P. (2020). Les faux souvenirs dans le vieillissement normal: Données empiriques du paradigme DRM et perspectives théoriques [False memory in normal aging: Empirical data from the DRM paradigm and theoretical perspectives]. *Gériatrie et Psychologie Neuropsychiatrie du Viellissement*, *18*(1), 65–75. https://doi.org/10.1684/pnv.2020.0862

Allen, M., & Preiss, R. (1993). Replication and meta-analysis: A necessary connection. *Journal of Social Behavior and Personality*, *8*(6), 9–20.

Anderson, J. R. (1983). A spreading activation theory of memory. *Journal of Verbal Learning and Verbal Behavior*, *22*(3), 261–295. https://doi.org/10.1016/S0022-5371(83)90201-3

Anderson, J. R., & Bower, G. H. (1973). *Human associative memory*. Winston & Sons.

Appelbaum, M., Cooper, H., Kline, R. B., Mayo-Wilson, E., Nezu, A. M., & Rao, S. M. (2018). Journal article reporting standards for quantitative research in psychology: The APA Publications and Communications Board task force report. *American Psychologist*, *73*(1), 3–25. https://doi.org/10.1037/amp0000191

Bauer, L. M., Olheiser, E. L., Altarriba, J., & Landi, N. (2009). Word type effects in false recall: Concrete, abstract, and emotion word critical lures. *The American Journal of Psychology*, *122*(4), 469–481. https://doi.org/10.2307/27784422

Baugerud, G. A., Howe, M. L., Magnussen, S., & Melinder, A. (2016). Maltreated and non-maltreated children's true and false memories of neutral and emotional word lists in the Deese/Roediger–McDermott task. *Journal of Experimental Child Psychology*, *143*, 102–110. https://doi.org/10.1016/j.jecp.2015.10.007

Beckwé, M., & Deroost, N. (2016). Worrying facilitates correct and false memories about negative information. *Journal of Psychology & Psychotherapy*, *6*(3), Article 268. https://doi.org/10.4172/2161-0487.1000268

Begg, C. B., & Mazumdar, M. (1994). Operating characteristics of a rank correlation test for publication bias. *Biometrics*, *50*(4), 1088–1101. https://doi.org/10.2307/2533446

Bernstein, E. M., & Putnam, F. W. (1986). Development, reliability, and validity of a dissociation scale. *The Journal of Nervous and Mental Disease*, *174*(12), 727–735. https://doi.org/10.1097/00005053-198612000-00004

Bland, C. E., Howe, M. L., & Knott, L. (2016). Discrete emotion-congruent false memories in the DRM paradigm. *Emotion*, *16*(5), 611–619. https://doi.org/10.1037/emo0000153

Bookbinder, S. H., & Brainerd, C. J. (2016). Emotion and false memory: The context–content paradox. *Psychological Bulletin*, *142*(12), 1315–1351. https://doi.org/10.1037/bul0000077

Bookbinder, S. H., & Brainerd, C. J. (2017). Emotionally negative pictures enhance gist memory. *Emotion*, *17*(1), 102–119. https://doi.org/10.1037/emo0000171

Bradley, M. M., & Lang, P. J. (1994). Measuring emotion: The self-assessment manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry*, *25*(1), 49–59. https://doi.org/10.1016/0005-7916(94)90063-9

Bradley, M. M., & Lang, P. J. (1999). *Affective norms for English words (ANEW): Instruction manual and affective ratings* (Technical Report C-1). The Center for Research in Psychophysiology, University of Florida.

Brainerd, C. J. (2013). Developmental reversals in false memory: A new look at the reliability of children's evidence. *Current Directions in Psychological Science*, *22*(5), 335–341. https://doi.org/10.1177/0963721413484468

Brainerd, C. J., & Bookbinder, S. H. (2019). The semantics of emotion in false memory. *Emotion*, *19*(1), 146–159. https://doi.org/10.1037/emo0000431

Brainerd, C. J., Chang, M., & Bialer, D. M. (2020). From association to gist. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *46*(11), 2106–2127. https://doi.org/10.1037/xlm0000938

Brainerd, C. J., Holliday, R., Reyna, V., Yang, Y., & Toglia, M. (2010). Developmental reversals in false memory: Effects of emotional valence and arousal. *Journal of Experimental Child Psychology*, *107*(2), 137–154. https://doi.org/10.1016/j.jecp.2010.04.013

Brainerd, C. J., & Reyna, V. F. (2005). *The science of false memory*. Oxford University Press.

Brainerd, C. J., Reyna, V. F., & Howe, M. L. (2009). Trichotomous processes in early memory development, aging, and neurocognitive impairment. *Psychological Review*, *116*(4), 783–832. https://doi.org/10.1037/a0016963

Brainerd, C. J., Stein, L. M., Silveira, R., Rohenkohl, G., & Reyna, V. F. (2008). How does negative emotion cause false memories? *Psychological Science*, *19*(9), 919–925. https://doi.org/10.1111/j.1467-9280.2008.02177.x

Brennen, T., Dybdahl, R., & Kapidzic, A. (2007). Trauma-related and neutral false memories in war-induced posttraumatic stress disorder. *Consciousness and Cognition*, *16*(4), 877–885. https://doi.org/10.1016/j.concog.2006.06.012

Brueckner, K., & Moritz, S. (2009). Emotional valence and semantic relatedness differentially influence false recognition in mild cognitive impairment, Alzheimer's disease, and healthy elderly. *Journal of the International Neuropsychological Society*, *15*(2), 268–276. https://doi.org/10.1017/S135561770909047X

Budson, A. E., Todman, R. W., Chong, H., Adams, E. H., Kensinger, E. A., Krangel, T. S., & Wright, C. I. (2006). False recognition of emotional word lists in aging and Alzheimer disease. *Cognitive and Behavioral Neurology*, *19*(2), 71–78. https://doi.org/10.1097/01.wnn.0000213905.49525.d0

Buratto, L. G., de Azevedo Gomes, C. F., da Silva Prusokowski, T., & Stein, L. M. (2013). Inter-item associations for the Brazilian version of the Deese/Roediger–McDermott paradigm. *Psicologia: Reflexão e Crítica*, *26*(2), 367–375. https://doi.org/10.1590/S0102-79722013000200017

Calado, B., Otgaar, H., & Muris, P. (2018). Are children better witnesses than adolescents? Developmental trends in different false memory paradigms. *Journal of Child Custody*, *15*(4), 330–348. https://doi.org/10.1080/15379418.2019.1568948

Calvillo, D. P., & Parong, J. A. (2016). The misinformation effect is unrelated to the DRM effect with and without a DRM warning. *Memory*, *24*(3), 324–333. https://doi.org/10.1080/09658211.2015.1005633

Carter, E. C., Schönbrodt, F. D., Gervais, W. M., & Hilgard, J. (2019). Correcting for bias in psychology: A comparison of meta-analytic methods. *Advances in Methods and Practices in Psychological Science*, *2*(2), 115–144. https://doi.org/10.1177/2515245919847196

