



## RESEARCH ARTICLE

WILEY

# Expert opinions on the smallest effect size of interest in false memory research

Paul Riesthuis<sup>1,2</sup>  | Ivan Mangiulli<sup>1,2</sup> | Nick Broers<sup>2</sup> | Henry Otgaar<sup>1,2</sup> 

<sup>1</sup>Leuven Institute of Criminology, KU Leuven, Leuven, Belgium

<sup>2</sup>Forensic Psychology Faculty of Psychology and Neuroscience, Maastricht University, Maastricht, Netherlands

## Correspondence

Paul Riesthuis, Leuven Institute of Criminology, KU Leuven, Oude Markt 13, 3000, Leuven, Belgium.  
Email: paul.riesthuis@kuleuven.be

## Funding information

The current manuscript has been supported by a C1 and FWO Research Project grant awarded to the last author.

## Abstract

In the present study, we used a new approach to establish the smallest effect size of interest (SESOI) for false memory research by asking memory researchers what they considered to be the SESOI in false memory research. They were presented with three hypothetical and three influential paper scenarios. These scenarios depicted studies examining the effects of certain manipulations (e.g., therapy) on false memory formation using well-known false memory paradigms: Deese/Roediger-McDermott, misinformation, and forced fabrication. Subsequently, they were asked for each scenario what they would consider to be the SESOI for practical and theoretical purposes and justify their decisions. We found that there was no clear consensus for the SESOI. However, memory researchers tended to accept smaller SESOIs or “any difference that leads to a  $p < .05$ ,” especially for theoretical ends. We argue that the lack of a general consensus is acceptable as long as proper justification is used. We discuss such rationales and provide recommendations for setting the SESOI.

## KEYWORDS

expert opinion, false memory, smallest effect size of interest

## 1 | INTRODUCTION

Using an effect size (ES; magnitude of a phenomenon) has become increasingly important in psychological science as an informative statistic to plan and interpret studies (e.g., power analysis), conduct meta-analyses, corroborate theories, and gauge the real-world implications of an effect (Cohen, 1988; Lakens, 2013). The latter aspect is especially important in areas where the stakes are high. For example, and of importance, one such area concerns the field of false memory (remembrance of a non-experienced event/detail) where false memories can lead to false accusations and even miscarriages of justice (Howe & Knott, 2015).

Take for instance the case of Holly Ramona who sued her father after she claimed to have clear and vivid memories of her father sexually abusing her during her childhood (Ramona v. Ramona, 1997). Upon closer inspection, it became clear that Holly's memories were

most likely false because they were recollected during therapy through the use of suggestive therapeutic techniques (i.e., sodium amylal interview). Of importance is to examine which manipulations (e.g., therapy, drugs, etc.) can lead to increased false memory rates. However, when can the results of such studies make practical implications such as that it needs to be taken into account when assessing the reliability of statements as in the case of Ramona? For example, imagine a study wherein researchers examine the effects of suggestive interviewing tactics on false memory formation. When can the researchers conclude that suggestive interviewing tactics should not be used during interrogations because they increase the susceptibility to false memories? Is a statistically significant result sufficient to make practical implications such as advocating against suggestive interviewing tactics or is a certain minimum effect size of increased false memories necessary? In addition, does this minimum effect size of false memories differ when results are aimed at theoretical

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2021 The Authors. *Applied Cognitive Psychology* published by John Wiley & Sons Ltd.

advancement, for example understanding the underlying mechanisms of false memory formation, instead of practical implications? In the present study, we examined what memory researchers consider the smallest effect size of interest (SESOI; Lakens, 2014) for practical and theoretical purposes in false memory research.

## 2 | SMALLEST EFFECT SIZE OF INTEREST

The SESOI can be established based on the smallest effect that (i) researchers personally care about, (ii) is practically meaningful, or (iii) theoretically relevant (Lakens, 2014). The SESOI differs from effects that are simply statistically significant as results can be statistically significant while effects are trivial (Anvari & Lakens, 2021). Determining the SESOI for a particular study can be achieved in several ways such as using objective (e.g., anchors-based methods, minimally clinically important differences), or subjective argumentations (e.g., Cohen's benchmarks, related studies, resource based) depending on what the researcher deems suitable (Anvari & Lakens, 2021; Lakens et al., 2018). A frequently used approach in psychological science and thus also in the field of false memory are Cohen's benchmarks, namely small, medium, and large effects sizes (e.g., Cohen's  $d$  of .2, .5, .8; Cohen, 1988; Schäfer & Schwarz, 2019). Although oftentimes not specifically used to establish the SESOI, researchers frequently use Cohen's benchmarks to perform a priori power analyses to calculate the required sample size that allows them to detect a certain effect size given a certain statistical power ( $1-\beta$ ) and alpha level. However, as Cohen (1988, p. 25) argued "The terms 'small', 'medium', and 'large' are relative, not only to each other, but also to the area of behavioral science or even more particularly to the specific content and research method being employed in any given investigation."

Following Cohen's recommendation, Bosco et al. (2015) revised Cohen's benchmarks for several areas of research in psychology by examining the published literature, extracting the effect sizes, and empirically establishing benchmarks for small, medium, and large effects. The authors found that the benchmarks varied greatly across research domains and tended to be smaller in comparison with Cohen's benchmarks. Moreover, it is likely that the empirically established effect sizes were larger than they really were because they were affected by publication bias, leading to even lower estimates of the small, medium, and large effect sizes (Bosco et al., 2015; Carter et al., 2019). Although the results of Bosco et al. (2015) gave insights about the variety of effect sizes observed in the literature, they did not clarify whether such effects sizes bear any practical meaning (Anvari & Lakens, 2021).

Another method to establish which effect sizes yield practical meaningfulness is using anchor-based methods such as the minimal clinically important difference (MCID; Anvari & Lakens, 2021; McGlothlin & Lewis, 2014). Specifically, the MCID is based on the smallest effect a patient personally experiences as an improvement (or decline). However, in false memory research, it is difficult to use such anchor-based methods because people might be unaware that

they have a false memory (Bernstein & Loftus, 2009). Thus, in the current paper, we propose another way to establish the SESOI which is by asking experts what they consider to be SESOI and examine whether there is general agreement among them. Although anchor-based methods are more frequently adopted to set the MCID in fields where possible such as medical research, expert consensus has also been used successfully to define the MCID (Mouelhi et al., 2020; van der Heijde et al., 2001; Wells et al., 2001). Ideally, anchor-based and expert consensus methods are implemented simultaneously to accurately estimate the MCID (Bonini et al., 2020). However, as in the field of false memory research where anchor-based methods are not feasible, expert consensus seems to be appropriate to set the SESOI. Establishing such expert consensus regarding the SESOI for false memory research can help clarify practical and theoretical relevance of (future) studies, but also guide future research in their sample size justification.

## 3 | FALSE MEMORY AND THE SESOI

To examine the SESOI, and contextualize Cohen's benchmarks specifically for the field of false memory research, we looked at three frequently used paradigms known to elicit false memories: Deese/Roediger-McDermott (DRM; Deese, 1959; Roediger & McDermott, 1995), misinformation (Loftus et al., 1978), and the forced fabrication paradigm (Ackil & Zaragoza, 1998). These paradigms differ in their ecological validity and experimental control (Wade et al., 2007), but also tap into different types of false memories with distinct underlying mechanisms (Ost et al., 2013), resulting in possibly different SESOIs for each paradigm. In the DRM paradigm, participants study words (e.g., *piano, jazz, and note*) that are associatively related to a critical lure that was not presented (i.e., *music*). On a subsequent memory task, participants oftentimes erroneously report having studied the critical lure (Gallo, 2006). We selected the DRM paradigm because it is frequently used to examine false memories generated by spontaneous mental associations (without external pressure), better known as *spontaneous* false memories (Otgaar et al., 2019). Although studies using the DRM paradigm have made practical contributions (e.g., Reyna et al., 2017), it is sometimes criticized for its lack of ecological validity (DePrince et al., 2004; Wade et al., 2007). We were interested what memory researchers considered to be the SESOI for studies using paradigms that have been criticized for its lack of ecological validity.

