Journal of Experimental Criminology
https://doi.org/10.1007/s11292-020-09432-z

Does constructing a facial composite affect eyewitness memory? A research synthesis and meta-analysis

Colin G. Tredoux[1,2] · Siegfried L. Sporer[3] · Annelies Vredeveldt[1,4] · Kate Kempen[1] · Alicia Nortje[1]

Published online: 06 June 2020

Abstract

Objectives We conducted a meta-analysis to assess whether the construction of facial composites affects witnesses' lineup identification decisions.
Methods We located 23 studies (56 effects, 2276 participants). We consider effects of constructing composites on (a) correct identifications, and (b) incorrect identifications, from target-present lineups, and (c) incorrect identifications from target-absent lineups. Log odds ratio effect sizes were entered into a random-effects meta-analysis. We also present novel signal detection theory analyses in an online supplement.
Results There were no significant negative effects of composite construction, but some weak evidence that composite construction reduced incorrect identifications in target-present lineups. Because effect sizes showed little hetereogeneity for any of the outcomes (after outlier removal), there were no moderator analyses. Results for SDT measures also showed no effects.
Conclusions Empirical evidence suggests no effects of composite construction on identifications. We identify gaps in knowledge and make recommendations for more ecologically valid research.

Keywords Face composite · Eyewitness · Composite construction · Identification · Meta-analysis

Electronic supplementary material The online version of this article contains supplementary material, which can be downloaded at https://doi.org/10.1007/s11292-020-09432-z

✉ Colin G. Tredoux
colin.tredoux@uct.ac.za

[1] Department of Psychology, University of Cape Town, Rondebosch, Cape Town 7701, South Africa

[2] CLLE, Université de Toulouse, CNRS, UT2J, Toulouse, France

[3] Department of Psychology and Sports Science, University of Giessen, Otto-Behaghel-Strasse 10F, Giessen, Germany

[4] Department of Criminal Law and Criminology, Vrije Universiteit Amsterdam, De Boelelaan 1105, Amsterdam 1081 HV, Netherlands

Mistaken eyewitness identifications have played a role in many wrongful convictions discovered through DNA testing (Garrett 2011; Innocence Project 2018). One potential source of misidentifications may be the use of face composites as part of the investigative process. In cases in which identification is at issue, the police may ask eyewitnesses to produce a likeness (or "composite") of the perpetrator. By publishing the composite, investigators hope that the person depicted in the composite will be recognized and reported to the police (Davies and Valentine 2007). Police services frequently rely on face composites when investigating crimes: For example, in a 2005 study of a police service in a city with 3.5 million inhabitants, the total number of face composites produced was over 500 (Schmidt and Tredoux 2006).

Despite the common use of face composites, laboratory research has often reported that composites may frequently not resemble the perpetrator (for a review, see Davies and Valentine 2007). We should note, though, that Frowd et al. (2015) summarize data on new generation composite systems and provide evidence that resemblance between composites and targets may be improving. Composites that do not resemble the perpetrator well may contribute to mistaken identifications in several ways. A low-quality likeness might lead to the arrest of an innocent person because that person resembles the composite. This appears to have happened in the case of Kirk Bloodsworth (Junkin 2004), who was arrested in large part based on his resemblance to a composite of a suspected child murderer. Bloodsworth was sentenced to two life terms but was eventually exonerated after 9 years of imprisonment.

Although the role of composites in mistaken identifications in the USA is not clear, in 2018 the Innocence Project reported that 69% of 367 cases involved eyewitness misidentification, and 27% of these cases involved the use of a composite sketch (Innocence Project 2018). Of course, one cannot draw causal conclusions from case studies regarding their role in miscarriages of justice (Horry et al. 2014).

A particular way in which a face composite may contribute to mistaken identifications is through contamination of witness memory: that is, constructing a composite damages a witness's memory in some way, leading especially to a reduced ability to identify the perpetrator afterwards. Why might such an effect occur? There are several possibilities worth considering, and these may help us understand the results of our meta-analysis later in the article.

Several authors of studies that investigated the effect of constructing composites have taken heed of the tradition of work on post-event information and its effects on memory (see Davis and Loftus 2007), framing the construction and viewing of composites as an example of information that creates a competing memory or one that updates the original (Sporer 1996; Wells et al. 2005). In this review, we will focus on whether constructing a composite, rather than simply being exposed to one created by another witness, affects witness memory. Whereas the literature on composite construction is reasonably homogenous methodologically, making a meta-analysis feasible, the literature on composite exposure is heterogenous, and it is infeasible to conduct a meta-analysis on it. We are preparing a narrative review of the composite exposure literature, which could be considered a companion article to the present review (Anonymous 2019).

Although misinformation research focuses on factors that make verbal post-event suggestion a source of misinformation (see Davis and Loftus 2007), there are at least two paradigms of research on visual sources of misinformation: studies of mugshot exposure effects on identification and studies of "unconscious transference" (Deffenbacher et al. 2006). In both paradigms, witnesses mistake a face they have seen somewhere for the face of the perpetrator in a later identification task. This is a phenomenon that seems analogous to negative effects of constructing a composite, but there is an important difference: misinformation studies explicitly test whether memory mistakes can be induced by planting misinformation, whereas studies that test composite construction effects investigate if the processes involved in constructing face composites may either inadvertently hamper the ability of witnesses to identify the target later or may improve it.

A popular explanation for the interference of face likenesses on memory tasks is source monitoring failure (Johnson et al. 1993). The witness encodes both the original face and the intervening composite and at test has the problem of attributing competing memories to the appropriate source. If the source of the memories is not clear, a witness may make a decision based on a feeling of familiarity, and the intervening stimulus may seem more familiar because it occurs later in the timeline. Also, the composite may be viewed repeatedly; hence, its influence may be stronger than the initial encoding. One should note that mugshot interference tasks are different in an important sense from composite construction tasks, since interference tasks typically induce witnesses to identify an incorrect person from a mugshot album, and later identification may be due to a commitment effect, rather than a contamination effect (see Goodsell et al. 2009).

Another explanation for possible effects of constructing a composite, relative to not constructing a composite, is that of a cautious shift against making an identification (cf. Clare and Lewandowsky 2004). Composite construction may well make participants more conservative than participants who do not construct a composite; participants who construct a composite realize the difficulty of the task and begin to doubt their memory and thus doubt their ability to identify the target, decreasing their willingness to choose a member from the lineup. Such a strategy will likely appear as a decrease in the hit rate when the target is in the lineup. For instance, in experiment 1 in Wells et al. (2005), 58% of participants in the composite construction condition made no selection, compared with only 10% in the control condition. However, when participants are forced to choose, they may be able to select the target (Kempen 2009).

Effects of constructing a composite

Studies that have examined the effect of constructing composites on witness memory have produced equivocal results. One group of studies (Davies et al. 1978; Davis et al. 2016; Holland et al. 1994; Yu and Geiselman 1993) showed that constructing a face composite does not affect later identification performance. A second group of studies (Davis et al. 2014; Mauldin and Laughery 1981; McClure and Shaw 2002; Wogalter et al. 1989, experiment 1) reported a positive effect of face composite construction on lineup identification performance—participants who built a composite were more likely to recognize the target. In contrast, a third group of studies produced tentative evidence for a negative effect of face composite construction on identification performance (Comish 1987; Kempen 2012; Topp-Manriquez et al. 2016; Wells et al. 2005; Wogalter et al. 1989, experiment 2)—but only the results of Wells et al. were statistically significant. The effect reported by Wells et al. (2005, experiment 1) for hits in target-present (TP) lineups was extremely strong—an odds ratio of 47:1. Participants who provided a description were much more likely to identify the target (84%) than participants who had constructed a composite (10%). Wells et al. concluded that building a face composite reduces identification of the target from a later lineup.

It is quite evident, then, that the literature on the effects of constructing face composites is inconsistent. An earlier meta-analysis of studies on the effects of composite construction (Meissner and Brigham 2001), which was part of a larger meta-analysis of verbal overshadowing effects, concluded that constructing a composite has a significant *positive* effect on later identification performance. Participants who constructed a facial composite were 1.56 times more likely to make a correct positive identification in an identification task than control participants. However, the meta-analysis was conducted on a small corpus of studies, and predates the Wells et al. (2005) article.

Although the evidence does not seem clear on whether constructing composites affects witness testimony, it is worth noting that the idea that constructing a composite "contaminates" witness memory is now commonly cited as authoritative in articles in law journals and in some courts. An explicit admonition appears in the Wisconsin draft "model policy and procedure for eyewitness identification," asserting that "… the process of making a composite can damage an eyewitness's ability to identify the true perpetrator in a later lineup" (Attorney General of the State of Wisconsin 2009, p. 27). On the other hand, in the case of State v Henderson (2011), the court declared "… without more accepted research, courts cannot make a finding on the effect the process of making a composite has on a witness" (p. 67).