Chang, M., Brainerd, C. J., Toglia, M. P., & Schmidt, S. R. (2021). Norms for emotion-false memory lists. *Behavior Research Methods*, *53*(1), 96–112. https://doi.org/10.3758/s13428-020-01410-7

Ching, T. H., Goh, W. D., & Tan, G. (2015). Exploring dimensionality in the contamination-relevant semantic network with simulated obsessions and

association splitting. *Journal of Obsessive-Compulsive and Related Disorders*, *6*, 39–48. https://doi.org/10.1016/j.jocrd.2015.06.003

Choi, H.-Y., Kensinger, E. A., & Rajaram, S. (2013). Emotional content enhances true but not false memory for categorized stimuli. *Memory & Cognition*, *41*(3), 403–415. https://doi.org/10.3758/s13421-012-0269-2

Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological Review*, *82*(6), 407–428. https://doi.org/10.1037/0033-295X.82.6.407

Dehon, H., Larøi, F., & Van der Linden, M. (2010). Affective valence influences participant's susceptibility to false memories and illusory recollection. *Emotion*, *10*(5), 627–639. https://doi.org/10.1037/a0019595

Deng, R., & Lu, A. (2022). Sleep modulates emotional effect on false memory. *Psychology in Russia: State of the Art*, *15*(1), 154–178. https://doi.org/10.11621/pir.2022.0110

DePrince, A. P., Allard, C. B., Oh, H., & Freyd, J. J. (2004). What's in a name for memory errors? Implications and ethical issues arising from the use of the term "false memory" for errors in memory for details. *Ethics & Behavior*, *14*(3), 201–233. https://doi.org/10.1207/s15327019eb1403_1

Dewhurst, S. A., Anderson, R. J., Berry, D. M., & Garner, S. R. (2018). Individual differences in susceptibility to false memories: The effect of memory specificity. *Quarterly Journal of Experimental Psychology*, *71*(7), 1637–1644. https://doi.org/10.1080/17470218.2017.1345961

Dewhurst, S. A., Anderson, R. J., & Knott, L. M. (2012). A gender difference in the false recall of negative words: Women DRM more than men. *Cognition & Emotion*, *26*(1), 65–74. https://doi.org/10.1080/02699931.2011.553037

Diliberto-Macaluso, K. A., Kazanas, S. A., Altarriba, J., O'Brien, E., Rivera, E., & Smith, J. (2016, November 17–20). *Emotion and emotion-laden words differ on both hits and false alarms: Insights from the DRM paradigm* [Poster presentation]. The 57th Annual Meeting of the Psychonomic Society, Boston, MA, United States.

Doss, M. K., Picart, J. K., & Gallo, D. A. (2020). Creating emotional false recollections: Perceptual recombination and conceptual fluency mechanisms. *Emotion*, *20*(5), 750–760. https://doi.org/10.1037/emo0000590

Duesenberg, M., Weber, J., Schaeuffele, C., Fleischer, J., Hellmann-Regen, J., Roepke, S., Moritz, S., Otte, C., & Wingenfeld, K. (2016). Effects of hydrocortisone on false memory recognition in healthy men and women. *Behavioral Neuroscience*, *130*(6), 635–642. https://doi.org/10.1037/bne0000170

Egger, M., Smith, G. D., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *BMJ*, *315*(7109), 629–634. https://doi.org/10.1136/bmj.315.7109.629

El Sharkawy, J. E., Groth, K., Vetter, C., Beraldi, A., & Fast, K. (2008). False memories of emotional and neutral words. *Behavioural Neurology*, *19*(1–2), 7–11. https://doi.org/10.1155/2008/587239

Esterhuizen, T. M., & Thabane, L. (2016). Con: Meta-analysis: Some key limitations and potential solutions. *Nephrology Dialysis Transplantation*, *31*(6), 882–885. https://doi.org/10.1093/ndt/gfw092

Foley, M. A., & Foy, J. (2008). Pictorial encoding effects and memory confusions in the Deese–Roediger–McDermott paradigm: Evidence for the activation of spontaneous imagery. *Memory*, *16*(7), 712–727. https://doi.org/10.1080/09658210802220054

Freyd, J. J., & Gleaves, D. H. (1996). "Remembering" words not presented in lists: Relevance to the current recovered/false memory controversy. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*(3), 811–813. https://doi.org/10.1037/0278-7393.22.3.811

Gallo, D. A. (2010). False memories and fantastic beliefs: 15 years of the DRM illusion. *Memory & Cognition*, *38*(7), 833–848. https://doi.org/10.3758/MC.38.7.833

Gallo, D. A., Foster, K. T., & Johnson, E. L. (2009). Elevated false recollection of emotional pictures in young and older adults. *Psychology and Aging*, *24*(4), 981–988. https://doi.org/10.1037/a0017545

Gallo, D. A., & Roediger, H. L. (2003). The effects of associations and aging on illusory recollection. *Memory & Cognition*, *31*(7), 1036–1044. https://doi.org/10.3758/BF03196124

Garg, A. X., Hackam, D., & Tonelli, M. (2008). Systematic review and meta-analysis: When one study is just not enough. *Clinical Journal of the American Society of Nephrology*, *3*(1), 253–260. https://doi.org/10.2215/CJN.01430307

Gomes, C. F. A., Brainerd, C. J., & Stein, L. M. (2013). Effects of emotional valence and arousal on recollective and nonrecollective recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(3), 663–677. https://doi.org/10.1037/a0028578

Griffin, N. R., & Schnyer, D. M. (2020). Memory distortion for orthographically associated words in individuals with depressive symptoms. *Cognition*, *203*, Article 104330. https://doi.org/10.1016/j.cognition.2020.104330

Harper, N. R. (2017). *The relationship between worry symptoms of generalized anxiety disorder and true memory, false memory, and metamemory* (Publication No. 10271999) [Doctoral dissertation, Southern Illinois University]. ProQuest Dissertations and Theses Global.