In the misinformation paradigm, participants witness an event (e.g., an unarmed robbery) and are then presented with post-event misleading information (e.g., the robber carried a gun; Loftus et al., 1978). The standard finding is that people falsely report seeing the misinformation (e.g., gun) during the witnessed event, a phenomenon known as the misinformation effect (Loftus, 2005). Currently, misinformation is omnipresent in everyday life as seen, for example, with the exposure of fake news about elections, politicians, and diseases (Lewandowsky et al., 2017). Hence, this paradigm closely resembles experiences we daily encounter, suggesting that the SESOI

for the DRM paradigm might not be appropriate when using the misinformation paradigm. Moreover, research showed that false memories elicited in the DRM and misinformation paradigms are rarely statistically correlated (Ost et al., 2013; Patihis et al., 2018; Zhu et al., 2013). In other words, a participant with false memories induced by the misinformation paradigm will not necessarily produce false memories in the DRM paradigm. Additionally, we selected the misinformation paradigm because it examines a different type of false memory namely: Suggestion-induced false memories which are false memories evoked by external misinformation (e.g., suggestive therapy, fake news, etc.; Loftus, 2005).

Finally, we examined the SESOI for the forced fabrication paradigm (Ackil & Zaragoza, 1998). In this paradigm, participants are presented with a short video (i.e., boy on a summer camp) and are then interviewed about it. During this interview, some participants are forced to answer all questions and guess if they do not know the answers (i.e., forced fabrication group), while other participants have to answer honestly and avoid guessing (i.e., honest group). Interestingly, participants are asked questions about details that actually occurred in the video but, more importantly, also about details that were not present. A recurrent finding is that participants who fabricate a response about non-presented details, form false memories for these fabrications (Zaragoza et al., 2007). The forced fabrication procedure illustrates real-life situations that sometimes occur during investigative interviewing, wherein investigators repeatedly ask the same question or force witnesses, or suspects, to answer a question (Kassin, 2006). Hence, establishing a SESOI for such paradigm where results have clear practical implications is vital. Moreover, in the forced fabrication paradigm a different kind of false memory is examined, which can be regarded as a mixture of spontaneous and suggestion-induced false memories. More specifically, participants produce false memories for self-generated information (i.e., fabrications), also known as internal misinformation, but this does not arise spontaneously because it is prompted by external pressure (e.g., investigators pressuring for answers).

Besides the SESOI for practical relevance, we also examined what memory researchers deemed a suitable SESOI for theoretical advancement. Ideally, theories are formal meaning that they can be expressed in mathematical terms and should be able to make specific predictions (Muthukrishna & Henrich, 2019). For example, a formal theory should be able to predict that on average an increase of  $x$  amount of false memories in a situation where  $y$  amount of suggestive interviewing techniques are used. However, as Gruijters and Peters (2020) argued, theories in the social sciences and thus also false memory research oftentimes are not able to make such predictions about the size of an effect in specific situations but simply whether there is an effect (informal theory; Meehl, 1967). This further complicates the decision which SESOI is interesting in support of a theory (Gruijters & Peters, 2020). Hence, we were interested what memory researchers would consider to be the SESOI for theoretical purposes.

Thus, in the current study, we investigated what memory researchers considered to be the SESOI for practical and theoretical purposes in false memory research. To examine this, we presented memory researchers with three hypothetical research designs and

procedures as well as with the design, procedure, and results of three influential false memory papers for each of the abovementioned paradigms. Then, they were asked about their expert opinion about the SESOI for practical and theoretical matters in terms of raw mean differences for each of the hypothetical and influential paper studies. Additionally, we asked experts what they considered small, medium, and large effects. The study was exploratory and, thus, we did not have any a priori hypotheses.

## 4 | METHOD

### 4.1 | Participants

We recruited a total of 75 memory researchers for our survey. Of those, 34 did not complete any of the hypothetical or influential paper scenarios and were thus excluded from any analyses. Of these 34 participants, eight indicated that they had not published peer reviewed articles in the field of false memory, while an additional one stopped at the question on whether the participant had published in the field of false memory, possibly because the participant had not. The remaining 25 participants who decided not to continue might have stopped for several reasons such as being not sure about several statistical issues or simply due to time constraints. As a result, we included 41 memory researchers in our analyses of which 27 completed all scenarios. Participants' age ranged from 23 to 77 years old ( $M_{age} = 44.5$ ,  $SD_{age} = 14.2$ ), and 65.9% were females (see Table 1). An a priori power analysis was not conducted as the study was exploratory. Moreover, because we aimed to target (false) memory researchers, our pool of participants was limited. Hence, we decided to recruit as many memory researchers as possible in the following two ways. First, we sent emails (initial email and two reminders) to a list of memory researchers that published peer-reviewed papers in the field of false memory based on the authors' knowledge and network. Moreover, we sent emails via the Society for Applied Research on Memory and Cognition and the European Association of Psychology and Law to recruit memory researchers that published peer-reviewed articles about false memories. Three 25-dollar Amazon vouchers were raffled among the participants who participated in the study.

The Social And Societal Ethics Committee and Privacy and Ethics Unit of KU Leuven approved this study (G-2021-3516-R2[MIN]). The data and supplemental materials are available on the Open Science Framework (OSF; <https://osf.io/8y5vt/>).<sup>1</sup>

### 4.2 | Materials

#### 4.2.1 | Hypothetical scenarios

We created three hypothetical experimental designs and procedures for the DRM, misinformation, and forced fabrication paradigm (see Supplemental Materials). In the hypothetical scenarios, participants were presented with a between-subject design using the typical

**TABLE 1** Demographical information of memory researchers

Characteristic	n	%
Gender		
Male	14	34.1
Female	27	65.9
Ethnicity		
Asian	4	9.7
Other (specify)	2	4.9
White	35	85.4
Nationality		
Australia	1	2.4
Canada	4	9.7
China	2	4.9
Czech Republic	1	2.4
Germany	1	2.4
Indonesia	1	2.4
Ireland	2	4.9
Italy	1	2.4
Netherlands	5	12.2
New Zealand	1	2.4
Spain	1	2.4
United Kingdom	2	4.9
United States of America	19	46.3
Education		
Bachelor's degree (e.g., BA, BS)	2	4.9
Master's degree (e.g., MA, MS, MEd)	3	7.3
Doctorate (e.g., PhD, MD, JD, EeD)	36	87.8
# of false memory peer reviewed publications		
1 article	9	22.0
2–5 articles	7	17.1
6–10 articles	8	19.5
More than 10 articles	17	41.5
Applied or theoretical research, or both		
Applied	8	19.5
Theoretical	8	19.5
Both	25	61.0

procedure of each paradigm. The hypothetical scenarios depicted a short explanation of each paradigm and the manipulation. Moreover, for the hypothetical scenarios the maximum amount of critical lures (DRM), misinformation details, and forced fabrication details were set to 10. This allowed us to make clear comparisons for the SESOI across the three paradigms. For example for the hypothetical scenario of the DRM paradigm the participants received the following:

“Imagine the following experiment: A researcher wants to examine whether a certain therapy can lead to increases (or decreases) in spontaneous false memory

formation using the DRM paradigm. Hence, participants are split up into two groups: “therapy” and “no-therapy.” All participants are instructed to study 10 DRM word lists consisting of 10 words each. This means that there are in total 10 critical lures (i.e., spontaneous false memories). Afterwards, participants in both groups will complete a free recall task. However, participants in the “therapy” group receive the therapy while recalling the previously studied words while the “no-therapy” group simply recalls what they remember.”

Following the description of the hypothetical scenario, participants were asked what they considered the SESOI should be in terms of raw mean difference for practical and theoretical matters. We also asked participants to give a rationale for the chosen SESOIs. Additionally, participants were asked what they considered small, medium, and large effects.