It is therefore of considerable importance to evaluate the literature that investigates effects of constructing face composites. We report a meta-analysis here of the relevant literature. When coding the data, we considered many methodological differences between studies, which may be potential moderators of effects. These include the system used to create the composite, the media used to display the original face and lineup, the instructions received before encoding, the exposure time to the original face, the delays between encoding and the composite and between the composite and the identification procedure, and the type of recognition task. We list the experimental variations that we coded for in Table 1. However, when we conducted the meta-analysis, we did not find convincing evidence for significant heterogeneity of effects, especially after removing two outlier effects, and we therefore did not follow up the initial analysis with a search for moderator effects. We do have a lengthy discussion of what could have been moderator variables, but because we did not conduct moderator analyses, we provide that discussion in the supplemental materials.

Table 1 Potential moderator variables coded for, with number of effect sizes available per coding category for categorical variables, and descriptive statistics for continuous variables

Variable	Number of studies		
	Hit	TP-foil	TA-FA
Composite system[a]			
Sketch artist	2	2	1
Identi-Kit	1	1	1
Photofit	2	0	0
Field Identification System	1	1	0
Mac-a-Mug Pro	2	2	0
Faces	9	9	6
Holistic (ID/EvoFit/E-Fit-V)	8	8	4
E-Fit	4	4	4
Medium of original stimulus			
Photo	12	10	6
Video/film	7	7	5
Slides	2	2	0
Live	2	2	1
Medium of lineup			
Photo	12	13	9
Video/film	7	5	3
Slides	2	3	0
Live	2	0	0
Encoding instructions			
Personality	8	8	5
Intentional	11	9	6
Incidental	2	2	0
Not specified	2	2	1
Exposure time			
Minimum seconds per face	2	2	3
Maximum seconds per face	180	180	180
Median seconds per face	16	20	18.5
Mean absolute deviation per face (s)	17.79	25.20	21.50
Delay between encoding and composite			
Minimum minutes	0	0	0
Maximum minutes	30,240	2880	2880
Median minutes	8	8	6.5
Mean absolute deviation (min)	6.67	6.67	4.45
Delay between composite and recognition			
Minimum days	0	0	0
Maximum days	35	35	35
Median days	2	2	2
Mean absolute deviation (days)	2.93	0.99	0

Table 1 (continued)

Variable	Number of studies		
	Hit	TP-foil	TA-FA
Lineup instructions			
Unbiased	20	20	12
Forced	2	0	0
Yes/no decision for each face	1	1	0
Number of lineup members[b]			
Minimum number	2	2	3
Maximum number	136	136	9
Median number	6	6	6
Mean absolute deviation	4.45	4.45	0

Hit identification of the target in a target-present lineup, *TP-foil* identification of a foil in a target-present lineup, *TA-FA* any positive identification in a TA lineup (false alarm)

[a] Three studies (Davis et al. 2014, experiment 1; Pike et al. 2018a; Sporer et al. 2016) compared two composite groups (one featural, one holistic) with the same control group. For those studies, we computed the average of the two composite groups for comparison with the control group, to avoid entering the same participants into the data file twice

[b] The minimum number does not represent a lineup, but a yes/no recognition task (McClure and Shaw 2002)

Dependent variables, hypotheses, and analytic plan

In our analysis, we considered three dependent measures: correct positive identifications of the target in TP lineups (hits), incorrect positive identifications of a foil from TP lineups (TP-foils), and incorrect positive identifications from TA lineups (false alarms; TA-FAs). For TA lineups, we were unable to distinguish between incorrect identifications of the "innocent suspect" and incorrect identifications of a foil, because most studies did not distinguish between these two types of misidentification (except for McClure and Shaw 2002, experiment 2). We added a fourth dependent measure, *d′* (*d* prime), for studies that reported both group-level hits and false alarms for conditions. Although authors may have reported group-level hits and false alarms, we needed estimates of the variance of *d′* values. We computed these using a bootstrapping method we improvised and a method originally proposed by Miller (1996) and implemented in R by Suero et al. (2017), with modifications. Because these analyses are exploratory, they are presented solely in the online supplemental material along with the R code used.

The key question of our meta-analysis is whether witnesses' construction of a composite of a previously encoded face will hamper or improve their subsequent ability to identify the face. We believe there is good reason from post-event misinformation studies to suggest that there will be an effect on later identification. However, we note that the effect can be expected to depend on several variables, as set out in our brief review of potential moderators. The three most important moderator types are likely to relate to the quality of encoding (a function of several factors, including exposure duration, and delays between encoding, composite construction, and recognition), quality of composite, and whether the perpetrator is present or absent in the identification task. The articles in our meta-analysis rarely measured the quality of encoding and the quality of composites explicitly, but there was considerable variation between studies in the conditions of encoding (e.g., in terms of exposure duration and encoding instructions) and in the conditions under which composites were produced (e.g., which composite system was used and how much time had passed before constructing the composite). Of course, since we are reporting a meta-analysis, we are at the mercy of the presence and distribution of these conditions across studies, which will affect our ability to draw firm conclusions.

Method

Inclusion/exclusion criteria

Studies were obtained by searching databases, including PsycInfo, MEDLINE, the *Social Sciences Citation Index*, and Google Scholar. To supplement the database search, we consulted the citation lists of all relevant manuscripts for additional sources and directly contacted researchers in the field, requesting unpublished manuscripts, conference proceedings, student dissertations, and other types of manuscript on this topic. Since the literature on facial composites is relatively small, we searched for all studies containing any of the following phrases in their title or abstract: "face composite," "facial composite," "identikit," "Identi-Kit," "photofit," "face reconstruction," and "face likeness." We enlarged our initial list by including additional studies from the reference lists of identified articles. We also used algorithms in the Thomson Reuters and Google Scholar databases to search for similar studies based on semantic similarity to our initial set.

To be included in the analysis, studies needed to employ the following general procedure: (1) During the encoding phase, participants were exposed to one or multiple faces, presented via photographs, via videos, or in person. (2) During the composite phase, participants in the experimental condition were asked to build a composite of the face(s) they had viewed during the encoding phase.[1] This condition was compared with a control condition in which participants did not construct a facial composite. (3) During the recognition phase, participants were asked to recognize the face(s) they viewed during the encoding phase. Most studies employed a lineup procedure, in which participants were exposed to the target face (or target replacement in the TA lineups) accompanied by a varying number of filler faces, whereas one study used a yes/no face recognition task (McClure and Shaw 2002).[2]

In cases in which the articles did not provide all the necessary data for each experimental condition, we contacted the authors to obtain it. Some studies had to be excluded, for instance, when the composite construction condition was manipulated as a within-participants factor (e.g., Bedillion 2017). This selection procedure resulted in a set of 56 effect size comparisons, drawn from 23 studies, reported in 14 articles (published and unpublished—marked with an asterisk in the reference list), involving 2276 participants in total.

Table 2 Study characteristics of all 23 studies, arranged by composite method

Authors	Program	Lineup bias[a]	Presentation medium	Test medium	Exposure time (s)	Encoding instruction[b]	Delay 1 (min)	Delay 2 (min)	Publication status
McClure and Shaw (2002), exp. 1)	Artist	Yes/no	Photo	Photo	3	Personality	3	5	Published
McClure and Shaw (2002), exp. 2)	Artist	Unbiased	Photo	Photo	3	Personality	9	9	Published
Davies et al. (1978), exp. 2)	Photofit	Forced	Photo	Photo	15	Intentional	1440	0	Published
Davies et al. (1978), exp. 2)	Photofit	Forced	Photo	Photo	15	Intentional	30,240	0	Published
Yu and Geiselman (1993)	Identi-Kit	Unbiased	Film	Film	7	Intentional	0	2880	Published
Wogalter et al. (1989, exp. 1)	FIS	Unbiased	Live	Slides	15	Intentional	0	0	Unpublished
Wogalter et al. (1989, exp. 2a)	Mac-a-Mug	Unbiased	Slides	Slides	4	Incidental	4	0	Unpublished
Wogalter et al. (1989, exp. 2b)	Mac-a-Mug	Unbiased	Slides	Slides	20	Incidental	4	0	Unpublished
Dumbell (2008)	Faces	Unbiased	Photo	Photo	2	Personality	15	2910	Unpublished
Wells et al. (2005, exp. 1)	Faces	Unbiased	Photo	Photo	180	Personality	8	2892	Published
Wells et al. (2005, exp. 2)	Faces	Unbiased	Film	Photo	21	Personality	8	2880	Published
Kempen (2009, exp. 1)	Faces	Unbiased	Photo	Photo	5	Personality	10	2880	Unpublished
Kempen (2012, exp. 2A)	Faces	Unbiased	Photo	Photo	5	Personality	10	2880	Unpublished
Maskow et al. (2007, exp. 1)	Faces	Unbiased	Photo	Photo	180	Personality	5	2880	Unpublished
Maskow et al. (2007, exp. 2a)	Faces	Unbiased	Photo	Photo	16	Personality	5	2880	Unpublished
Maskow et al. (2007, exp. 2b)	Faces	Unbiased	Photo	Photo	16	Personality	5	2880	Unpublished
Sporer et al. (2016)	Faces/ID	Unbiased	Photo	Photo	20	Personality	20	2880	Unpublished
Davis et al. (2014, exp. 1)	E-Fit/E-Fit-V	Unbiased	Film	Film	78		0	35	Published
Davis et al. (2014, exp. 2)	E-Fit-V	Unbiased	Film	Film	78		30	1914	Published
Davis et al. (2016)	E-Fit-V	Unbiased	Film	Film	60	Intentional	30	5	Published