Hauschildt, M., Peters, M. J., Jelinek, L., & Moritz, S. (2012). Veridical and false memory for scenic material in posttraumatic stress disorder. *Consciousness and Cognition*, *21*(1), 80–89. https://doi.org/10.1016/j.concog.2011.10.013

Hellenthal, M. V., Knott, L. M., Howe, M. L., Wilkinson, S., & Shah, D. (2019). The effects of arousal and attention on emotional false memory formation. *Journal of Memory and Language*, *107*, 54–68. https://doi.org/10.1016/j.jml.2019.03.010

Higgins, J. P., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine*, *21*(11), 1539–1558. https://doi.org/10.1002/sim.1186

Houben, S. T., Otgaar, H., Roelofs, J., Smeets, T., & Merckelbach, H. (2020). Increases of correct memories and spontaneous false memories due to eye movements when memories are retrieved after a time delay. *Behaviour Research and Therapy*, *125*, Article 103546. https://doi.org/10.1016/j.brat.2019.103546

Howe, M. L. (2007). Children's emotional false memories. *Psychological Science*, *18*(10), 856–860. https://doi.org/10.1111/j.1467-9280.2007.01991.x

Howe, M. L., Candel, I., Otgaar, H., Malone, C., & Wimmer, M. C. (2010). Valence and the development of immediate and long-term false memory illusions. *Memory*, *18*(1), 58–75. https://doi.org/10.1080/09658210903476514

Howe, M. L., Cicchetti, D., Toth, S. L., & Cerrito, B. M. (2004). True and false memories in maltreated children. *Child Development*, *75*(5), 1402–1417. https://doi.org/10.1111/j.1467-8624.2004.00748.x

Howe, M. L., & Knott, L. M. (2015). The fallibility of memory in judicial processes: Lessons from the past and their modern consequences. *Memory*, *23*(5), 633–656. https://doi.org/10.1080/09658211.2015.1010709

Howe, M. L., & Malone, C. (2011). Mood-congruent true and false memory: Effects of depression. *Memory*, *19*(2), 192–201. https://doi.org/10.1080/09658211.2010.544073

Howe, M. L., Toth, S. L., & Cicchetti, D. (2011). Can maltreated children inhibit true and false memories for emotional information? *Child Development*, *82*(3), 967–981. https://doi.org/10.1111/j.1467-8624.2011.01585.x

Howe, M. L., Wimmer, M. C., Gagnon, N., & Plumpton, S. (2009). An associative-activation theory of children's and adults' memory illusions. *Journal of Memory and Language*, *60*(2), 229–251. https://doi.org/10.1016/j.jml.2008.10.002

Huff, M. J., McNabb, J., & Hutchison, K. A. (2015). List blocking and longer retention intervals reveal an influence of gist processing for lexically ambiguous critical lures. *Memory & Cognition*, *43*(8), 1193–1207. https://doi.org/10.3758/s13421-015-0533-3

Ioannidis, J. P., & Trikalinos, T. A. (2007). The appropriateness of asymmetry tests for publication bias in meta-analyses: A large survey. *Canadian Medical Association Journal*, *176*(8), 1091–1096. https://doi.org/10.1503/cmaj.060410

Irwanda, D. Y., & Maulina, D. (2018, September 13–14). *False memory in traffic accident context: The effect of word types and gender* [Paper presentation]. Proceedings of the 2nd International Conference on Intervention and Applied Psychology (ICIAP 2018), Depok, West Java, Indonesia.

Jarosz, A. F., & Wiley, J. (2014). What are the odds? A practical guide to computing and reporting Bayes factors. *The Journal of Problem Solving*, *7*(1), 2–9. https://doi.org/10.7771/1932-6246.1167

Jelinek, L., Hottenrott, B., Randjbar, S., Peters, M. J., & Moritz, S. (2009). Visual false memories in post-traumatic stress disorder (PTSD). *Journal of Behavior Therapy and Experimental Psychiatry*, *40*(2), 374–383. https://doi.org/10.1016/j.jbtep.2009.02.003

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

Joormann, J., Teachman, B. A., & Gotlib, I. H. (2009). Sadder and less accurate? False memory for negative material in depression. *Journal of Abnormal Psychology*, *118*(2), 412–417. https://doi.org/10.1037/a0015621

Kensinger, E. A. (2004). Remembering and emotional experiences: The contribution of valence and arousal. *Reviews in the Neurosciences*, *15*(4), 241–251. https://doi.org/10.1515/REVNEURO.2004.15.4.241

Knott, L. M., Howe, M. L., Toffalini, E., Shah, D., & Humphreys, L. (2018). The role of attention in immediate emotional false memory enhancement. *Emotion*, *18*(8), 1063–1077. https://doi.org/10.1037/emo0000407

Knott, L. M., & Shah, D. (2019). The effect of limited attention and delay on negative arousing false memories. *Cognition and Emotion*, *33*(7), 1472–1480. https://doi.org/10.1080/02699931.2018.1556153

Knott, L. M., & Thorley, C. (2014). Mood-congruent false memories persist over time. *Cognition and Emotion*, *28*(5), 903–912. https://doi.org/10.1080/02699931.2013.860016

Knott, L. M., Threadgold, E., & Howe, M. L. (2014). Negative mood state impairs false memory priming when problem-solving. *Journal of Cognitive Psychology*, *26*(5), 580–587. https://doi.org/10.1080/20445911.2014.922091

Kvarven, A., Strømland, E., & Johannesson, M. (2020). Comparing meta-analyses and preregistered multiple-laboratory replication projects. *Nature Human Behaviour*, *4*(4), 423–434. https://doi.org/10.1038/s41562-019-0787-z

Laws, K. R. (2016). Psychology, replication & beyond. *BMC Psychology*, *4*(1), Article 30. https://doi.org/10.1186/s40359-016-0135-2

Lewis, M., Mathur, M. B., Van der Weele, T. J., & Frank, M. C. (2022). The puzzling relationship between multi-laboratory replications and meta-analyses of the published literature. *Royal Society Open Science*, *9*(2), Article 211499. https://doi.org/10.1098/rsos.211499

Lo, J. C., Sim, S. K., & Chee, M. W. (2014). Sleep reduces false memory in healthy older adults. *Sleep*, *37*(4), 665–671. https://doi.org/10.5665/sleep.3564

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

Loftus, E. F., & Pickrell, J. E. (1995). The formation of false memories. *Psychiatric Annals*, *25*(12), 720–725. https://doi.org/10.3928/0048-5713-19951201-07

Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: A user's guide*. Cambridge University Press.

Maulina, D., Irwanda, D. Y., Sekarmewangi, T. H., Putri, K. M. H., & Otgaar, H. (2021). How accurate are memories of traffic accidents? Increased false memory levels among motorcyclists when confronted with accident-related word lists. *Transportation Research Part F: Traffic Psychology and Behaviour*, *80*, 275–294. https://doi.org/10.1016/j.trf.2021.04.015

McDermott, K. B. (1996). The persistence of false memories in list recall. *Journal of Memory and Language*, *35*(2), 212–230. https://doi.org/10.1006/jmla.1996.0012

McHugh, M. L. (2012). Interrater reliability: The kappa statistic. *Biochemia Medica*, *22*(3), 276–282. https://doi.org/10.11613/BM.2012.031

McKeon, S., Pace-Schott, E. F., & Spencer, R. M. (2012). Interaction of sleep and emotional content on the production of false memories. *PLoS ONE*, *7*(11), Article e49353. https://doi.org/10.1371/journal.pone.0049353

Meeks, J. T., Taul, M. L., Rice, R. A., Posey, Z. W., & Harper, N. R. (2019). Negative mood reduces negative false memories after a brief mindfulness exercise. *Mindfulness*, *10*(12), 2507–2521. https://doi.org/10.1007/s12671-019-01223-6

Merckelbach, H., Muris, P., Schmidt, H., Rassin, E., & Horselenberg, R. (1998). De creatieve ervaringen vragenlijst als maat voor "fantasy proneness" [The creative experiences questionnaire (CEQ) as a measure of "fantasy proneness"]. *De Psycholoog*, *33*(5), 204–208.