#### 4.2.2 | Influential paper scenarios

Three published papers using each specific paradigm were chosen (see Supplemental Materials; Ackil & Zaragoza, 1998; Assefi & Garry, 2003; Payne et al., 2009). Articles were chosen because they used a between subjects design, were highly cited (i.e., more than 100 citations), and made practical recommendations based on their results. For instance, the influential paper scenario of the DRM paradigm was as follows:

A study by Payne et al. (2009) examined the effects of sleep on false memory formation using the DRM paradigm. In one of their experiments, they divided the participants into two groups: Wake and sleep group. All participants listened to a recording of 8 DRM word lists. This means that there were in total 8 critical lures (i.e., spontaneous false memories). However, participants in the “wake” group studied the word-lists at 9 a.m. and recalled the words at 9 p.m. that same day while participants in the “sleep” group studied the word-lists at 9 PM and recalled the words the day after at 9 AM. The authors found that participants in the “sleep” group falsely recalled more critical lures than participants in the “wake” group,  $t(138) = 2.8$ ,  $p = .005$ , raw mean difference of .7 critical lures.

Following the description of the influential paper scenario, participants were asked what they considered the SESOI should be in terms of raw mean differences for practical and theoretical matters. For the influential paper scenarios, the observed effect of the study was included as an option for the SESOI. We also asked participants to give a rationale for the chosen SESOIs. Moreover, they were asked what they considered small, medium, and large effects.

### 4.3 | Procedure

Participants were sent an email containing a Qualtrics link which directed them to the questionnaire. Before the study, participants gave their informed consent and subsequently answered some demographical questions. Then participants received information explaining what the SESOI is and a reminder of what the Deese/Roediger-McDermott, misinformation, and forced fabrication paradigms consist of (see Supplemental Materials). Participants were able to consult this information throughout the experiment. Then, participants were presented with the hypothetical and influential paper scenarios one-by-one in complete random order. Participants received the following instructions:

On the following slides you will be presented with 6 scenarios: 3 hypothetical and 3 from peer-reviewed published articles. The hypothetical scenarios consist of the design and procedure of false memory studies. It is possible that the hypothetical scenarios will deviate from peer-reviewed studies in minor ways. The scenarios from peer-reviewed articles will entail the design, procedure, and results from the original study without any deviations. For each scenario, we will ask you some questions concerning your perspective on the smallest effect size of interest. For both types of scenario you can assume that the within group variance is small as we are mainly interested in what you consider to be the smallest effect size of interest between groups.

Upon completion participants were debriefed and thanked for their participation.

### 4.4 | Coding scheme

To examine the SESOIs for the various paradigms, we examined the responses of memory researchers in terms of the value they gave as the SESOI for practical and theoretical matters. Upon inspection of the data, we observed that only a few respondents opted for “2 raw mean difference,” “3 raw mean difference,” or “4 or more raw mean difference” across scenarios. To enhance the interpretation of our results, we collapsed these scores and referred to them as larger SESOIs meaning 2 or more raw mean differences. Additionally, we examined the given rationales for the chosen SESOI. To do so, we categorized the rationales into nine different categories by inspecting the participants' answers: (1) Any effect can be interesting/have major consequences, (2) Subjective opinion (e.g., 10% misinformation effect would be interesting to me), (3) Convention (e.g., Cohen's standards), (4) Any effect if replicable/reliable, (5) Low ecological validity/paradigm related, (6) Recommendations require large effects, (7) Need more info (e.g., design, other variables), (8) No Rationale (9) Other (see Supplemental Materials for all rationales and examples for each category). Then, we assigned the rationales given by the memory researchers to one of these nine categories. Finally, we calculated the averages of what memory researchers considered a small, medium,

and large effect in raw mean differences for each hypothetical and influential paper scenario.

## 5 | RESULTS

### 5.1 | SESOI for practical implications for hypothetical scenarios

Our results showed that although there was no clear consensus, there was a tendency for a smaller SESOI (e.g., “1 raw mean difference”) or to indicate “any difference that leads to a  $p < .05$ ” as the SESOI across the three different paradigms for practical purposes (see Table 2).

#### 5.1.1 | DRM paradigm

For the DRM paradigm, 20.6% of participants (7/34) indicated “any difference that leads to a  $p < .05$ ” as the SESOI and 17.6% of participants (6/34) considered “1 critical lure difference” as the SESOI, 23.6% of participants (8/34) requested a larger SESOI (“2 or more critical lures difference”), while 38.2% of participants (13/34) indicated the option “other.” The “other” responses showed that 8.8% of participants (3/34) used Cohen's benchmarks, another 8.8% of participants chose smaller raw mean differences (.01 or .5 raw mean difference) than the given options, one participant out of 34 (2.9%) chose a larger raw mean difference (6 raw mean differences), one participant chose any effect as the SESOI as long as it was reliable, and 14.7% of participants (5/34) did not provide a SESOI because they either disagreed with the paradigm or needed more information.

#### 5.1.2 | Misinformation paradigm

For the misinformation paradigm, 21.9% of participants (7/32) considered “any difference that leads to a  $p < .05$ ” to be the SESOI, 15.6% of participants (5/32) believed it should be “1 misinformation detail difference,” 28.1% of participants (9/32) wanted at least 2 or more critical lures difference as the SESOI, and 34.4% of participants (11/32) responded “other.” The “other” responses revealed that 12.5% of participants (4/32) used Cohen's benchmarks, two out of 32 participants (6.3%) chose .5 misinformation detail difference, one participant out of 32 (3.1%) chose a larger raw mean difference (6 raw mean differences), two participants (6.3%) considered any effect as the SESOI as long as it was reliable, and two participants needed more information or disagreed with the given scenario.

#### 5.1.3 | Forced fabrication paradigm

For the forced fabrication paradigm, 20.6% of participants (7/32) indicated that “any difference that leads to a  $p < .05$ ” should be the SESOI, 32.4% of participants (11/32) believed it should be at least “1

**TABLE 2** Experts opinion on smallest effect size of interest (SESOI) for hypothetical scenarios

DRM	Any difference that leads to a $p < .05$		1 critical lure		2 critical lures		3 critical lures		4 or more critical lures		Other	
	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%
SESOI practical	7	20.6	6	17.6	4	11.8	2	5.9	2	5.9	13	38.2
SESOI theoretical	10	30.3	7	21.2	3	9.1	0	0.0	2	6.1	11	33.3
Misinformation	Any difference that leads to a $p < .05$		1 misinformation detail		2 misinformation details		3 misinformation details		4 misinformation details		Other	
	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%
SESOI practical	7	21.9	5	15.6	7	21.9	1	3.1	1	3.1	11	34.4
SESOI theoretical	12	37.5	5	15.6	1	3.1	2	6.3	1	3.1	11	34.4
Forced fabrication	Any difference that leads to a $p < .05$		1 fabrication detail		2 fabrication details		3 fabrication details		4 or more fabrication details		Other	
	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%
SESOI practical	7	20.6	11	32.4	4	11.8	0	0.0	1	2.9	11	32.4
SESOI theoretical	11	33.3	6	18.2	2	6.1	1	3.0	1	3.0	12	36.4

Note: *n* stands for number of participants and % stands for percentage of participants that indicated a specific agreement. The total participants might differ between paradigms because not all participants completed the whole experiment.

fabrication detail difference, 14.7% of participants (5/32) believed the SESOI should be at least 2 or more fabrication details difference, and 32.4% of participants (11/32) indicated “other.” The “other” responses showed that 9.4% of participants (3/32) used Cohen's benchmarks, one participant out of 32 (3.1%) chose .01 fabrication detail difference as the SESOI, another participant chose 6 fabrication detail difference as the SESOI, 6.3% of participants (2/32) considered any effect as the SESOI as long as it was reliable, and 12.5% of participants (4/32) did not provide a SESOI because they needed more information or disagreed with the given scenario.