Table 2 (continued)

Authors	Program	Lineup bias[a]	Presentation medium	Test medium	Exposure time (s)	Encoding instruction[b]	Delay 1 (min)	Delay 2 (min)	Publication status
Pike et al. (2018a)	E-Fit/E-Fit-V	Unbiased	Film	Photo	103	Intentional	2880	25,920	Published
Pike et al. (2018b, exp. 1)	E-Fit	Unbiased	Live	Photo	60	Intentional	12.5	50,400	Published
Pike et al. (2018b, exp. 2)	E-Fit	Unbiased	Film	Film	60	Intentional	5	12,240	Published

[a] Lineup bias: *Forced* participants required to choose someone, *Unbiased* participants could reject the lineup, *Yes/no* participants made a decision for each face: yes (seen before) or no (not seen before)

[b] Encoding instruction: *Intentional* participants notified that they would be tested later, *Incidental* participants not aware of the pending test, but no other special instruction given, *Personality* participants required to rate stimuli on personality traits during encoding

Coding procedure

The first and third authors completed all coding independently, following Wilson's (2009) recommendations for double-coding. All moderator variables are listed in Table 1, and Table 2 shows the moderator distribution over studies. Some variables were originally coded into fine-grained categories, which were collapsed into broader categories for the final analysis, to obtain larger cell sizes for comparisons.

Intercoder reliabilities for categorical variables were estimated using Cohen's kappa, as it controls for chance agreement. For continuous variables, we calculated the intraclass correlation coefficient (ICC(C, 1), Gamer et al. 2017), which takes systematic differences between coders into account and treats studies as a random effect (Orwin and Vevea 2009; Sporer and Cohn 2011). Overall, intercoder agreement was satisfactory, with almost all coefficients indicating either perfect agreement or good agreement. The lowest *kappa* value was .84 (corrected for maximum possible value) and all except two kappa values were greater than .90. The lowest ICC was .89 (log-corrected), and all other ICC values were greater than .90. There were thus few disagreements, and where they occurred, they were resolved among the coders by discussion.

Effect sizes

The effect size statistic of choice regarding *identification accuracy* (a binary dependent variable), when testing differences between proportions, is the odds ratio (Fleiss and Berlin 2009). The odds of an event are defined as the probability of an event (p) divided by the probability of the event not occurring ($1-p$). For example, when the target in a target-present lineup is correctly identified by 75% of the participants in the control group (pCG = .75), the odds of a correct identification in the control group are odds = (pCG/[1 − pCG]), that is, 3 to 1. If the target is only correctly identified in the experimental (composite construction) group by 50% (pEG = .50), the odds are (pEG/[1 − pEG]), that is, 1 to 1.

The odds ratio (OR) compares the two groups in terms of their relative odds:

$$\mathrm{OR} = (\mathrm{pCG}/[1{-}\mathrm{pCG}])/(\mathrm{pEG}/[1{-}\mathrm{pEG}]) = \\ (.75/.25)/(.50/.50) = 3.0/1.0 = 3.00,$$

that is, the chances of a correct identification are 3 times as likely in the control group compared with the composite construction group.

To obtain more precise estimates, raw frequencies were reconstructed, and calculations were based on those rather than the rounded proportions usually reported in publications. It is customary to perform all analyses on the natural log of the odds ratio (LOR) and to exponentiate (or "back transform") these to the original ratio, when interpreting (Fleiss and Berlin 2009; Lipsey and Wilson 2001). Table 3 provides an approximate transformation between different types of effect sizes.

For the current meta-analysis, weighted mean effect sizes and their inverse variance weights were calculated, following the recommendations by Cooper et al. (2009) and by Lipsey and Wilson (2001). We then conducted a random-effects meta-analysis, following recommendations outlined by Borenstein et al. (2009).

Table 3 Approximate transformations between effect sizes: odds ratio (OR), logged odds ratio (LOR), Cohen's *d*, and point-biserial r_{pb}

OR	LOR	Cohen's *d*	r_{pb}
1.00	0.00	0.00	0.00
1.50	*0.41*	*0.22*	*0.11*
2.00	0.69	0.38	0.19
2.50	*0.92*	*0.51*	*0.24*
3.00	1.10	0.61	0.29
3.50	1.25	0.69	0.33
4.00	1.39	0.76	0.36
4.50	*1.50*	*0.83*	*0.38*
5.00	1.61	0.89	0.41
5.50	1.70	0.94	0.43
6.00	1.79	0.99	0.44

"Small," "medium," and "large" effect sizes according to Cohen's (1988) recommendations are marked in italics (from Sporer & Martschuk, 2014, p. 12; reprinted with permission of the authors). The transformations to r_{pb} assume equal 50% base rates

After a preliminary search for outliers, the first goal in the analysis was to compute an appropriately weighted aggregate effect size, with an accompanying confidence interval: This allowed us to answer the question of the direction and magnitude of any effects due to constructing composites. The second goal of the analysis was to compute measures of the heterogeneity of the effect size (I^2 and Q, in particular). In the case of high homogeneity, the search for moderators is contraindicated. If one finds high heterogeneity, one should explore the reasons for this, including a search for moderator effects. Although many meta-analyses do this informally, often with the assistance of descriptive tables of results, such methods of "slicing and dicing" are considered too opportunistic by many authors—they leave other moderators uncontrolled, and report overlapping analyses, whose dependence is not modeled (Pigott 2012). We decided to use mixed linear meta-regression for potential moderator analysis, if indicated, as this controls for covariate moderators and for dependencies of effect sizes.

Calculations were performed independently by the first two authors, using the R language and environment for statistical computing (R Core Team 2018) and the R package metafor (Viechtbauer 2010) and with the macros provided by Lipsey and Wilson (2001). Results were virtually identical except that the mean ESs were slightly smaller in the analyses by macro. This difference may be because Lipsey and Wilson use the method of moments, while the method of fitting in metafor is restricted maximum likelihood (REML).

Results

We report separate meta-analyses for three dependent measures, namely, correct identifications of targets from TP lineups (hits), incorrect identifications of foils from TP lineups (TP-foils), and incorrect identifications of suspects or foils (false alarms) from TA lineups (TA-FAs). Descriptive summaries are given for each of the dependent measures in Table 4, per experimental and control group. We report aggregate proportions, weighted by sample sizes rather than by inverse variances, and caution readers that the aggregates we report in Table 4 are for summary purposes only and should not be interpreted as meta-analytic estimates of effect sizes.

Table 4 Proportions of hits and foil IDs in target-present lineups and false IDs (both foil and suspect IDs) in target-absent lineups in 23 studies

Authors	Target-present					Target-absent		
	Hits		Foil IDs		*N*	False IDs		*N*
	EG	CG	EG	CG		EG	CG	
Davies et al. (1978), exp. 2)	.80	.90			20			
Davies et al. (1978), exp. 2)	.40	.60			20			
Davis et al. (2014, exp. 1)	.67	.45	.14	.28	141	.57	.62	127
Davis et al. (2014, exp. 2)	.49	.35	.17	.31	198			
Davis et al. (2016)	.35	.32	.15	.32	67			
Dumbell (2008)	.46	.64	.25	.23	50			
Kempen (2009, exp. 1)	.44	.65	.07	.12	86	.09	.05	86
Kempen (2012, exp. 2A)	.69	.83	.06	.03	72	.03	.11	72
Maskow et al. (2007, exp. 1)	.98	1.00	.02	.00	120	.05	.08	120
Maskow et al. (2007, exp. 2a)	.77	.80	.07	.07	45	.27	.33	45
Maskow et al. (2007, exp. 2b)	.83	.93	.00	.00	45	.37	.53	45
McClure and Shaw (2002), exp. 1)	.85	.76	.14	.17	135			
McClure and Shaw (2002), exp. 2)	.78	.66	.13	.08	165	.45	.46	165
Pike et al. (2018a)	.83	.90	.05	.10	39	.19	.20	38
Pike et al. (2018b, exp. 1)	.70	.63	.10	.11	37	.25	.39	36
Pike et al. (2018b, exp. 2)	.58	.44	.16	.22	34	.17	.11	31
Sporer et al. (2016)	.68	.78	.05	.02	90			
Wells et al. (2005, exp. 1)	.10	.84	.30	.06	100			
Wells et al. (2005, exp. 2)	.18	.60	.20	.04	100	.26	.20	100
Wogalter et al. (1989, exp. 1)	.74	.60	.03	.06	58			
Wogalter et al. (1989, exp. 2a)	.78	.94	.10	.13	36			
Wogalter et al. (1989, exp. 2b)	.89	1.00	.16	.13	36			
Yu and Geiselman (1993)	.40	.37	.10	.30	47	.50	.31	49
Mean weighted effect sizes	.62	.66	.12	.14		.29	.30	

We used sample sizes as weights for the mean effect sizes, rather than inverse variances (see the text)

EG experimental group (created composite), *CG* no composite control group

Correct identifications from target-present lineups (hits)

Figure 1 reports effect sizes LOR (ln[odds ratio]) for hits included in our meta-analysis, per study. The analysis was based on $k = 23$ effects and $N = 1741$. The average LOR is -0.40 (SE = 0.25), 95% CI [-0.89, 0.10], which equates to an OR of 0.67, 95% CI [0.41, 1.11]. A normal theory test of whether this effect size differs from 1 (equally likely outcomes) was not significant ($Z = -1.58$, $p = .115$). In other words, the observed effect size is compatible with the view that constructing a composite does not affect hit rates.