Meusel, L. A., Macqueen, G. M., Jaswal, G., & McKinnon, M. C. (2012). Youth are more vulnerable to false memories than middle-aged adults due to liberal response bias. *Journal of the Canadian Academy of Child and Adolescent Psychiatry*, *21*(4), 289–295. https://pubmed.ncbi.nlm.nih.gov/23133463/

Miano, A., Schulze, K., Moritz, S., Wingenfeld, K., & Roepke, S. (2022). False memory in posttraumatic stress disorder and borderline personality disorder. *Psychiatry Research*, *314*, Article 114547. https://doi.org/10.1016/j.psychres.2022.114547

Monds, L. A., Paterson, H. M., & Kemp, R. I. (2017). Do emotional stimuli enhance or impede recall relative to neutral stimuli? An investigation of two "false memory" tasks. *Memory*, *25*(8), 945–952. https://doi.org/10.1080/09658211.2016.1237653

Monds, L. A., Paterson, H. M., Kemp, R. I., & Bryant, R. A. (2013). Individual differences in susceptibility to false memories for neutral and trauma-related words. *Psychiatry, Psychology and Law*, *20*(3), 399–411. https://doi.org/10.1080/13218719.2012.692932

Moritz, S., Woodward, T. S., Cuttler, C., Whitman, J. C., & Watson, J. M. (2004). False memories in schizophrenia. *Neuropsychology*, *18*(2), 276–283. https://doi.org/10.1037/0894-4105.18.2.276

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (2004). The University of South Florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, & Computers*, *36*(3), 402–407. https://doi.org/10.3758/BF03195588

Neuschatz, J. S., Benoit, G. E., & Payne, D. G. (2003). Effective warnings in the Deese–Roediger–McDermott false-memory paradigm: The role of identifiability. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*(1), 35–41. https://doi.org/10.1037/0278-7393.29.1.35

Newbury, C. R. (2019). *The role of sleep in memory consolidation: Effects of lateralisation and emotion* (Publication No. 28278014) [Doctoral dissertation, Lancaster University]. ProQuest Dissertations and Theses Global.

Nie, A., Li, M., Li, M., Xiao, Y., & Wang, S. (2022). Together we lose or gain: Ongoing and enduring impacts of collaboration in episodic memory of emotional DRM lists. *Current Psychology*, *42*(32), 27965–27982. https://doi.org/10.1007/s12144-022-03940-z

Nieuwenstein, M. R., Wierenga, T., Morey, R. D., Wicherts, J. M., Blom, T. N., Wagenmakers, E. J., & van Rijn, H. (2015). On making the right choice: A meta-analysis and large-scale replication attempt of the unconscious thought advantage. *Judgment and Decision Making*, *10*(1), 1–17. https://doi.org/10.1017/S1930297500003144

Norris, C. J., Leaf, P. T., & Fenn, K. M. (2019). Negativity bias in false memory: Moderation by neuroticism after a delay. *Cognition and Emotion*, *33*(4), 737–753. https://doi.org/10.1080/02699931.2018.1496068

Nosek, B. A., & Errington, T. M. (2020). What is replication? *PLoS Biology*, *18*(3), Article e3000691. https://doi.org/10.1371/journal.pbio.3000691

Nuijten, M. B., van Assen, M. A. L. M., Hartgerink, C. H. J., Epskamp, S., & Wicherts, J. M. (2017, November 16). *The validity of the tool "statcheck"*

in discovering statistical reporting inconsistencies. https://doi.org/10.31234/osf.io/tcxaj

Oliveira, H. M., Albuquerque, P. B., & Saraiva, M. (2019). Associative strength or gist extraction: Which matters when DRM lists have two critical lures? *Quarterly Journal of Experimental Psychology*, *72*(3), 570–578. https://doi.org/10.1177/1747021818761002

Otgaar, H., Alberts, H., & Cuppens, L. (2012). Ego depletion results in an increase in spontaneous false memories. *Consciousness and Cognition*, *21*(4), 1673–1680. https://doi.org/10.1016/j.concog.2012.09.006

Otgaar, H., Candel, I., & Merckelbach, H. (2008). Children's false memories: Easier to elicit for a negative than for a neutral event. *Acta Psychologica*, *128*(2), 350–354. https://doi.org/10.1016/j.actpsy.2008.03.009

Otgaar, H., Howe, M. L., Brackmann, N., & Smeets, T. (2016). The malleability of developmental trends in neutral and negative memory illusions. *Journal of Experimental Psychology: General*, *145*(1), 31–55. https://doi.org/10.1037/xge0000127

Otgaar, H., Howe, M. L., Mangiulli, I., & Bücken, C. (2020). The impact of false denials on forgetting and false memory. *Cognition*, *202*, Article 104322. https://doi.org/10.1016/j.cognition.2020.104322

Otgaar, H., Howe, M. L., & Muris, P. (2017). Maltreatment increases spontaneous false memories but decreases suggestion-induced false memories in children. *British Journal of Developmental Psychology*, *35*(3), 376–391. https://doi.org/10.1111/bjdp.12177

Otgaar, H., Howe, M. L., Peters, M., Sauerland, M., & Raymaekers, L. (2013). Developmental trends in different types of spontaneous false memories: Implications for the legal field. *Behavioral Sciences & the Law*, *31*(5), 666–682. https://doi.org/10.1002/bsl.2076

Otgaar, H., Moldoveanu, G., Wang, J., & Howe, M. L. (2017). Exploring the consequences of nonbelieved memories in the DRM paradigm. *Memory*, *25*(7), 922–933. https://doi.org/10.1080/09658211.2016.1272701

Otgaar, H., Peters, M., & Howe, M. L. (2012). Dividing attention lowers children's but increases adults' false memories. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(1), 204–210. https://doi.org/10.1037/a0025160

Ouzzani, M., Hammady, H., Fedorowicz, Z., & Elmagarmid, A. (2016). Rayyan—A web and mobile app for systematic reviews. *Systematic Reviews*, *5*(1), Article 210. https://doi.org/10.1186/s13643-016-0384-4

Palmer, J. E., & Dodson, C. S. (2009). Investigating the mechanisms fuelling reduced false recall of emotional material. *Cognition & Emotion*, *23*(2), 238–259. https://doi.org/10.1080/02699930801976663

Pardilla-Delgado, E., & Payne, J. D. (2017). The Deese–Roediger–McDermott (DRM) task: A simple cognitive paradigm to investigate false memories in the laboratory. *Journal of Visualized Experiments*, *119*(119), Article e54793. https://doi.org/10.3791/54793