## 5.2 | SESOI for theoretical implications for hypothetical scenarios

We found that more memory researchers accepted a smaller SESOI or “any difference that leads to a  $p < .05$ ” as their SESOI when implications were made in regard to theory.

### 5.2.1 | DRM paradigm

Specifically, for the DRM paradigm, 30.3% of participants (10/33) considered “any difference that leads to a  $p < .05$ ” as their SESOI, 21.2% of participants (7/33) believed the SESOI should be “1 critical lure difference,” 15.2% of participants (5/33) indicated that the SESOI should be 2 or more critical lures difference, and 33.3% of participants (11/33) indicated “other.” The “other” responses revealed that 9.1% of participants (3/33) used Cohen's benchmarks, 9.1% of participants used smaller raw mean differences (.01 or .5 critical lure difference) than the given options, 6.1% of participants (2/33) chose any effect as the SESOI as long as it is reliable, and 9.1% of participants did not give a SESOI because they either disagreed with the scenario or needed more information.

### 5.2.2 | Misinformation paradigm

For the misinformation paradigm, 37.5% of participants (12/33) believed the SESOI should be “any difference that leads to a  $p < .05$ ,” 15.6% of participants (5/33) asked for at least a “1 misinformation detail difference” for the SESOI, 12.6% of participants (4/33) wanted at least a SESOI of 2 or more misinformation details difference, while 34.4% of participants (11/33) indicated “other.” The “other” responses showed that 15.2% of participants (5/33) used Cohen's benchmarks, 6.1% of participants (2/33) indicated smaller raw mean differences (.01 or .5 misinformation detail difference) than the given options, 9.1% of participants (3/33) considered any effect as the SESOI as long as it was reliable, and one participant out of 33 participants (3.0%) did not agree with the given scenario.

### 5.2.3 | Forced fabrication paradigm

For the forced fabrication paradigm, 33.3% of participants (11/33) chose “any difference that leads to a  $p < .05$ ” as their SESOI and 18.2% of participants (6/33) believed the SESOI should be “1 fabrication detail difference,” 12.2% of participants (4/33) indicated that SESOI should be 2 or more fabrication details difference, while 36.4% of participants (12/33) indicated “other.” The “other” responses revealed that 12.1% of participants (4/33) used Cohen's benchmarks, 6.1% of participants (2/33) chose smaller raw mean differences (.01 or .5 fabrication detail difference), one participant out of 33 (3.0%) indicated a larger raw mean difference (6 fabrication details difference), 6.1% of participants (2/33) considered any effect as the SESOI as long as it is reliable, and 9.1% of participants (3/33) did not provide a SESOI because they needed more information or disagreed with the given scenario.

**TABLE 3** Expert rationales for SESOI hypothetical scenarios

Rationale	DRM paradigm				Misinformation paradigm				Forced fabrication paradigm			
	SESOI practical		SESOI theoretical		SESOI practical		SESOI theoretical		SESOI practical		SESOI theoretical	
	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%
Any effect can be interesting/have major consequences	5	14.7	6	18.1	6	18.8	9	28.1	11	32.4	6	18.1
Subjective opinion	5	14.7	6	18.1	4	12.5	5	15.6	6	17.6	4	12.1
Convention (e.g., Cohen's standards)	3	8.8	3	9.1	5	15.6	4	12.5	4	11.8	4	12.1
Any effect if replicable/reliable	3	8.8	5	15.2	2	6.3	3	9.4	2	5.9	6	18.1
Low ecological validity/paradigm related	4	11.8	4	12.1	1	3.3	1	3.1	/	/	/	/
Recommendations require large effects	2	5.9	1	3.0	3	9.4	1	3.1	1	2.9	4	12.1
Need more info (e.g., design)	3	8.8	3	9.1	2	6.3	1	3.1	1	2.9	3	9.1
No rationale	3	8.8	1	3.0	5	15.6	3	9.4	2	5.9	2	6.1
Other	6	17.6	4	12.1	4	12.5	5	15.6	7	20.6	4	12.1

Note: Based on the rationales given by the experts, nine different categories were identified. Then we allocated the rationales of the experts that coincided most with one of the categories.

**TABLE 4** Experts opinion on smallest effect size of interest (SESOI) for influential paper scenarios

DRM paradigm	Any difference that leads to a $p < .05$		.70 critical lures		1 critical lure		2 critical lures		3 critical lures		4 or more critical lures		Other	
	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%
SESOI practical	6	18.8	2	6.3	6	18.8	4	12.5	1	3.1	3	9.4	10	31.3
SESOI theoretical	11	33.3	1	3.0	3	9.1	1	3.0	2	6.1	1	3.0	14	42.4

Misinformation paradigm	Any difference that leads to a $p < .05$		.75 misinformation details		1 misinformation detail		2 misinformation details		3 misinformation details		4 or more misinformation details		Other	
	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%
SESOI practical	7	24.1	4	13.8	6	20.7	1	3.4	1	3.4	0	0.0	10	34.5
SESOI theoretical	12	40.0	4	13.3	1	3.3	2	6.7	0	0.0	1	3.3	10	33.3

Forced fabrication paradigm	Any difference that leads to a $p < .05$		0.54 fabrication details		1 fabrication detail		2 fabrication details		3 or more fabrication details		Other	
	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%
SESOI Practical	4	13.8	5	17.2	5	17.2	1	3.5	2	6.9	12	41.4
SESOI Theoretical	8	28.6	4	14.3	0	0.0	2	7.1	2	7.1	12	42.9

Note: *n* stands for number of participants and % stands for percentage of participants that indicated a specific agreement. The total participants might differ between paradigms because not all participants completed the whole experiment.

Lastly, we also examined the rationales the memory researchers gave for their chosen SESOI. We found that their SESOI justification varied from “no rationale” to taking into account the specific paradigm at hand (see Table 3)

### 5.3 | SESOI for practical implications for influential paper scenarios

In line with the hypothetical scenarios depicted above, our results showed that there was no clear consensus among memory

researchers for the SESOI but they leaned towards smaller SESOIs (e.g., “1 raw mean difference) or “any difference that leads to a  $p < .05$ ” as their SESOI for practical implications (see Table 4).

#### 5.3.1 | DRM paradigm

Specifically, for the DRM paradigm, we found that 18.8% of participants (6/32) indicated “any difference that leads to a  $p < .05$ ” as their SESOI, two out of the 32 participants (6.3%) indicated the observed effect (.70 critical lures difference) as the SESOI, 18.8% of participants

(6/32) believed the SESOI should be “1 critical lure difference,” 25% of participants (8/32) demanded a SESOI of at least 2 or more critical lures difference, while 31.3% of participants (10/32) indicated the option “other.” The “other” responses showed that 6.3% of participants (2/32) used Cohen's benchmarks, one participant out of 32 (3.1%) indicated a .01 critical lure difference as the SESOI, one participant indicated .8 critical lure difference, one participant indicated a 6 critical lure difference, another participant deemed any effect as the SESOI as long as it is reliable, and 12.5% of participants (4/32) did not provide a SESOI because they needed more information or disagreed with the scenario.

### 5.3.2 | Misinformation paradigm

For the misinformation paradigm, 24.1% of participants (7/29) accepted “any difference that leads to a  $p < .05$ ” as their SESOI, 13.8% of participants (4/29) indicated the observed effect size (.75 misinformation details difference), 20.7% of participants (6/29) believed the SESOI should be “1 misinformation detail difference,” two out of 29 participants (6.8%) demanded a SESOI greater than 2 or more misinformation details, while 34.5% of participants (10/29) indicated “other.” The “other” responses revealed that 6.9% of participants (2/29) used Cohen's benchmarks, 13.8% of participants (4/29) indicated smaller raw mean differences (.4, .5, or less than 1 misinformation detail difference) as the SESOI, 6.9% of participants deemed any effect as the SESOI as long as it is reliable, and 6.9% of participants did not provide a SESOI because they needed more information or disagreed with the given scenario.