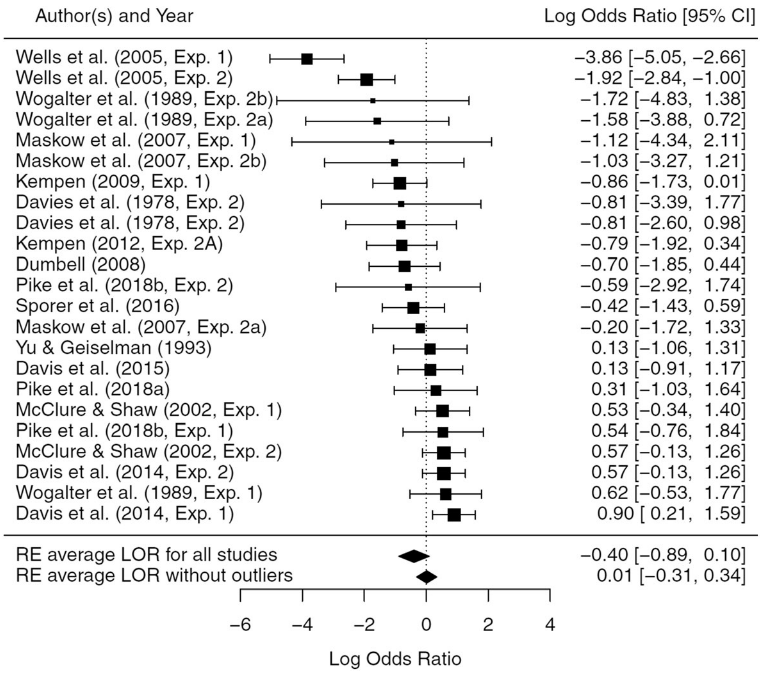

Fig. 1 Forest plot of LORs and their respective 95% CIs for correct identifications from target-present lineups (hits) of all individual studies as well as the weighted mean effect size with and without outliers. Negative values denote a decrease and positive values an increase in hits as a function of composite construction. Effects for which the confidence intervals included 0 were not significant. RE random effects

However, it is important to compute and test the heterogeneity of effect sizes, as high heterogeneity may indicate the presence of subgroups with different mean effect sizes. Indeed, measures of heterogeneity were high. Thus, $I^2 = 73.95\%$, 95% CI [50.98, 85.46], suggesting that the amount of variability in true effect size is between 51 and 85%. Similarly, Q was statistically significant, $Q(22) = 83.88$, $p < .001$. This pointed to the need for a moderator analysis, but when we explored Fig. 1, as well as model residuals, we noted several outlying effect sizes. Figure 1 shows that the 95% confidence interval around the average effect includes 0, which is reason for concluding that the overall effect is not significant. It is apparent from the figure that effect sizes vary from study to study, but with wide confidence intervals in most cases, suggesting greater random than systematic variability in effect size. Only one article, which contributes two effects, is significant in the direction suggesting contamination (Wells et al. 2005), and one study (Davis et al. 2014, experiment 1) is significant in the other direction.

Additionally, we examined the funnel plot of effects (Fig. 2) to assess potential "file drawer" problems. Most studies were distributed in a pyramid-like structure around the mean effect size, except that there appeared to be fewer studies with positive effects than would be expected, and two outliers with negative effects.

Finally, examination of model diagnostic statistics (studentized residuals, Cook's distances, and parameter deletion statistics; see Viechtbauer and Cheung 2010) suggested the presence of two cases that were both influential values and outliers. These were identified as the two effects reported in the article by Wells et al. (2005). We performed

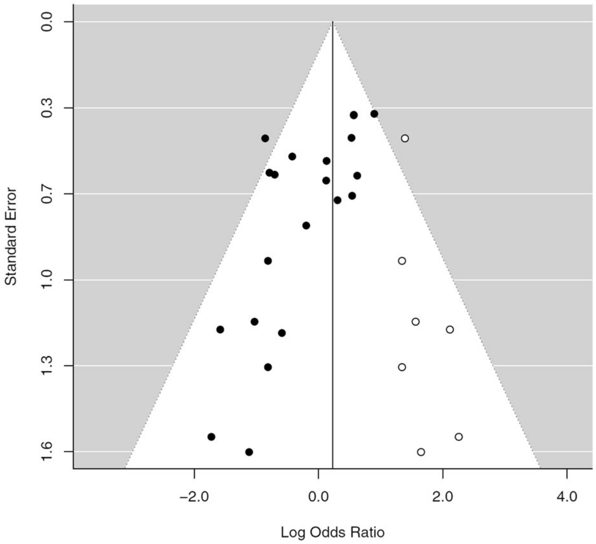

Fig. 2 Funnel plot of LORs for correct identifications from target-present lineups (hits), suggesting under-reporting of positive effects and presence of outlying effects. Open circles are imputed data, following Duval and Tweedy (2000)

a sensitivity analysis with and without Wells et al.'s (2005) effects and tested model adequacy, the funnel plot, and residual diagnostics. Removing the outliers improved the model considerably, and a further residual analysis suggested no remaining outliers.

On re-computing the main meta-analysis statistics, we found an average effect LOR = 0.014 (SE = 0.17), $k = 21$, $N = 1701$, which was not significantly different from zero ($Z = 0.08$, $p = .933$; 95% CI [− 0.31, 0.34]; OR = 1.01, 95% CI [0.73, 1.40]). The Q statistic for heterogeneity was much reduced and no longer significant, $Q(20) = 27.73$, $p = .116$. The I^2 statistic computed to 34.23%, which was well down from the 73.95% computed before removal of outliers. This set of results suggests that there is no need for a moderator analysis, since most of the initial heterogeneity in effect sizes was due to these two outliers.

In summary, a random-effects meta-analysis showed no significant effect of composite construction on the ability to identify a target from a TP lineup. There was initial evidence of heterogeneity in effect sizes, but after correcting for two outlying effects from one study, a test of heterogeneity was nonsignificant and the aggregate effect was close to zero.

Foil identifications from target-present lineups (TP-foils)

Figure 3 shows effect sizes for foil identifications in TP lineups, per study. The analysis was based on $k = 21$ effects and $N = 1701$. A random-effects analysis showed an average LOR = − 0.04 (SE = 0.23; 95% CI [− 0.48, 0.40]; OR = 0.96, 95% CI [0.62, 1.49]). A normal theory test of whether the LOR effect size differs from 0 (equally likely outcomes) was not significant ($Z = − 0.17$, $p = .862$). In other words, the observed effect size is compatible with the view that there is no effect of creating composites on foil identifications in TP lineups.

Standard measures suggested some heterogeneity, but not a great deal. Thus, I^2 = 37.74%, 95% CI [0, 62.65], suggesting that the amount of variability in true effect size is between 0 and 63%. *Q* evaluated to 28.49, which is not significantly different from 0, *Q*(20) = 28.49, *p* = .098. When we explored the plot of effect sizes (Fig. 3), as well as the model residuals, we noted that there appeared to be two potential outliers, although these were not as large as those we observed for the analysis of hits. Figure 3 also shows that only one study, which contributed two effects, was significant in the direction suggesting contamination (i.e., Wells et al. 2005).

Additionally, we examined the funnel plot of effects (Fig. 4), to assess potential file drawer problems. Almost all the studies appear to be contained within a pyramid-like structure around the mean effect size, suggesting a reasonable sampling of the domain and no systematic under-representation of negative or positive effects.