Pezdek, K., & Lam, S. (2007). What research paradigms have cognitive psychologists used to study "false memory," and what are the implications of these choices? *Consciousness and Cognition*, *16*(1), 2–17. https://doi.org/10.1016/j.concog.2005.06.006

Piguet, O., Connally, E., Krendl, A. C., Huot, J. R., & Corkin, S. (2008). False memory in aging: Effects of emotional valence on word recognition accuracy. *Psychology and Aging*, *23*(2), 307–314. https://doi.org/10.1037/0882-7974.23.2.307

Pittelkow, M.-M., Hoekstra, R., Karsten, J., & van Ravenzwaaij, D. (2021). Replication target selection in clinical psychology: A Bayesian and qualitative reevaluation. *Clinical Psychology: Science and Practice*, *28*(2), 210–221. https://doi.org/10.1037/cps0000013

Porter, S., Bellhouse, S., McDougall, A., ten Brinke, L., & Wilson, K. (2010). A prospective investigation of the vulnerability of memory for positive and negative emotional scenes to the misinformation effect. *Canadian Journal of Behavioural Science*, *42*(1), 55–61. https://doi.org/10.1037/a0016652

Porter, S., Spencer, L., & Birt, A. R. (2003). Blinded by emotion? Effect of the emotionality of a scene on susceptibility to false memories. *Canadian Journal of Behavioural Science*, *35*(3), 165–175. https://doi.org/10.1037/h0087198

Porter, S., Taylor, K., & ten Brinke, L. (2008). Memory for media: Investigation of false memories for negatively and positively charged public events. *Memory*, *16*(6), 658–666. https://doi.org/10.1080/09658210802154626

Porter, S., ten Brinke, L., Riley, S. N., & Baker, A. (2014). Prime time news: The influence of primed positive and negative emotion on susceptibility to false memories. *Cognition and Emotion*, *28*(8), 1422–1434. https://doi.org/10.1080/02699931.2014.887000

Quas, J. A., Rush, E. B., Yim, I. S., Edelstein, R. S., Otgaar, H., & Smeets, T. (2016). Stress and emotional valence effects on children's versus adolescents' true and false memory. *Memory*, *24*(5), 696–707. https://doi.org/10.1080/09658211.2015.1045909

Reyna, V. F., & Brainerd, C. J. (1995). Fuzzy-trace theory: An interim synthesis. *Learning and Individual Differences*, *7*(1), 1–75. https://doi.org/10.1016/1041-6080(95)90031-4

Riesthuis, P., Mangiulli, I., Bogaard, G., & Otgaar, H. (2022). The impact of fabrication on recognition memory: An experimental study. *New Ideas in Psychology*, *67*, Article 100966. https://doi.org/10.1016/j.newideapsych.2022.100966

Rife, S. C., Nuijten, M. B., Epskamp, S. (2016). *statcheck: Extract statistics from articles and recompute p-values* [Web application]. http://statcheck.io

Rodríguez-Ferreiro, J., Martínez, C., & Cuetos, F. (2019). Differential effects of negative and positive emotional content over veridical and false recognition in aging and Alzheimer's disease. *Journal of Neurolinguistics*, *49*, 109–116. https://doi.org/10.1016/j.jneuroling.2018.10.001

Roediger, H. L., & McDermott, K. B. (1995). Creating false memories: Remembering words not presented in lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*(4), 803–814. https://doi.org/10.1037/0278-7393.21.4.803

Roediger, H. L., & McDermott, K. B. (2000). Tricks of memory. *Current Directions in Psychological Science*, *9*(4), 123–127. https://doi.org/10.1111/1467-8721.00075

Roediger, H. L., Watson, J. M., McDermott, K. B., & Gallo, D. A. (2001). Factors that determine false recall: A multiple regression analysis. *Psychonomic Bulletin & Review*, *8*(3), 385–407. https://doi.org/10.3758/BF03196177

Rohatgi, A. (2021). *Webplotdigitizer: Version 4.5* [Web application]. https://automeris.io/WebPlotDigitizer

Ruci, L., Tomes, J. L., & Zelenski, J. M. (2009). Mood-congruent false memories in the DRM paradigm. *Cognition and Emotion*, *23*(6), 1153–1165. https://doi.org/10.1080/02699930802355420

Schmidt, F. L., & Hunter, J. E. (1996). Measurement error in psychological research: Lessons from 26 research scenarios. *Psychological Methods*, *1*(2), 199–223. https://doi.org/10.1037/1082-989X.1.2.199

Schooler, J. (2011). Unpublished results hide the decline effect. *Nature*, *470*(7335), Article 437. https://doi.org/10.1038/470437a

Seamon, J. G., Luo, C. R., Kopecky, J. J., Price, C. A., Rothschild, L., Fung, N. S., & Schwartz, M. A. (2002). Are false memories more difficult to forget than accurate memories? The effect of retention interval on recall and recognition. *Memory & Cognition*, *30*(7), 1054–1064. https://doi.org/10.3758/BF03194323

Shah, D., & Knott, L. M. (2018). The role of attention at retrieval on the false recognition of negative emotional DRM lists. *Memory*, *26*(2), 269–276. https://doi.org/10.1080/09658211.2017.1349803

Sharma, P. R., Wade, K. A., & Jobson, L. (2023). A systematic review of the relationship between emotion and susceptibility to misinformation. *Memory*, *31*(1), 1–21. https://doi.org/10.1080/09658211.2022.2120623

Simons, D. J. (2014). The value of direct replication. *Perspectives on Psychological Science*, *9*(1), 76–80. https://doi.org/10.1177/1745691613514755

Smeets, T., Otgaar, H., Candel, I., & Wolf, O. T. (2008). True or false? Memory is differentially affected by stress-induced cortisol elevations and sympathetic activity at consolidation and retrieval. *Psychoneuroendocrinology*, *33*(10), 1378–1386. https://doi.org/10.1016/j.psyneuen.2008.07.009

Snodgrass, J. G., & Corwin, J. (1988). Pragmatics of measuring recognition memory: Applications to dementia and amnesia. *Journal of Experimental*

Psychology: General, *117*(1), 34–50. https://doi.org/10.1037/0096-3445.117.1.34

Stanislaw, H., & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior Research Methods, Instruments, & Computers*, *31*(1), 137–149. https://doi.org/10.3758/BF03207704

Stanley, T. D., Jarrell, S. B., & Doucouliagos, H. (2010). Could it be better to discard 90% of the data? A statistical paradox. *The American Statistician*, *64*(1), 70–77. https://doi.org/10.1198/tast.2009.08205

Stea, J. N., Lee, S. M., & Sears, C. R. (2013). Enhancement of false memory for negative material in dysphoria: Mood congruency or response bias? *Cognitive Therapy and Research*, *37*(6), 1189–1200. https://doi.org/10.1007/s10608-013-9557-9