### 5.3.3 | Forced fabrication paradigm

For the forced fabrication paradigm, 13.8% of participants (4/29) considered “any difference that leads to a  $p < .05$ ” as the SESOI, 17.2% of participants (5/29) chose the observed effect (.54 fabrication details difference) and another 17.2% indicated “1 fabrication detail difference” as their SESOI, 3 out of the 29 participants (10.3%) demanded at least 2 or more fabrication details difference for the SESOI, while 41.4% of participants (12/29) indicated “other.” The “other” responses showed that 10.3% of participants (3/29) used Cohen's benchmarks, 13.8% of participants (4/29) indicated smaller raw mean differences (between .01 and .5 fabrication detail difference), 6.9% of participants (2/29) deemed any effect as the SESOI as long as it is reliable, and 10.3% of participants did not provide a SESOI because they needed more information or disagreed with the given scenario.

## 5.4 | SESOI for theoretical implications for influential paper scenarios

As seen with the hypothetical scenarios, in comparison with the SESOI for practical implications, more memory researchers accepted

smaller SESOIs or “any difference that leads to a  $p < .05$ ” as their SESOI for theoretical purposes.

### 5.4.1 | DRM paradigm

For the DRM paradigm, 33.3% of participants (11/33) indicated that “any difference that leads to a  $p < .05$ ” should be the SESOI, one out of the 33 participants (3.0%) indicated the observed effect (.70 critical lures difference) as the SESOI, 9.1% of participants (3/33) believed the SESOI should be at least “1 critical lure difference,” 12.1% of participants (4/33) wanted at least two or more critical lures difference for the SESOI, while 42.4% of participants (14/33) indicated “other.” The “other” responses showed that 12.1% of participants (4/33) used Cohen's benchmarks, 18.1% of participants (6/33) indicated smaller raw mean differences (.01, .4, .5, or .8 critical lure difference), 6.1% of participants (2/33) considered any effect as the SESOI as long as it was reliable, and 6.1% of participants did not provide a SESOI because they needed more information or disagreed with the given scenario.

### 5.4.2 | Misinformation paradigm

For the misinformation paradigm, 40% of participants (12/30) deemed “any difference that leads to a  $p < .05$ ” sufficient as the SESOI, 13.3% of participants (4/30) chose the observed effect (.75 misinformation details difference), and 3.3% of participants (1/30) indicated “1 misinformation detail difference” as the SESOI, three out of the 30 participants (10.0%) demanded at least two or more misinformation details difference for the SESOI, while 33.3% of participants (10/30) indicated “other.” The “other” responses showed that 6.7% of participants (2/30) used Cohen's benchmarks, 16.7% of participants (5/30) indicated smaller raw mean differences (between .01 and .5 misinformation detail difference), 6.7% of participants considered any effect as the SESOI as long as it was reliable, and one participant out of 30 (3.3%) did not provide a SESOI because of disagreement with the given scenario.

### 5.4.3 | Forced fabrication paradigm

For the forced fabrication paradigm, 28.6% of participants (8/28) accepted “any difference that leads to a  $p < .05$ ” as the SESOI, 14.3% of participants (4/28) believed that the observed effect (.54 fabrication details) should be the SESOI, 14.2% of participants indicated that the SESOI should be at least two or more fabrication details, while 42.9% of participants (12/29) indicated “other.” The “other” responses revealed that 7.1% of participants (2/28) used Cohen's benchmarks, 17.8% of participants (5/28) indicated smaller raw mean differences (between .01 and .5 fabrication detail difference), one participant out of 28 (3.6%) indicated the maximum amount of fabrication detail differences, 7.1% of participants considered any effect as



the SESOI as long as it was reliable, and 7.1% of participants did not provide a SESOI because they needed more information or disagreed with the given scenario.

In line with the hypothetical scenarios we found that SESOI justification between memory researchers were mixed (see Table 5).

## 5.5 | Small, medium, and large effects

We also examined what memory researchers considered small, medium, and large effects in raw mean differences for each of the given scenarios. For the hypothetical scenarios we found that memory researchers had a general consensus across paradigms for what is deemed a small effect ( $M_{\text{DRM}} = 1.15$ ,  $SD_{\text{DRM}} = .70$ ;  $M_{\text{misinformation}} = 1.06$ ,  $SD_{\text{misinformation}} = .68$ ;  $M_{\text{forced fabrication}} = 1.20$ ,  $SD_{\text{forced fabrication}} = .71$ ) and medium effect ( $M_{\text{DRM}} = 2.34$ ,  $SD_{\text{DRM}} = 1.27$ ;  $M_{\text{misinformation}} = 2.43$ ,  $SD_{\text{misinformation}} = 1.38$ ;  $M_{\text{forced fabrication}} = 2.66$ ,  $SD_{\text{forced fabrication}} = 1.29$ ).

However, memory researchers demanded a larger effect in the forced fabrication paradigm ( $M_{\text{forced fabrication}} = 4.43$ ,  $SD_{\text{forced fabrication}} = 2.28$ ) as compared with the DRM ( $M_{\text{DRM}} = 4.02$ ,  $SD_{\text{DRM}} = 2.24$ ) and the misinformation paradigms ( $M_{\text{misinformation}} = 4.01$ ,  $SD_{\text{misinformation}} = 2.55$ ) (Table 6).

For the influential paper scenarios, we found that memory researchers demanded a greater small effect size for the DRM scenario ( $M_{\text{DRM}} = .93$ ,  $SD_{\text{DRM}} = .58$ ) as compared with the misinformation ( $M_{\text{misinformation}} = .70$ ,  $SD_{\text{misinformation}} = .63$ ) and forced fabrication scenarios ( $M_{\text{forced fabrication}} = .74$ ,  $SD_{\text{forced fabrication}} = .63$ ). The memory researchers also demanded a greater medium effect size for the DRM scenario ( $M_{\text{DRM}} = 2.02$ ,  $SD_{\text{DRM}} = 1.06$ ) than the misinformation ( $M_{\text{misinformation}} = 1.27$ ,  $SD_{\text{misinformation}} = .76$ ) and the forced fabrication paradigms ( $M_{\text{forced fabrication}} = 1.21$ ,  $SD_{\text{forced fabrication}} = .91$ ). The same was observed for large effect sizes wherein memory researchers wanted a greater effect for the DRM scenario ( $M_{\text{DRM}} = 3.18$ ,  $SD_{\text{DRM}} = 1.54$ ) as compared with the misinformation

**TABLE 5** Expert rationales for SESOI influential paper scenarios

Rationale	DRM paradigm				Misinformation paradigm				Forced fabrication paradigm			
	SESOI practical		SESOI theoretical		SESOI practical		SESOI theoretical		SESOI practical		SESOI theoretical	
	n	%	n	%	n	%	n	%	n	%	n	%
Any effect can be interesting/have major consequences	5	15.6	4	12.1	6	20.7	5	16.7	4	13.8	3	10.7
Subjective opinion	2	6.3	4	12.1	2	6.9	2	6.7	4	13.8	4	14.3
Convention (e.g., Cohen's standards)	5	15.6	8	24.2	4	13.8	6	20	3	10.3	6	21.4
Any effect if replicable/reliable	2	6.3	5	15.2	3	10.3	7	23.3	2	6.9	5	17.9
Low ecological validity/paradigm related	6	18.8	2	6.1	/	/	/	/	2	6.9	/	/
Recommendations require large effects	2	6.3	1	3.0	/	/	1	3.3	/	/	2	7.1
Need more info (e.g., design)	3	9.4	2	6.1	2	6.9	2	6.7	5	17.2	3	10.7
No rationale	4	12.5	3	9.1	6	20.7	5	16.7	4	13.8	1	3.6
Other	3	9.4	4	12.1	6	20.7	2	6.7	5	17.2	4	14.3

Note: Based on the rationales given by the experts, nine different categories were identified. Then we allocated the rationales of the experts that coincided most with one of the categories.