Finally, examination of residual scores (studentized residuals, Cook's distances, and parameter deletion statistics) suggested the presence of one clear outlying score and perhaps a second. These were identified as the two effects reported in the article by Wells et al. (2005). When we deleted these outlying effects, a further residual analysis suggested no remaining outlying scores. On re-computing the main meta-analysis statistics, we found an average effect LOR = − 0.34 (SE = 0.17), *k* = 19, *N* = 1501, which is just significantly different from zero (*Z* = − 1.97, *p* = .049; 95% CI [− 0.68, − 0.001]; OR = 0.71, 95% CI [0.51, 1.00]). The *Q* statistic for heterogeneity was not significant, *Q*(18) = 12.06, *p* = .844. The I^2 statistic computed to 0.54%. This set of

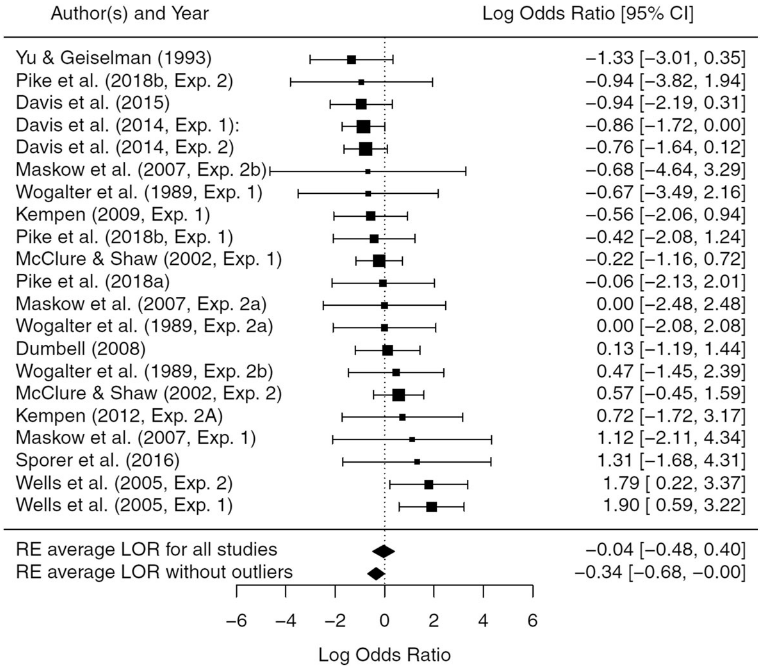

Fig. 3 Forest plot of LORs and their respective 95% CIs for foil identifications from target-present lineups (TP-foils), for all individual studies as well as the weighted mean effect size. Negative values denote a decrease and positive values an increase in foil identifications as a function of composite construction. Effects for which the confidence intervals included 0 were not significant

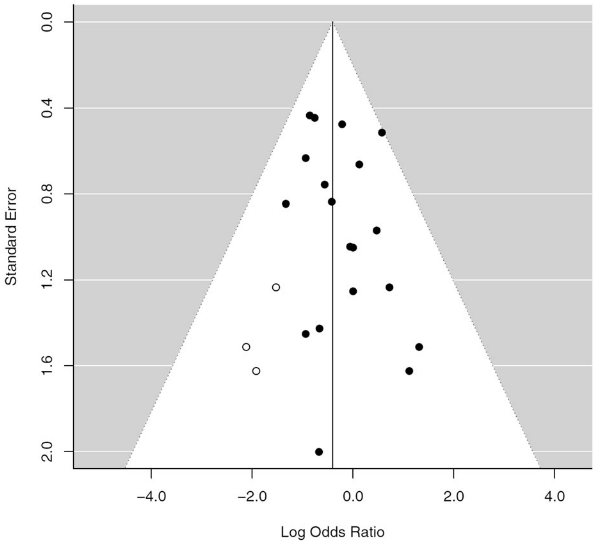

Fig. 4 Funnel plot of LORs for foil identifications from target-present lineups (TP-foils). Open circles are imputed data, following Duval and Tweedy (2000)

results suggests that there is little need for a moderator analysis, with some slight evidence that composite construction *reduces* the number of foil identifications that witnesses make from TP lineups.

In summary, a random-effects meta-analysis showed a nonsignificant, near-zero effect of composite construction on mistaken identification of a foil in a TP lineup. There was initial marginal evidence of heterogeneity in effect sizes ($p < .1$), but after correcting for one outlying effect, a test of heterogeneity was nonsignificant. The correction for the outlying effect resulted in a weak but significant effect of constructing a composite on subsequent identification of foils in a subsequent TP lineup. However, this effect was opposite to the hypothesized direction, showing that participants who constructed face composites were less likely to choose foils than participants in control conditions.

Incorrect identifications from target-absent lineups (TA-FAs)

We computed a meta-analysis of all positive identifications on TA recognition tasks (false alarms). The false identification of an innocent suspect is the most serious type of error a witness can make. Because most studies did not distinguish identifications of a foil and identifications of a designated innocent suspect from TA lineups, we were only able to investigate whether constructing a composite increased the chance that witnesses made any identification from a TA lineup, which is per definition incorrect.

Figure 5 reports effect sizes for false alarms in TA lineups, per study. The analysis was based on $k = 12$ effects and $N = 914$. A random-effects analysis showed an average LOR = − 0.07 (SE = 0.16; 95% CI [− 0.39, 0.25]; OR = 0.93, 95% CI [0.68, 1.28]). None of the individual effects reported were significant. A

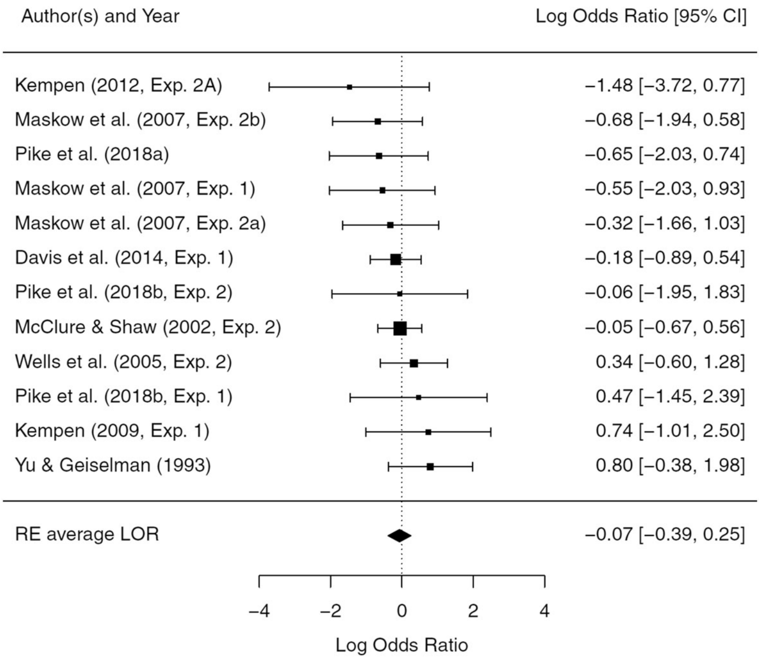

Fig. 5 Forest plot of LORs and their respective 95% CIs for incorrect identifications from target-absent lineups (TA-FAs), for all individual studies as well as the weighted mean effect size. Negative values denote a decrease and positive values an increase in false alarms as a function of composite construction. Effects for which the confidence intervals included 0 were not significant

test of whether the LOR differed from 0 was nonsignificant ($Z = -0.41$, $p = .679$). In other words, the observed effect size is compatible with the view that there is no effect of creating composites on false alarms in TA lineups. Standard measures of heterogeneity suggested little heterogeneity: Q evaluated to 7.65, which is not significantly different from 0, $Q(11) = 7.65$, $p = .744$; $I^2 = 0\%$. When we explored the plot of effect sizes (Fig. 5), we did not note influential cases, or outliers, or indeed any whose 95% confidence interval exceeded 0. In other words, no studies showed a significant effect on false alarms.

Additionally, we examined the funnel plot of effects (Fig. 6) to assess potential file drawer problems. All studies appear to be contained within a pyramid-like structure around the mean effect size, suggesting a reasonable sampling of the domain and no systematic under-representation of negative or positive effects.

Finally, examination of residual scores suggested the presence of one influential value. This was identified as an effect reported in the article by McClure and Shaw (2002), experiment 2). Although the McClure and Shaw study had high scores on three indicators of influence, the residual effect size was well within the expected range, and the study was thus not an outlier in terms of effect size, but rather exerted a strong influence on the model. There was thus no reason to remove it, but a sensitivity analysis with and without the effect size in question showed that very little changed when it was omitted (average LOR = − 0.07, SE = 0.19; 95% CI [− 0.45, 0.30]; OR = 0.93, 95% CI [0.64, 1.35]).