Storbeck, J., & Clore, G. L. (2005). With sadness comes accuracy; with happiness, false memory: Mood and the false memory effect. *Psychological Science*, *16*(10), 785–791. https://doi.org/10.1111/j.1467-9280.2005.01615.x

Talmi, D. (2013). Enhanced emotional memory: Cognitive and neural mechanisms. *Current Directions in Psychological Science*, *22*(6), 430–436. https://doi.org/10.1177/0963721413498893

Thijssen, J., Otgaar, H., Howe, M. L., & de Ruiter, C. (2013). Emotional true and false memories in children with callous-unemotional traits. *Cognition & Emotion*, *27*(4), 761–768. https://doi.org/10.1080/02699931.2012.744300

Thornton, A., & Lee, P. (2000). Publication bias in meta-analysis: Its causes and consequences. *Journal of Clinical Epidemiology*, *53*(2), 207–216. https://doi.org/10.1016/S0895-4356(99)00161-4

van Damme, I. (2013). Mood and the DRM paradigm: An investigation of the effects of valence and arousal on false memory. *Quarterly Journal of Experimental Psychology*, *66*(6), 1060–1081. https://doi.org/10.1080/17470218.2012.727837

Van Damme, I., & Smets, K. (2014). The power of emotion versus the power of suggestion: Memory for emotional events in the misinformation paradigm. *Emotion*, *14*(2), 310–320. https://doi.org/10.1037/a0034629

Van Elk, M., Matzke, D., Gronau, Q., Guang, M., Vandekerckhove, J., & Wagenmakers, E.-J. (2015). Meta-analyses are no substitute for registered replications: A skeptical perspective on religious priming. *Frontiers in Psychology*, *6*, Article 1365. https://doi.org/10.3389/fpsyg.2015.01365

Vannucci, M., Nocentini, A., Mazzoni, G., & Menesini, E. (2012). Recalling unpresented hostile words: False memories predictors of traditional and cyberbullying. *European Journal of Developmental Psychology*, *9*(2), 182–194. https://doi.org/10.1080/17405629.2011.646459

Velsor, S. F. (2016). *The roles of emotion regulation and working memory in the relationship between depressive symptoms and false memory for negative information* (Publication No. 10128862) [Master's thesis, Southern Illinois University]. ProQuest Dissertations and Theses Global.

Vevea, J. L., & Woods, C. M. (2005). Publication bias in research synthesis: Sensitivity analysis using a priori weight functions. *Psychological Methods*, *10*(4), 428–443. https://doi.org/10.1037/1082-989X.10.4.428

Viechtbauer, W. (2005). Bias and efficiency of meta-analytic variance estimators in the random-effects model. *Journal of Educational and Behavioral Statistics*, *30*(3), 261–293. https://doi.org/10.3102/10769986030003261

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, *36*(3), 1–48. https://doi.org/10.18637/jss.v036.i03

Vohs, K. D., Schmeichel, B. J., Lohmann, S., Gronau, Q. F., Finley, A. J., Ainsworth, S. E., Alquist, J. L., Baker, M. D., Brizi, A., Bunyi, A., Butschek, G. J., Campbell, C., Capaldi, J., Cau, C., Chambers, H., Chatzisarantis, N. L. D., Christensen, W. J., Clay, S. L., Curtis, J., … Albarracín, D. (2021). A multi-site preregistered paradigmatic test of the ego depletion effect. *Psychological Science*, *32*(10), 1566–1581. https://doi.org/10.1177/0956797621989733

Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods*, *45*(4), 1191–1207. https://doi.org/10.3758/s13428-012-0314-x

Watson, J. M., Bunting, M. F., Poole, B. J., & Conway, A. R. A. (2005). Individual differences in susceptibility to false memory in the Deese–Roediger–McDermott paradigm. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(1), 76–85. https://doi.org/10.1037/0278-7393.31.1.76

Welliver, J. M. (2015). *Effects of weapon focus on true and false recall of emotional words* (Publication No. 1593271) [Master's thesis, University of Louisiana]. ProQuest Dissertations and Theses Global.

Wells, G. L., & Windschitl, P. D. (1999). Stimulus sampling and social psychological experimentation. *Personality and Social Psychology Bulletin*, *25*(9), 1115–1125. https://doi.org/10.1177/01461672992512005

White, C. N., Kapucu, A., Bruno, D., Rotello, C. M., & Ratcliff, R. (2014). Memory bias for negative emotional words in recognition memory is driven by effects of category membership. *Cognition and Emotion*, *28*(5), 867–880. https://doi.org/10.1080/02699931.2013.858028

Wixted, J. T., & Stretch, V. (2000). The case against a criterion-shift account of false memory. *Psychological Review*, *107*(2), 368–376. https://doi.org/10.1037/0033-295X.107.2.368

Wright, D. B., Startup, H. M., & Mathews, S. A. (2005). Mood, dissociation and false memories using the Deese–Roediger–McDermott procedure. *British Journal of Psychology*, *96*(3), 283–293. https://doi.org/10.1348/000712605X49916

Yablonski, A. J. (2016). *A novel use of the Deese–Roediger–McDermott paradigm: Distinguishing between differential memory mechanisms in emotional literature [Senior Honors Project, James Madison University]*. JMU Scholarly Commons.

Yonelinas, A. P. (2002). The nature of recollection and familiarity: A review of 30 years of research. *Journal of Memory and Language*, *46*(3), 441–517. https://doi.org/10.1006/jmla.2002.2864

Yüvrük, E., & Kapucu, A. (2022). False (or biased) memory: Emotion and working memory capacity effects in the DRM paradigm. *Memory & Cognition*, *50*(7), 1443–1463. https://doi.org/10.3758/s13421-022-01298-y

Yuvruk, E., Turan, H., & Kapucu, A. (2019). Development of Turkish DRM lists with emotional words. *Studies in Psychology*, *39*(2), 245–266. https://doi.org/10.26650/SP2019-0026

Zhang, W. (2017). *The effect of emotion on false memories in the Deese-Roediger-McDermott (DRM) paradigm* [Doctoral dissertation, University of Otago]. University of Otago Archive.