**TABLE 6** Averages for expert opinions for a small, medium, and large effect size for hypothetical scenarios

DRM paradigm	Small effect in critical lures M (SD)	Medium effect in critical lures M (SD)	Large effect in critical lures M (SD)
	1.15 (.70)	2.34 (1.27)	4.02 (2.24)
Misinformation paradigm	Small effect in misinformation details M (SD)	Medium effect in misinformation details M (SD)	Large effect in misinformation details M (SD)
	1.06 (.68)	2.43 (1.38)	4.01 (2.55)
Forced fabrication paradigm	Small effect in fabricated details M (SD)	Medium effect in fabricated details M (SD)	Large effect in fabricated details M (SD)
	1.20 (.71)	2.66 (1.29)	4.43 (2.28)

Note: M is mean and SD is standard deviation. Effect sizes are given as averages of the raw mean differences given by the experts.

DRM paradigm	Small effect in critical lures M (SD)	Medium effect in critical lures M (SD)	Large effect in critical lures M (SD)
	.93 (.58)	2.02 (1.06)	3.18 (1.54)
Misinformation paradigm	Small effect in misinformation details M (SD)	Medium effect in misinformation details M (SD)	Large effect in misinformation details M (SD)
	.70 (.63)	1.27 (.76)	1.67 (.88)
Forced fabrication paradigm	Small effect in fabricated details M (SD)	Medium effect in fabricated details M (SD)	Large effect in fabricated details M (SD)
	.74 (.63)	1.21 (.91)	1.60 (.90)

**TABLE 7** Averages for expert opinions for a small, medium, and large effect size for influential paper scenarios

Note: *M* is mean and *SD* is standard deviation. Effect sizes are given as averages of the raw mean differences given by the experts. Some participants were excluded because they indicated a raw mean difference that exceeded the maximum amount. For example, in the influential paper scenario of the DRM paradigm the maximum was 8 critical lures and some participants indicated that a large effect had to be 10 critical lures.

( $M_{\text{misinformation}} = 1.67$ ,  $SD_{\text{misinformation}} = .88$ ) and forced fabrication scenarios ( $M_{\text{forced fabrication}} = 1.60$ ,  $SD_{\text{forced fabrication}} = .90$ ) (Table 7).

## 6 | DISCUSSION

Establishing a credible SESOI in psychological studies is an important step to increase the practical and theoretical significance of psychological research. Across and within various false memory paradigms, and for practical and theoretical purposes, we found no clear consensus for such SESOI among memory experts. However, across all scenarios, memory experts leaned towards smaller effects as the SESOI (1 raw mean difference) or “any difference that leads to a  $p < .05$ ” as the SESOI, especially for theoretical purposes. Moreover, our results showed that the rationales for the chosen SESOIs varied substantially across and within scenarios and ranged from relying on conventions such as Cohen's benchmarks to concerns about ecological validity.

One of the main findings of the present study is that no clear consensus emerged for the SESOI across and within scenarios. However, we argue that this should not be a reason for concern. That is, the SESOI can differ based on the context (e.g., area of research) but also factors such as the used stimuli, whether research is practically or theoretically focused, and the philosophy of science (Cohen, 1988). Indeed, we found that part of the observed differences in the given SESOI within and across scenarios stem from that type of reasoning. That is, some experts argued that recommendations for either practical or theoretical implications need large effect sizes, some referred to the used paradigm, while others asserted that small effects can have major consequences. However, other discrepancies were caused by a substantial amount of experts relying on conventions (e.g., Cohen's benchmarks), subjective reasoning (e.g., “I believe that a 1 raw mean difference is interesting”), or not providing any rationale. We argue that it can be problematic when disagreements about fundamental

issues such as the SESOI arise when experts provide no rationale, rely only on subjective reasoning, or use conventions (e.g., Cohen's benchmarks) because such rationales have been argued to be weak justifications (Hill et al., 2008; Lakens et al., 2018). Our study thus suggests that there is no clear consensus for the SESOI in false memory research among experts. However, it is possible that experts come to a general agreement on such matters when different consensus development methods such as the Delphi or nominal group technique (NGT) are used (Black et al., 1999). In short, in the Delphi and NGT methods, experts are provided with a summary of the groups' responses after they shared their perspective on, for example, the SESOI. Then, they can adjust their answers, and in the NGT they even interact and discuss face-to-face with other experts before giving their final views on the SESOI. Even though the Delphi and NGT methods have been criticized for its inconsistent use and poor standardization (Humphrey-Murto et al., 2017), future research could investigate whether using such methods can lead to a general consensus among researchers for the SESOI in false memory research.

Another interesting result is that a considerable amount of experts considered any effect that leads to a  $p < .05$  as the SESOI to make practical recommendations such as advocating against the use of suggestive interviewing tactics. That is, based on our findings, memory researchers seemed to conflate statistical significance (i.e.,  $p < .05$ ) with a practically meaningful effect size. This is concerning because statistical significance is not the same as a practically meaningful effect. More specifically, on the one hand, as sample sizes increase, smaller and possibly trivial effects can become statistically significant (Anvari & Lakens, 2021). On the other hand, previous research has shown that results obtained from small sample sizes can be unreliable, difficult to replicate, and inflate the effect sizes (Button et al., 2013). Hence, basing the SESOI for practical implications on an arbitrary  $p < .05$  cutoff is inappropriate and, in combination with the observed weak justifications given for the chosen SESOIs, might

indicate a lack of statistical knowledge about these topics among memory researchers. An interesting follow-up study could examine whether providing memory researchers with relevant statistical training about the SESOI beforehand allows them to give more educated answers to the given scenarios used in the current study.

Noteworthy is that some of the experts pointed out that it was more important whether the effect is reliable (e.g., replicated) rather than it being “large” (see Tables 3 & 5). It is true that small effects can have real-life implications (Funder & Ozer, 2019), such as the effect of aspirin consumption on heart attack occurrences as one of the experts referred. However, this does not automatically imply that such implications of small effects of aspirin consumption on heart attack occurrences bears similar meaning in false memory research. Hence, we argue that such effects should be regarded in the context of the area of research (Cohen, 1988; Cortina & Landis, 2009). One expert argued that .01 raw mean differences in critical lures, misinformation, or forced fabrication details, caused by for example therapy, should be the SESOI because 1 detail in 100 therapy sessions could have severe consequences (e.g., miscarriages of justice). However, it is difficult to understand what a .01 raw mean difference truly means and how to interpret this when explaining it in terms of practical implications. Moreover, this is quite different from the aspirin example as therapy is a high-cost intervention (e.g., time, effort) and also brings benefits (e.g., mental health). Hence, in this case, it might be harmful disregarding a certain therapy based such a small increased susceptibility of false memories.

However, an important difference concerns whether or not certain effects bear practical implications, or whether they advance our theoretical understanding. For that reason, we examined whether experts hold different standards for the SESOI for practical versus theoretical matters. Indeed, our results showed that across and within scenarios, the experts tended to be more accepting of smaller effects when it is considered for theoretical purposes. A possible explanation for a lower SESOI for theoretical ends (vs. practical) is that it is more based around the accumulation of knowledge. This was endorsed by some experts who argued that the stakes, especially for false memory research, are lower for theoretical compared with practical implications. Also, theories in social science are typically verbal (as opposed to formal) meaning that it cannot make quantitative predictions but rather whether there is a certain effect, possibly leading memory researchers to accept any difference that leads to a  $p < .05$  as the SESOI for theoretical matters (Gruijters & Peters, 2020).

Interestingly, it seemed that memory researchers demanded a greater SESOI and small, medium, and large effects for the scenarios concerning the DRM paradigm as compared with the misinformation and forced fabrication paradigms. Moreover, in line with previous research (Wade et al., 2007), a returning argument of the experts was that studies using the DRM paradigm lack ecological validity and were hesitant to make causal inferences based on such research. Hence, especially for the DRM paradigm, the memory researchers considered the context of what truly constitutes a false memory and justified their SESOI accordingly.