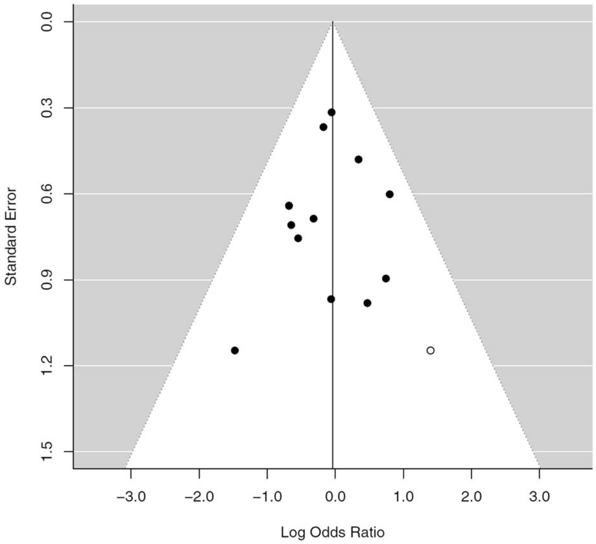

Fig. 6 Funnel plot of LORs for incorrect identifications (false alarms) from target-absent lineups (TA-FAs). Open circles are imputed data, following Duvall and Tweedy (2000)

In summary, a random-effects meta-analysis showed a nonsignificant, near-zero effect of composite construction on mistaken identification of a foil in a TA lineup. Further, there was little evidence of heterogeneity in the effect sizes.

Discussion

Our meta-analyses do not show a significant effect of creating a composite on subsequent lineup identifications for any of the outcome variables we considered: identifications of targets from TP lineups, or misidentifications of foils as targets, in either TP or TA lineups. Although an initial analysis of correct identifications from TP lineups showed considerable variation in effect sizes across (and within) studies, suggesting the presence of subgroups of studies with either positive or negative effects, this was due to one study (Wells et al. 2005), which reported two strong negative effects of composite construction. Removal of that study from the meta-analyses showed that the aggregate effect size for each of the outcome variables was near-zero, with little heterogeneity in effect size.

Our results are surprising. When witnesses produce a face composite, they are self-constructing post-event information, which, if the composite is poor, one might expect to hamper later attempts at identification (and conversely, if it is good, this might aid later identification attempts). A possible outcome, when the composite is poor—which is what extant literature suggests composites typically are—is a decreased rate of guilty-suspect identifications in TP lineups and an increased rate of innocent-suspect identifications in TA lineups. However, this is not what we found in our meta-analysis of studies.

Before we discuss reasons for the surprising absence of effects of constructing a composite on witness memory, we note that there is an important exception, in that one study found a strong negative effect for constructing a face composite (Wells et al.

2005). Eight different tests to determine outlying values within the random-effects analysis modeling framework suggested that the Wells et al. (2005) effect sizes were likely to be outliers that should not be considered in a meta-analysis. We searched for a methodological explanation for why the study might be unusual. We can only offer some speculative possibilities. First, as acknowledged by the authors, the stimulus faces in experiment 1 were identical at study and test, which may explain the high hit rate in the control condition (84% accuracy, vs. 10% in the composite construction condition). It has become unusual to use identical stimulus faces across conditions in face recognition experiments, since this confounds picture and face memory (see Bruce 1982). Second, this study differs from all the other studies in our corpus, except one (Maskow et al. 2007), in terms of the encoding time afforded participants. In Wells et al.'s study, witnesses were required to spend 180 s encoding each face, judging 10 personality traits, whereas the median encoding time across studies was 16 s. It is notable that the control group in Wells et al.'s experiment 1 achieved 84% recognition accuracy in the TP condition and that the experimental group achieved only 10% accuracy; it might be that the lengthy encoding time boosted the control group's performance but did not do the same for the experimental group. We note that Maskow et al. conducted a close replication of this experiment, in which the same exposure time of 180 s was used. They similarly found not only very high control group performance but also very high experimental group performance in the TP condition (100% and 98% recognition accuracy, respectively).

Ultimately, we cannot explain why the study by Wells et al. (2005) found such different results even though their method was similar in many ways to other studies in the corpus. We do know that one unpublished attempt to replicate their study closely was not successful (Maskow et al. 2007). An anonymous reviewer has suggested that the effect sizes reported by Wells et al. (2005) are implausible, and we think that such a conclusion may be warranted on the basis of the meta-analytic evidence we have gathered.

Potential moderators

We argued earlier that three critical factors are likely to moderate effects of constructing face composites: quality of encoding, quality of composite, and whether the perpetrator is present or absent in the identification task. Our meta-analysis revealed considerable variation between studies in the conditions under which the faces were encoded and the composites produced. There were also a number of studies that included both TP and TA conditions. Very few studies, however, reported information about the similarity of foils to targets or to suspects, and very few studies included information on the quality of the composites created and their relation to identification performance. In the study by Holland et al. (1994), persons who had constructed better composites, as determined by a matching task of independent raters, recognized more targets (84.4%) correctly than participants who had constructed poorer resemblances (67.5%). One normally searches for moderator effects when effect sizes exhibit considerable heterogeneity. Our analysis found no significant heterogeneity of effect sizes once outlying effects had been removed, for each of our dependent measures. This does not definitively indicate that there is no heterogeneity between effect sizes in our corpus but does suggest that an attempt to identify them formally would be unsuccessful. Hence, we are unable to specifically test the theoretical predictions we made about potential moderators. This does not mean that the theoretical predictions are wrong, just that the current literature does not provide a basis to test them.

Possible reasons for the absence of composite construction effects

What are some reasons for the absence of an aggregate effect? First, it might be that the quality of studies in this area has not been sufficiently high to detect the effect. Under-powered studies might have been unable to detect effects that were in fact present. It is also clear that several of the studies are unpublished and/or are from student theses, and it is fair to say that unpublished, unreviewed work is likely to be of lower quality than published work that has withstood scrutiny by peers. We do not think that study quality can explain the lack of an aggregate effect, though. Six of the studies we reviewed had larger sample sizes than that of Wells et al., and if we assume the same population effect size and population variability in effect size, would likely have had equivalent or greater statistical power. Of the 23 effect sizes reported in Table 2, 10 were harvested from unpublished studies and the remaining 13 from studies published in peer-reviewed journals. It does not seem likely that all studies except one are of poor quality, and our inspection of the manuscripts does not suggest that. Some of the unpublished studies in Table 2 have been submitted for publication but may have been rejected due to a bias against null findings (Cooper 2010). Therefore, most meta-analysts try particularly hard to locate unpublished manuscripts for this reason.

Second, we have argued that the quality of the composite is likely an important moderator of any effects due to constructing face composites. That is, if a composite does not resemble the face of the perpetrator, this may constitute misleading information and may compete at the time of retrieval with the originally encoded face memory or even overwrite it. It may be that in the studies we reviewed the composites were not typically misleading. The problem with assessing this explanation is that we cannot tell from the original studies what the quality of composites was. Researchers have not usually assessed the quality of composites or, where they have, have made it difficult to aggregate or compare across studies.

Third, whereas source monitoring failure is a likely explanation for misinformation effects in general (cf. Johnson et al. 1993), it may not apply in the case of composites. For instance, when mugshots are viewed between encoding and retrieval, witnesses may have access to the original face memory and the face memory derived from the mugshot but be confused about where they encoded the respective memories and may thus end up opting for the latter since its trace is stronger. A similar argument could be made for exposure to a misleading composite that the witness has created himself: The witness mistakes the memory of the composite they created for the original face they encoded. However, because most composites do not resemble real faces—at least not in the set of studies we have reviewed, which mostly involve construction systems developed before 2005—they would not pass as plausible alternatives for the original face. In the misinformation literature, it is well established that misinformation is unlikely to have an effect when witnesses detect discrepancies between the misinformation and the original stimulus (Tousignant et al. 1986). Thus, detection of the discrepancy between a memory of a face and a memory of a composite likely compensates for possible source monitoring failures.

There may also be several aspects of the composite construction process used in the studies under review that make effects less likely than one might first assume. It is well established in studies of misinformation and false memory that levels of false recall and false recognition are lower following visual presentation than other modes of presentation, and this is in fact a prediction from opponent-process theory (Brainerd and Reyna 2005). The compositing process studied in the literature we have reviewed involves visual exposure to original information (the face of a target) and visual exposure to self-constructed post-event information (the composite image). Many of the classic demonstrations of the misinformation effect rely on the presentation of stimuli within a narrative or story (e.g., Loftus and Palmer 1974), probably because narratives encourage reliance on familiarity or gist memory (Garry and Wade 2005). Composite studies do not have a clear narrative structure in the same sense: Usually what is at issue is just the memory for a single face.

Study space analysis

Meta-analyses are important aides to taking stock of a literature, by quantifying effects in it, as well as sources of variation in effect sizes across studies. However, meta-analyses are also significantly confined by the range of variables and conditions studied in the literature they review. Malpass et al. (2008) introduced the concept of a “study space” to capture this important limitation of review methods (including meta-analysis). The key idea in a study space analysis is to evaluate a corpus of studies according to how effectively and rigorously the studies have sampled a domain of inquiry. We do not attempt a full-blown study space analysis here but instead identify some limitations of the research to date, based on our knowledge of the practices and exigencies of constructing face composites in law enforcement.