Zhang, W., Gross, J., & Hayne, H. (2017). The effect of mood on false memory for emotional DRM word lists. *Cognition and Emotion*, *31*(3), 526–537. https://doi.org/10.1080/02699931.2016.1138930

Zhang, W., Gross, J., & Hayne, H. (2018). If you're happy and you know it: Positive moods reduce age-related differences in false memory. *Child Development*, *89*(4), e332–e341. https://doi.org/10.1111/cdev.12890

Zhang, W., Gross, J., & Hayne, H. (2019). Mood impedes monitoring of emotional false memories: Evidence for the associative theories. *Memory*, *27*(2), 198–208. https://doi.org/10.1080/09658211.2018.1498107

Zhang, W., Gross, J., & Hayne, H. (2021). An age-related positivity effect in semantic true memory but not false memory. *Emotion*, *21*(3), 526–535. https://doi.org/10.1037/emo0000715

Zhong, Y. P., Zhang, W. J., Li, Y. L., & Fan, W. (2018). Influence of time stress on mood-congruent false memories. *Acta Psychologica Sinica*, *50*(9), 929–939. https://doi.org/10.3724/SP.J.1041.2018.00929

Zwaan, R. A., Pecher, D., Paolacci, G., Bouwmeester, S., Verkoeijen, P., Dijkstra, K., & Zeelenberg, R. (2018). Participant nonnaiveté and the reproducibility of cognitive psychology. *Psychonomic Bulletin & Review*, *25*(5), 1968–1972. https://doi.org/10.3758/s13423-017-1348-y

(*Appendices follow*)

Appendix A

Search Terms for the Meta-Analysis

Ovid, APA PsycInfo, 1806 to February Week 4, 2021

1. (Deese ADJ2 McDermot* OR DRM).ti,ab,id.
2. emotional states/ or emotions/ or affective valence/ or negative emotions/ or positive emotions/ OR (Mood OR Affect* OR negativ* OR emotion* OR positiv* OR arousal OR valence OR feel*).ti,ab,id.
3. 1 AND 2

The search terms for Medline were:

Ovid MEDLINE, including e-publications ahead of print, in-process & other nonindexed citations and Ovid MEDLINE Daily, 1946 to February 26, 2021

1. ((Deese ADJ2 McDermot*) OR (DRM ADJ3 (task OR paradigm* OR list* OR procedure))).ti,ab,kf.
2. emotions/ or affect/ OR (Mood OR Affect* OR negativ* OR emotion* OR positiv* OR arousal OR valence OR feel*).ti,ab,kf.
3. 1 AND 2

The search terms for PsycInfo were:

Ovid, APA PsycInfo, 1806 to February Week 4, 2021

1. (Deese ADJ2 McDermot* OR DRM).ti,ab,id.
2. emotional states/ or emotions/ or affective valence/ or negative emotions/ or positive emotions/ OR (Mood OR Affect* OR negativ* OR emotion* OR positiv* OR arousal OR valence OR feel*).ti,ab,id.
3. 1 AND 2

For Google Scholar, the following search terms were used:

"Deese/Roediger-McDermott"|"Deese-Roediger-McDermott"|"Deese Roediger McDermott"|"Deese/Rödiger-McDermott" & Mood|Affect|negative|emotion|positive|arousal|valence|feelings.

Due to the decreased report volume of the OSF server and ProQuest, only the following terms were used.

OSF server:

(Roediger AND McDermot*) OR DRM

ProQuest:

Roediger AND McDermot* AND (negativ* OR valenc* OR arous*)

Appendix B

Recall Test Including Zhang (2017)

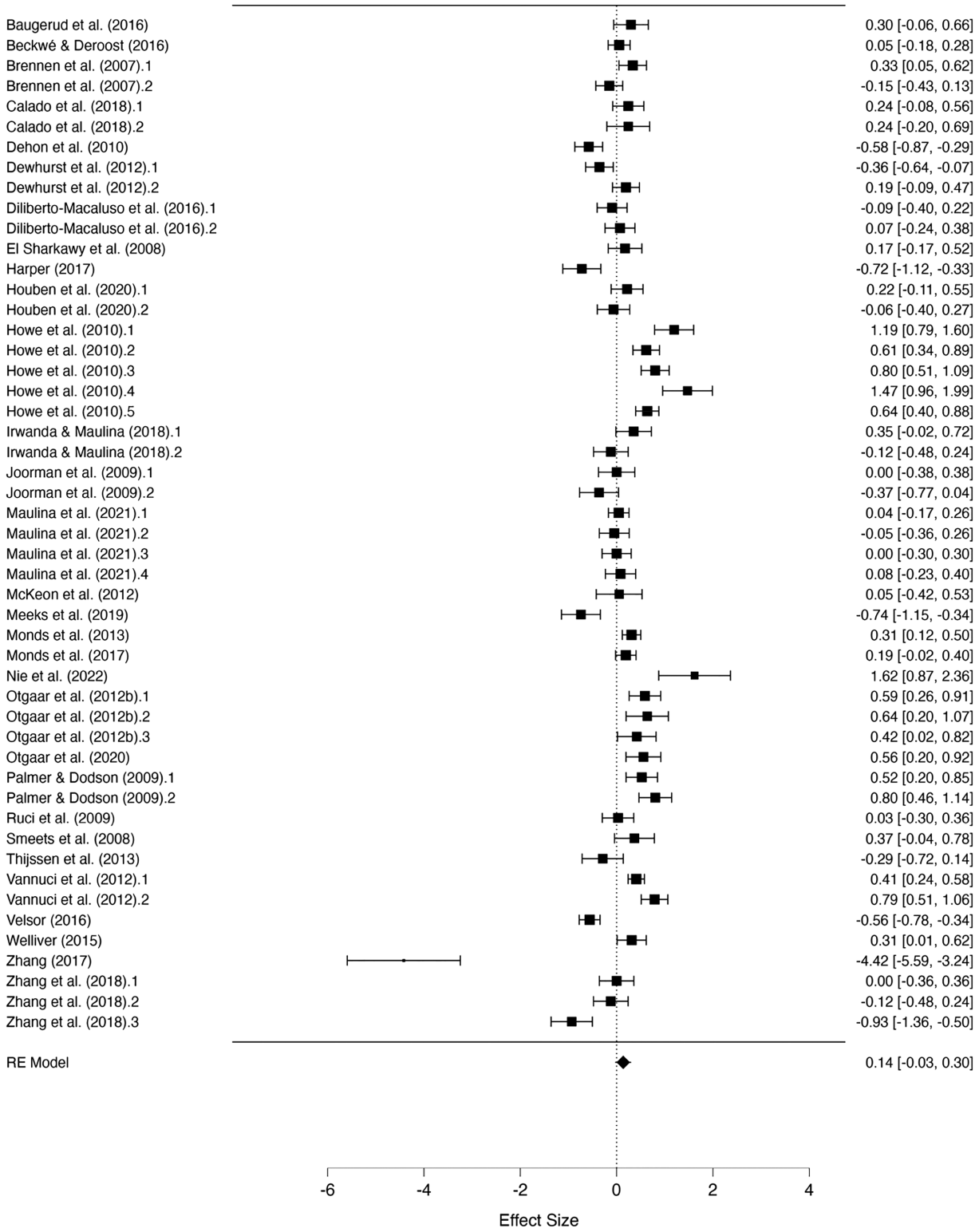
Overall Effect Size

We computed the average standardized paired difference across all extracted effect sizes ($k = 50$, $n = 2{,}239$). For the frequentist analysis, the results revealed a nonsignificant effect size, $d = 0.14$, 95% CI [−0.03, 0.30], $p = .104$. The effect sizes of the single studies and the average effect size across all studies are provided in Figure B1. Additionally, the Bayesian model-averaged meta-analysis showed inconclusive evidence ($BF_{01} = 2.85$) and, thus, did not clearly evidence one hypothesis over the other.