One of the difficulties some memory researchers had with the current study was that they needed additional information (e.g., information on correct details, specifics of the manipulation) to decide what they

considered the SESOI to be. In the current study, we intentionally presented several scenarios using different paradigms with basic information to examine what experts' choice of the SESOI is in various contexts within the area of false memory research. However, a caveat of the current study is that we did not provide exhaustive details (e.g., costs and benefits of manipulation, exact methodology) for one specific study. Additionally, we gave memory researchers the specific options “any difference that leads to a  $p < .05$ ,” “1, 2, 3, or 4 or more raw mean differences” or a deviating answer from the given options by choosing “other” to indicate what they considered the SESOI to be. Even though many experts used the “other” option to give a different or a more continuous response, it is possible that framing the answers in whole numbers affected the responses of some experts. Hence, future research could assess whether providing memory researchers with a more detailed study, for example using a study they performed in the past, and a continuous scale (e.g., “On a scale from 0 to 4 raw mean differences, what would you consider to be the SESOI?”) allows them to more accurately establish the SESOI.

Interestingly, one memory researcher was concerned that the SESOI will become another publication gatekeeping device which does not add anything to our understanding of psychological science. The expert further argued that when using appropriate conclusions (e.g., “therapy leads to, on average, [point estimate] more/fewer spontaneous false memories.”) the SESOI becomes meaningless or purely subjective and that “it is just as interesting/important to show that a manipulation leads to no appreciable difference.” However, we argue that it is precisely the SESOI that is crucial to establish whether a manipulation leads to a meaningful difference or not, especially for practical implications. In the example of the memory researcher reported before, it might be that a certain therapy leads to, on average, .0001 more/fewer spontaneous false memories and that this is a statistically significant effect ( $p < .05$ ) given a large sample size. Even though this effect might be statistically significant it is difficult to suggest it bears practical significance (Anvari & Lakens, 2021). The SESOI is a fundamental statistical tool which can be used in, for example, equivalence tests (Lakens, 2017) to help assess whether certain manipulations have practical and/or theoretical relevance or, just as interesting and as the expert alluded to, which manipulations do not.

Taken together, our study showed that memory researchers do not have a clear consensus about what the SESOI should be for false memory research. We suggest that such a lack of general consensus is acceptable as long as a proper justification is given for the chosen SESOI. However, our findings showed that a substantial amount of researchers relied on weaker justifications (i.e., conventions, subjective reasoning, no rationale) for the SESOI (Lakens et al., 2018). Moreover, several experts alluded to the idea that small effects can be practically meaningful, but failed to explain or contextualize how this works in the field of false memory research. Additionally, we found that experts tended to endorse any effect that leads to a  $p < .05$  as the SESOI, especially for theoretical purposes which is justifiable given that theories are generally verbal in social sciences (Gruijters & Peters, 2020). However, we believe that following such norms can lead to making practical implications based on trivial differences and that in the process of justifying the SESOI the costs and benefits

and/or harms of a manipulation/intervention should be considered (e.g., benefits/harms of a therapy). To counter such issues, we recommend researchers to contextualize their SESOI in regard to the studied phenomenon at hand, especially in fields such as false memory research where its results can yield far-reaching consequences (e.g., miscarriages of justice; Howe & Knott, 2015). One way to do so is to refer to unstandardized effect sizes whenever possible (Funder & Ozer, 2019) or explain the observed effects more intuitively (e.g., common language explanation; Magnusson, 2021). By doing so, the magnitude of an observed or to be studied effect can be understood in terms of the study at hand instead of the decontextualized Cohen's benchmarks.

Imagine for instance a team of researchers that decides to examine the effects of imagination on false memory formation to understand whether it should be used during therapy. To do so, the researchers decide to use the DRM paradigm. They present the participants with 10 DRM lists consisting of 10 words each. The words are studied one-by-one but some participants are instructed to generate a vivid image of each word (imagination) to help memorize the words while others simply study the words one-by-one. Afterwards, they have to recollect which words they studied before. Because the DRM paradigm has been criticized by some for its lack of ecological validity (DePrince et al., 2004; Wade et al., 2007; but see Brainerd et al., 2008), the researchers can decide that any effect leading to a  $p < .05$  is not sufficient as it might lead to effects that have no practical implications. Moreover, with the knowledge that such a study can have severe consequences, the researchers might argue that the effect of imagination on false memory formation should at least be, on average, a difference of one more detail falsely recollected. They can argue that such an unstandardized effect clearly conveys the adverse effects of imagination on memory by increasing the amount of falsely recollected details, on average, by one. The researchers can then use this unstandardized SESOI to evaluate observed effect sizes in the literature or use it to calculate a standardized effect size and subsequently perform a power analysis. The SESOI and the given justifications may differ based on the used paradigm, the specific conditions of the paradigm (e.g., amount of DRM lists), whether the results are meant for practical or theoretical purposes, or personal philosophy science. However, we argue that using such unstandardized or meaningful effect sizes allows the researcher to contextualize the SESOI and accurately communicate the possible practical or theoretical implications for the studied phenomenon at hand.

#### CONFLICT OF INTEREST

All authors declare no conflict of interest.

#### ENDNOTE

<sup>1</sup> We did not preregister anything for this study.

#### DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available on the Open Science Framework at [https://osf.io/8y5vt/?view\\_only=a3d965d44b164b7ab311de8e603899b9](https://osf.io/8y5vt/?view_only=a3d965d44b164b7ab311de8e603899b9)

#### ORCID

Paul Riesthuis  <https://orcid.org/0000-0001-6520-2453>

Henry Otgaar  <https://orcid.org/0000-0002-2782-2181>

#### REFERENCES

- Ackil, J. K., & Zaragoza, M. S. (1998). Memorial consequences of forced confabulation: Age differences in susceptibility to false memories. *Developmental Psychology, 34*, 1358–1372. <https://doi.org/10.1037/0012-1649.34.6.1358>
- Anvari, F., & Lakens, D. (2021). Using anchor-based methods to determine the smallest effect size of interest. *Journal of Experimental Social Psychology, 96*, 104159. <https://doi.org/10.1016/j.jesp.2021.104159>
- Assefi, S. L., & Garry, M. (2003). Absolut<sup>®</sup> memory distortions: Alcohol placebos influence the misinformation effect. *Psychological Science, 14*, 77–80. <https://doi.org/10.1111/1467-9280.01422>
- Bernstein, D. M., & Loftus, E. F. (2009). How to tell if a particular memory is true or false. *Perspectives on Psychological Science, 4*, 370–374. <https://doi.org/10.1111/j.1745-6924.2009.01140.x>
- Black, N., Murphy, M., Lamping, D., McKee, M., Sanderson, C., Askham, J., & Marteau, T. (1999). Consensus development methods: A review of best practice in creating clinical guidelines. *Journal of Health Services Research & Policy, 4*, 236–248. <https://doi.org/10.1177/135581969900400410>
- Bonini, M., di Paolo, M., Bagnasco, D., Baiardini, I., Braido, F., Caminati, M., Carpagnano, E., Contoli, M., Corsico, A., del Giacco, S., Heffler, E., Lombardi, C., Menichini, I., Milanese, M., Scichilone, N., Senna, G., & Canonica, G. W. (2020). Minimal clinically important difference for asthma endpoints: An expert consensus report. *European Respiratory Review, 29*, 190137. <https://doi.org/10.1183/16000617.0137-2019>
- Bosco, F. A., Aguinis, H., Singh, K., Field, J. G., & Pierce, C. A. (2015). Correlational effect size benchmarks. *Journal of Applied Psychology, 100*, 431. <https://doi.org/10.1037/a0038047>
- Brainerd, C. J., Reyna, V. F., & Ceci, S. J. (2008). Developmental reversals in false memory: A review of data and theory. *Psychological Bulletin, 134*, 343. <https://doi.org/10.1037/0033-2909.134.3.343>
- Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience, 14*, 365–376. <https://doi.org/10.1038/nrn3475>
- 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*, 115–144. <https://doi.org/10.1177/2515245919847196>
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Erlbaum.
- Cortina, J. M., & Landis, R. S. (2009). When small effect size tell a big story, and when large effect sizes don't. In C. E. Lance & R. J. Vandenberg (Eds.), *Statistical and methodological myths and urban legends: Received doctrine, verity, and fable in the organizational and social sciences* (pp. 269–288). Routledge.
- Deese, J. (1959). On the prediction of occurrence of particular verbal intrusions in immediate recall. *Journal of Experimental Psychology, 58*, 17–22. <https://doi.org/10.1037/h0046671>
- 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*, 201–233. [https://doi.org/10.1207/s15327019eb1403\\_1](https://doi.org/10.1207/s15327019eb1403_1)
- Funder, D. C., & Ozer, D. J. (2019). Evaluating effect size in psychological research: Sense and nonsense. *Advances in Methods and Practices in Psychological Science, 2*, 156–168. <https://doi.org/10.1177/2515245919847202>
- Gallo, D. A. (2006). *Associative illusions of memory: False memory research in DRM and related tasks*. Psychology Press.