Composite system

There is evidence that the new generation of holistic systems, typically based on statistical models of face variation, seems to be better at face reconstruction than earlier systems (Frowd et al. 2015; Tredoux et al. 2006). Newer studies also attempt to improve interviewing techniques to align the cognitive processes involved in recalling a face and composite construction (Skelton et al. in press). Although new generation systems constitute a minority of the studies in the literature on composite construction effects, law enforcement in the UK and in some other countries are moving to these new systems, ahead of the literature on composite construction effects. The three studies in our sample that directly compared the featural (older) and holistic composite (newer) construction systems revealed no significant differences (Davis et al. 2014, experiment 1; Pike et al. 2018a; Sporer et al. 2016), in terms of composite construction effects. We clearly need to know more about these systems, but the available evidence does not suggest that they contaminate witness memory.

Delays

The delay between observing a perpetrator and constructing a likeness of their face with a composite system is in police practice usually quite significant—certainly 24 to 48 h at a typical minimum. However, in laboratory studies, it is inconvenient to implement such long delays, and as such, few studies have investigated delay in a realistic manner, preferring instead to engage participants in the reconstruction within minutes of having witnessed the simulated crime. The literature on delay effects on eyewitness memory (see Deffenbacher et al. 2008), however, makes it clear that memory after very brief delays is quite different from memory after significant delays. The nonlinearity of the relation between delay and memory fidelity means that memory after 24 or 48 h could be dramatically worse than immediately after encoding the stimulus. This is likely to mean that composites constructed in laboratory studies are more likely to be accurate than in natural environments, but without conducting appropriate studies, this must be considered speculation.

Expertise of the operator

Another notable shortcoming of most of the studies we reviewed is that they typically used student participants to create composites directly with software programs and usually with very little training. In contrast, most law enforcement agencies have trained and experienced police operators construct the composite interactively with witnesses. It is not clear whether this practice would make composites more accurate or whether it would lead to contamination, but it certainly warrants further investigation.

Quality of the composite

An important question about a composite concerns its quality, that is, the resemblance between the composite and the target (a question the police appear to ask routinely). It is likely that the quality of the composite moderates the effect of the composite on identification accuracy, but we presently know little about it. Composite quality can be assessed in two major ways: self-ratings of likeness to a target by witnesses creating the composite and ratings of similarity to a target by "mock witnesses." The former is perhaps more akin to a confidence rating by the witness in the quality of the composite and is clearly not an independent estimate of the quality of the composite. That is, a witness with a poorly encoded memory might not be able to give a good indication of the accuracy of the composite. However, witnesses with better-encoded memories ought to be able to do this, and since they likely have unique information about the identity of the target or perpetrator, their rating may be useful to consider.

A related question is whether eyewitnesses' confidence in their lineup decisions is affected by composite construction or exposure. Unfortunately, most studies in the literature under consideration do not disaggregate confidence-accuracy correlations according to whether participants chose someone from the lineup (see e.g., Sporer et al. 1995). Clearly, more data are needed to explore this further.

Similarity relations within lineups

The similarity of lineup foils to the suspect and the similarity of the suspect to the perpetrator are two other important potential (but rarely studied) moderators of composite construction effects. Although there is some evidence in the literature that these variables are important (e.g., Franzen and Sporer 1994a, 1994b; Sporer 1996), questions about the effects of foil-suspect-target similarity have not been studied systematically. It is clear from several lines of enquiry that similarity relations within lineups are indeed strongly predictive of witness identification accuracy (Fitzgerald et al. 2013; Tredoux 2002).

Stress

Vredeveldt et al. (2015) worked closely with law enforcement in a field study of face composite operators and interview methods. One of the striking observations in that study, emphasized repeatedly by police officers, was the degree of anxiety and stress witnesses exhibited while attempting to create face composites. Police officers reported having to postpone interviews, having witnesses break down with “uncontrollable crying,” and other debilitating conditions. Most witnesses who construct face composites have themselves been the victims of crimes against their persons (e.g., assault, rape, theft of property on their persons, home invasion burglaries), and many believe that it will be dangerous for them to attempt a later lineup identification. Creating a face composite is thus often a highly anxiety-provoking experience for witnesses. Like most other studies in the eyewitness literature, this degree of anxiety and stress has not been well simulated in studies examining the effect of composite construction on eyewitness identification. Where studies have been able to inculcate high levels of stress and study its effects on eyewitnesses, the evidence shows that this negatively impacts later recognition (e.g., Deffenbacher et al. 2004; Valentine and Mesout 2009). This underscores the continuing importance of finding ways of studying eyewitness identification under high-stress conditions, including the possible moderating role of stress in composite construction effects.

Joint construction of composites by witnesses

Some law enforcement agencies use multiple witnesses to construct a joint composite of the face they saw (cf., Vredeveldt et al. 2015). It is not clear whether this practice increases the risk of misinformation or whether it will yield better composites, since witnesses may be more likely to correct each other’s errors (e.g., Vredeveldt et al. 2017). To our knowledge, there are no studies that assess the effect of joint construction on the quality of the composite and subsequent eyewitness identifications.

Multiple-perpetrator composites

In general, eyewitness research has focused almost exclusively on scenarios in which a crime is committed by a single perpetrator. In cases with multiple perpetrators, witnesses may have to construct multiple composites, and it is possible that source confusion may affect the construction of composites. This could be a significant source of effects on witness memory, but one on which there are yet no data.

Practical implications

An important lacuna in the literature is the absence of field studies. Many aspects of real-life police investigations have not yet been studied in the context of composite construction. For instance, eyewitnesses may have their own personal copy of the composite they have produced and may review it many times (a form of rehearsal), which could increase the likelihood of contamination but may also preserve or improve memory if the composite is a good likeness. Additionally, witnesses may receive feedback from police, which could inflate their confidence in the composite and thus exacerbate negative composite construction effects (cf. post-identification feedback effects; e.g., Steblay et al. 2014).

Conclusion

The current meta-analysis produced what some may consider a surprising result—creating a facial composite generally did not affect subsequent identification performance. We emphasize that this conclusion is tentative: the literature we reviewed is quite small, consisting of 23 studies. Moreover, the ecological validity of studies to date has generally been low and some important variables have not yet been studied sufficiently. At this point in time, the body of work in this area does not allow for solid, evidence-based policy recommendations regarding the use of facial composites in police investigations.

One might expect there to be a clear effect of composite construction on identification performance, given well-established findings that misinformation presented to witnesses after an event, and before retrieval, can have significant adverse consequences. We think that part of the reason for this surprising result has to do with the way extant studies have been constructed: Work in the eyewitness and memory literatures suggests that encoding quality, composite quality, and the structure of the identification task (target presence and suspect-foil-target similarity relations) ought to be important determinants of whether there is an effect of constructing a composite. Unfortunately, few studies have manipulated these variables directly and never in concert. The important variable of composite quality has rarely been measured. The finding of no or little effect of constructing composites thus does not contradict well-established findings about the hazards of post-event misinformation.

However, as we have indicated in our discussion, there are also reasons for expecting that composite construction might not have contaminating effects on witness memory: We do not expect source misattribution, and we expect discrepancy detection to be heightened by the long encoding that typically happens at the time of constructing a composite, together with the evident fact that composites look different from real faces and are consciously perceived as such.

We make one final point, namely, that the studies in this area were designed to answer a practical, not a theoretical question: If one simulates an eyewitness crime in a laboratory, and gets witnesses to construct a composite afterwards, will their later identification be affected adversely? If we assume that the variety of encoding conditions and composite construction conditions has been representative and the lineup tasks have been fair, then it seems reasonable to conclude that the practice of asking witnesses to construct composites has little adverse consequence on their ability at an identification task. At the same time, we can reasonably conclude that we do not know what the boundary conditions are of possible composite construction effects and that we need well-structured experimental studies to rigorously determine these.

Notes

1. Note that in one study, participants constructed faces from memory by sketching them (McClure and Shaw 2002). This is similar to studies in which sketches were drawn by a police operator, except that they were unsupervised in this instance. Many other studies also allowed participants to construct composites unsupervised by police operators (e.g., Wells et al. 2005).
2. We computed an effect size for the study by McClure and Shaw (2002) by computing the total proportion of accurate choices from other study information and calculating the log odds ratio. This seemed a better option than estimating an effect size from the means and standard deviations reported by the authors, since such estimation formulae are known to be inaccurate when the underlying distribution is not normal (Sánchez-Meca, Marín-Martínez, & Chacón-Moscoso, 2003).

Acknowledgments This article was made possible with the support from the South African Centre for Digital Language Resources (SADiLaR). SADiLaR is a research infrastructure established by the Department of Science and Technology of the South African government as part of the South African Research Infrastructure Roadmap (SARIR). The research was also supported by two grants to the second author: Recognizing Faces of Other Ethnic Groups by the Deutsche Forschungsgemeinschaft (German Science Foundation): Grant Nr. Sp272/9-1) and Reducing False Identifications of Other-race People through new Computer Techniques (SUA 07/028) by the National Research Foundation of South Africa and the German Ministry of Science and Education.

References

Note: studies with asterisks (*) are included in the meta-analysis.