(*Appendices continue*)

Figure B1

Forest Plot for the Recall Test Outcome including Zhang (2017)



Note. RE = Random-Effects.

(*Appendices continue*)

Appendix C

Imputed Correlation Sensitivity Analyses Overview on the Recognition Outcome Controlling for Potential Response Bias

Table C1

Effect Size, p-Value, and Bayes Factor per Imputed Correlation Reanalysis

Imputed correlation	Cohen's *d*	*p*	Bayes factor (BF_{01})
1. Average, 2. Average, 3. 1 *SD* below	−0.05	.400	9.80
1. Average, 2. Average, 3. 1 *SD* above	−0.05	.475	9.09
1. Average, 2. 1 *SD* below, 3. Average	−0.06	.410	9.43
1. Average, 2. 1 *SD* below, 3. 1 *SD* below	−0.06	.379	9.71
1. Average, 2. 1 *SD* below, 3. 1 *SD* above	−0.06	.451	9.09
1. Average, 2. 1 *SD* above, 3. Average	−0.05	.456	9.62
1. Average, 2. 1 *SD* above, 3. 1 *SD* below	−0.05	.423	10.00
1. Average, 2. 1 *SD* above, 3. 1 *SD* above	−0.05	.500	9.17
1. 1 *SD* below 2. Average, 3. Average	−0.06	.418	9.52
1. 1 *SD* below 2. Average, 3. 1 *SD* below	−0.05	.387	9.71
1. 1 *SD* below 2. Average, 3. 1 *SD* above	−0.06	.460	9.09
1. 1 *SD* below 2. 1 *SD* below, 3. Average	−0.06	.398	9.43
1. 1 *SD* below 2. 1 *SD* below, 3. 1 *SD* below	−0.06	.368	9.62
1. 1 *SD* below 2. 1 *SD* below, 3. 1 *SD* above	−0.06	.439	9.01
1. 1 *SD* below 2. 1 *SD* above, 3. Average	−0.05	.440	9.62
1. 1 *SD* below 2. 1 *SD* above, 3. 1 *SD* below	−0.05	.409	9.90
1. 1 *SD* below 2. 1 *SD* above, 3. 1 *SD* above	−0.05	.482	9.17
1. 1 *SD* above 2. Average, 3. Average	−0.05	.446	9.52
1. 1 *SD* above 2. Average, 3. 1 *SD* below	−0.05	.414	9.90
1. 1 *SD* above 2. Average, 3. 1 *SD* above	−0.05	.490	9.17
1. 1 *SD* above 2. 1 *SD* below, 3. Average	−0.06	.422	9.43
1. 1 *SD* above 2. 1 *SD* below, 3. 1 *SD* below	−0.05	.391	9.71
1. 1 *SD* above 2. 1 *SD* below, 3. 1 *SD* above	−0.06	.463	9.09
1. 1 *SD* above 2. 1 *SD* above, 3. Average	−0.05	.474	9.62
1. 1 *SD* above 2. 1 *SD* above, 3. 1 *SD* below	−0.05	.440	10.00
1. 1 *SD* above 2. 1 *SD* above, 3. 1 *SD* above	−0.05	.520	9.17

Note. 1. Correlation between neutral critical and neutral unrelated; 2. Correlation between negative critical and negative unrelated; 3. Correlation of the difference between neutral critical and neutral unrelated, and negative critical and negative unrelated.

Appendix D

Exploratory Metaregression Analyses: The Effect of Matched Arousal

Prior literature suggests that high arousal in negative lists (vs. low) may further heighten the effect of negative valence on false memory formation (vs. neutral; e.g., Brainerd et al., 2010; Hellenthal et al., 2019). To tackle the limitations of the confirmatory arousal analysis that assessing the arousal level of negative critical lures may not give information about the differences in arousal between valence conditions, we conducted an exploratory analysis in which we compared findings of studies that explicitly reported having matched arousal versus those that did not. For this, all studies that did not match arousal levels or did not report having matched arousal levels were coded as 0 (i.e., not matched; 1 = matched). We used mixed-effects models with REML estimation for the amount of (residual) heterogeneity.

Results

For the recall outcome, we found a nonsignificant effect in the metaregression analysis when including matched arousal ($Z = -1.38$, $p = .169$). The average effect size for the difference between neutral and negative valence for studies that did not match arousal levels was .26, and .03 when arousal was matched. In recognition, similarly, the effect of arousal was not significant ($Z = 0.20$, $p = .845$) with an average effect size of −.24 for studies that did not match arousal levels, and −.23 when arousal was matched. Lastly, when looking at the effect of arousal on the recognition outcome controlling for potential response bias, we found a significant effect ($Z = -2.84$, $p = .005$). Here, the average effect size for the difference between neutral and negative valence for studies that did not match arousal levels was .08, and −.31 when arousal was matched. Hence, when controlling for potential response bias in recognition, negative valence (vs. neutral) was associated with more false memories when conditions were matched on arousal levels.

(*Appendices continue*)

Appendix E

Formulas for Calculating the Response Bias Index and the Discriminability Index

Response Bias Index

$$c'_{\text{studied–critical difference in neutral condition}} = -[z_{\text{hit rate, neutral}} + z_{\text{critical false alarm rate, neutral}}]/2$$

$$c'_{\text{studied–critical difference in negative condition}} = -[z_{\text{hit rate, negative}} + z_{\text{critical false alarm rate, negative}}]/2$$

$$c'_{\text{studied–unrelated difference in neutral condition}} = -[z_{\text{hit rate, neutral}} + z_{\text{unrelated false alarm rate, neutral}}]/2$$

$$c'_{\text{studied–unrelated difference in negative condition}} = -[z_{\text{hit rate, negative}} + z_{\text{unrelated false alarm rate, negative}}]/2$$

Discriminability Index

$$d'_{\text{studied–critical difference in neutral condition}} = z_{\text{hit rate, neutral}} - z_{\text{critical false alarm rate, neutral}}$$

$$d'_{\text{studied–critical difference in negative condition}} = z_{\text{hit rate, negative}} - z_{\text{critical false alarm rate, negative}}$$

$$d'_{\text{studied–unrelated difference in neutral condition}} = z_{\text{hit rate, neutral}} - z_{\text{unrelated false alarm rate, neutral}}$$

$$d'_{\text{studied–unrelated difference in negative condition}} = z_{\text{hit rate, negative}} - z_{\text{unrelated false alarm rate, negative}}$$

Received November 14, 2022
Revision received October 3, 2023
Accepted October 15, 2023 ■