- Grujters, S. L., & Peters, G. J. Y. (2020). Meaningful change definitions: Sample size planning for experimental intervention research. *Psychology & Health, 1*-16. <https://doi.org/10.1080/08870446.2020.1841762>
- Hill, C. J., Bloom, H. S., Black, A. R., & Lipsey, M. W. (2008). Empirical benchmarks for interpreting effect sizes in research. *Child Development Perspectives, 2*, 172-177. <https://doi.org/10.1111/j.1750-8606.2008.00061.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*, 633-656. <https://doi.org/10.1080/09658211.2015.1010709>
- Humphrey-Murto, S., Varpio, L., Wood, T. J., Gonsalves, C., Ufholz, L. A., Mascioli, K., Wang, C., & Foth, T. (2017). The use of the Delphi and other consensus group methods in medical education research: A review. *Academic Medicine, 92*, 1491-1498. <https://doi.org/10.1097/ACM.0000000000001812>
- Kassin, S. M. (2006). A critical appraisal of modern police interrogations. In T. Williamson (Ed.), *Investigative interviewing: Rights, research, regulation* (pp. 207-228). Willan.
- Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. *Frontiers in Psychology, 4*, 863. <https://doi.org/10.3389/fpsyg.2013.00863>
- Lakens, D. (2014). Performing high-powered studies efficiently with sequential analyses. *European Journal of Social Psychology, 44*, 701-710. <https://doi.org/10.1002/ejsp.2023>
- Lakens, D. (2017). Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social Psychological and Personality Science, 8*, 355-362. <https://doi.org/10.1177/1948550617697177>
- Lakens, D., Scheel, A. M., & Isager, P. M. (2018). Equivalence testing for psychological research: A tutorial. *Advances in Methods and Practices in Psychological Science, 1*, 259-269. <https://doi.org/10.1177/2515245918770963>
- Lewandowsky, S., Ecker, U. K., & Cook, J. (2017). Beyond misinformation: Understanding and coping with the "post-truth" era. *Journal of Applied Research in Memory and Cognition, 6*, 353-369. <https://doi.org/10.1016/j.jarmac.2017.07.008>
- Loftus, E. F. (2005). Planting misinformation in the human mind: A 30-year investigation of the malleability of memory. *Learning & Memory, 12*, 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*, 19. <https://doi.org/10.1037/0278-7393.4.1.19>
- Magnusson, K. (2021). *Interpreting Cohen's d effect size: An interactive visualization* (Version 2.5.1) [Web App]. R Psychologist. <https://rpsychologist.com/cohend/>
- McGlothlin, A. E., & Lewis, R. J. (2014). Minimal clinically important difference: Defining what really matters to patients. *JAMA, 312*, 1342-1343. <https://doi.org/10.1001/jama.2014.13128>
- Meehl, P. E. (1967). Theory-testing in psychology and physics: A methodological paradox. *Philosophy of Science, 34*, 103-115. <https://doi.org/10.1086/288135>
- Mouelhi, Y., Jouve, E., Castelli, C., & Gentile, S. (2020). How is the minimal clinically important difference established in health-related quality of life instruments? Review of anchors and methods. *Health and Quality of Life Outcomes, 18*, 1-17. <https://doi.org/10.1186/s12955-020-01344-w>
- Muthukrishna, M., & Henrich, J. (2019). A problem in theory. *Nature Human Behaviour, 3*, 221-229. <https://doi.org/10.1038/s41562-018-0522-1>
- Ost, J., Blank, H., Davies, J., Jones, G., Lambert, K., & Salmon, K. (2013). False memory ≠ false memory: DRM errors are unrelated to the misinformation effect. *PLoS One, 8*, e57939. <https://doi.org/10.1371/journal.pone.0057939>
- Otgaar, H., Howe, M. L., Muris, P., & Merckelbach, H. (2019). Associative activation as a mechanism underlying false memory formation. *Clinical Psychological Science, 7*, 191-195. <https://doi.org/10.1177/2167702618807189>
- Patihis, L., Frenda, S. J., & Loftus, E. F. (2018). False memory tasks do not reliably predict other false memories. *Psychology of Consciousness: Theory, Research and Practice, 5*, 140. <https://doi.org/10.1037/cns0000147>
- Payne, J. D., Schacter, D. L., Propper, R. E., Huang, L. W., Wamsley, E. J., Tucker, M. A., Walker, M. P., & Stickgold, R. (2009). The role of sleep in false memory formation. *Neurobiology of Learning and Memory, 92*, 327-334. <https://doi.org/10.1016/j.nlm.2009.03.007>
- Ramona v. Ramona, B111565 (1997). <https://caselaw.findlaw.com/ca-court-of-appeal/1122202.html>
- Reyna, V. F., Mills, B., Estrada, S., & Brainerd, C. J. (2017). False memory in children: Data, theory, and legal implications. In *Handbook of eyewitness psychology* (pp. 479-508). Psychology Press.
- 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*, 803. <https://doi.org/10.1037/0278-7393.21.4.803>
- Schäfer, T., & Schwarz, M. A. (2019). The meaningfulness of effect sizes in psychological research: Differences between sub-disciplines and the impact of potential biases. *Frontiers in Psychology, 10*, 813. <https://doi.org/10.3389/fpsyg.2019.00813>
- Van der Heijde, D., Lassere, M., Edmonds, J., Kirwan, J., Strand, V., & Boers, M. (2001). Minimal clinically important difference in plain films in RA: Group discussions, conclusions, and recommendations. OMERACT imaging task force. *The Journal of Rheumatology, 28*, 914-917.
- Wade, K. A., Sharman, S. J., Garry, M., Memon, A., Mazzoni, G., Merckelbach, H., & Loftus, E. F. (2007). False claims about false memory research. *Consciousness and Cognition, 16*, 18-28. <https://doi.org/10.1016/j.concog.2006.07.001>
- Wells, G., Beaton, D., Shea, B., Boers, M., Simon, L., Strand, V., Brooks, P., & Tugwell, P. (2001). Minimal clinically important differences: Review of methods. *The Journal of Rheumatology, 28*, 406-412.
- Zaragoza, M. S., Belli, R. F., & Payment, K. E. (2007). Misinformation effects and the suggestibility of eyewitness memory. Do justice and let the sky fall: Elizabeth Loftus and her contributions to science, law, and academic freedom, 35-63
- Zhu, B., Chen, C., Loftus, E. F., Lin, C., & Dong, Q. (2013). The relationship between DRM and misinformation false memories. *Memory & Cognition, 41*, 832-838. <https://doi.org/10.3758/s13421-013-0300-2>

**How to cite this article:** Riesthuis, P., Mangiulli, I., Broers, N., & Otgaar, H. (2022). Expert opinions on the smallest effect size of interest in false memory research. *Applied Cognitive Psychology, 36*(1), 203-215. <https://doi.org/10.1002/acp.3911>