Anonymous (2019). Does exposure to facial composites damage eyewitness memory? A comprehensive review. Manuscript in preparation.

Attorney General of the State of Wisconsin. (2009). Model policy and procedure for eyewitness identification. Retrieved from http://www.doj.state.wi.us/sites/default/files/2009-news/eyewitness-public-20091105.pdf

Bedillion, C. N. (2017). The effects of the verbal overshadowing effect and independently creating composite sketches on eyewitness identification accuracy. (B.Sc. Thesis), Allegheny College Meadville, Pennsylvania, USA. Retrieved from http://hdl.handle.net/10456/45385

Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Chichester: John Wiley & Sons, Ltd. https://doi.org/10.1002/9780470743386.

Brainerd, C. J. (2005). The science of false memory. In V. F. Reyna (Ed.), *Oxford University Press*. https://doi.org/10.1093/acprof:oso/9780195154054.001.0001.

Bruce, V. (1982). Changing faces: visual and non-visual coding processes in face recognition. *British Journal of Psychology, 73*, 105–116. https://doi.org/10.1111/j.2044-8295.1982.tb01795.x.

Clare, J., & Lewandowsky, S. (2004). Verbalizing facial memory: criterion effects in verbal overshadowing. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30*(4), 739–755. https://doi.org/10.1037/0278-7393.30.4.739.

Comish, S. E. (1987). Recognition of facial stimuli following an intervening task involving the Identi-kit. *Journal of Applied Psychology, 72*, 488–491. https://doi.org/10.1037/0021-9010.72.3.488.

Cooper, H. (2010). *Research synthesis and meta-analysis. A step-by-step approach*. Los Angeles: Sage.

Cooper, H., Hedges, L. V., & Valentine, J. C. (2009). *The handbook of research synthesis and meta-analysis*. New York: Russel Sage Foundation.

*Davies, G. M., Ellis, H. C., & Shepherd, J. (1978). Face identification: The influence of delay upon accuracy of photofit construction. *Journal of Police Science & Administration, 6*, 35–42.

Davies, G. M., & Valentine, T. (2007). Facial composites: forensic utility and psychological research. In R. C. L. Lindsay, D. F. Ross, J. D. Read, & M. P. Toglia (Eds.), *The handbook of eyewitness psychology, Vol II: memory for people* (pp. 59–83). Mahwah: Lawrence Erlbaum Associates Publishers.

Davis, D., & Loftus, E. F. (2007). Internal and external sources of misinformation in adult witness memory. In R. C. L. Lindsay, D. F. Ross, J. D. Read, & M. P. Toglia (Eds.), *The handbook of eyewitness psychology, Vol I: memory for events* (pp. 195–237). Mahwah: Lawrence Erlbaum Associates Publishers.

*Davis, J. P., Gibson, S., & Solomon, C. (2014). The positive influence of creating a holistic facial composite on video line-up identification. Applied Cognitive Psychology, 28, 634–639. https://doi.org/10.1002/acp.3045.

*Davis, J. P., Thorniley, S., Gibson, S., & Solomon, C. (2016). Holistic facial composite construction and subsequent lineup identification accuracy: comparing adults and children. The Journal of Psychology, 150, 102–118. https://doi.org/10.1080/00223980.2015.1009867.

Deffenbacher, K. A., Bornstein, B. H., McGorty, E. K., & Penrod, S. D. (2008). Forgetting the once-seen face: estimating the strength of an eyewitness's memory representation. *Journal of Experimental Psychology: Applied, 14*, 139–150. https://doi.org/10.1037/1076-898x.14.2.139

Deffenbacher, K. A., Bornstein, B. H., & Penrod, S. D. (2006). Mugshot exposure effects: retroactive interference, mugshot commitment, source confusion, and unconscious transference. *Law and Human Behavior, 30*(3), 287–307. https://doi.org/10.1007/s10979-006-9008-1.

Deffenbacher, K. A., Bornstein, B. H., Penrod, S. D., & McGorty, E. K. (2004). A meta-analytic review of the effects of high stress on eyewitness memory. *Law and Human Behavior, 28*, 687–706. https://doi.org/10.1007/s10979-004-0565-x.

*Dumbell, E. (2008). Face composite production effects on witness memory: Case not closed. (Unpublished honor's thesis), University of Cape Town.

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

Fitzgerald, R. J., Price, H. L., Oriet, C., & Charman, S. D. (2013). The effect of suspect-filler similarity on eyewitness identification decisions: a meta-analysis. *Psychology, Public Policy, and Law, 19*(2), 151–164. https://doi.org/10.1037/a0030618.

Fleiss, J. L., & Berlin, J. A. (2009). Effect sizes for dichotomous data. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 237–253). New York: Russell Sage.

Franzen, S., & Sporer, S. L. (1994a). Personenverwechslungen durch irrefuehrende Rekonstruktionsbilder: Zum Einfluss nachtraeglicher Informationen und der Wiederherstellung des Wahrnehmungskontextes (Person mixups as a function of misleading composites: on the influence of postevent information and context reinstatement). In S. L. Sporer & D. Meurer & (Eds.), *Die Beeinflussbarkeit von Zeugenaussagen* (Influencing eyewitness testimony) (pp. 207–236). Marburg: N.G. Elwert.

Franzen, S., & Sporer, S. L. (1994b). Personenverwechslungen und Moeglichkeiten ihrer Vermeidung: Koennen Augenzeugen durch Visualisierung gegen den Einfluss von irrefuehrenden Rekonstruktionsbildern immunisiert werden (Person mixups and possible countermeasures: can eyewitnesses be inoculated against misleading composites through visualization)? In S. L. Sporer & D. Meurer (Eds.), *Die Beeinflussbarkeit von Zeugenaussagen (Influencing eyewitness testimony)* (pp. 237–283). Marburg: N.G. Elwert.

Frowd, C. D., Valentine, T., & Davis, J. P. (2015). Facial composites and techniques to improve image recognizability. In T. Valentine & J. P. Davis (Eds.), *Forensic facial identification: theory and practice of identification from eyewitnesses, composites and CCTV* (pp. 43–70). Chichester: Wiley-Blackwell.

Gamer, M., Lemon, J., & Singh, I. F. P. (2017). irr: Various coefficients of interrater reliability and agreement. R package version 0.84. https://CRAN.Rproject.org/package=irr

Garrett, B. (2011). *Convicting the innocent*. Cambridge: Harvard University Press. https://doi.org/10.4159/harvard.9780674060982.

Garry, M., & Wade, K. A. (2005). Actually, a picture is worth less than 45 words: narratives produce more false memories than photographs do. *Psychonomic Bulletin & Review, 12*(2), 359–366. https://doi.org/10.3758/bf03196385.

Goodsell, C. A., Neuschatz, J. S., & Gronlund, S. D. (2009). Effects of mugshot commitment on lineup performance in young and older adults. *Applied Cognitive Psychology, 23*(6), 788–803. https://doi.org/10.1002/acp.1512.

Holland, K., Otzen, H., & Sporer, S. L. (1994). Der Einfluss der Rekonstruktion und der Beschreibung von Gesichtern auf das spätere Wiedererkennen (The influence of constructing a face composite and describing a face on later face recognition). In S. L. Sporer & D. Meurer (Eds.), *Die Beeinflussbarkeit von Zeugenaussagen (Influencing eyewitness testimony)* (pp. 154–206). Marburg: N. G. Elwert.

Horry, R., Halford, P., Brewer, N., Milne, R., & Bull, R. (2014). Archival analysis of eyewitness identification test outcomes: what can they tell us about eyewitness memory? *Law and Human Behavior, 38*, 94–108. https://doi.org/10.1037/lhb0000060.

Innocence Project (2018). Eyewitness misidentification. Retrieved August 13, 2018, from https://www.innocenceproject.org/dna-exonerations-in-the-united-states/

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.

Junkin, T. (2004). *Bloodsworth: the true story of the first death row inmate exonerated by DNA*. Chapel Hill: Algonquin.

*Kempen, K. (2009). “It’s the thought that counts”: face composite production can hamper recognition performance. (Unpublished honour’s thesis). University of Cape Town.

*Kempen, K. (2012). “What big teeth you have” – Red riding hood and the face recognition failure: the effects of isolated featural and configural composite construction on recognition accuracy. (Unpublished master’s thesis). University of Cape Town.

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. London: Sage Publications.

Loftus, E. F., & Palmer, J. C. (1974). Reconstruction of automobile destruction: an example of the interaction between language and memory. *Journal of Verbal Learning and Verbal Behavior, 13*(5), 585–589.

Malpass, R. S., Tredoux, C. G., Compo, N. S., McQuiston-Surrett, D., MacLin, O. H., Zimmerman, L. A., & Topp, L. D. (2008). Study space analysis for policy development. *Applied Cognitive Psychology, 22*(6), 789–801.

*Maskow, M., Schmidt, H., Tredoux, C. G., & Nunez, D. T. (2007). Face composite production does not affect eyewitness identification accuracy. Unpublished manuscript, University of Cape Town.

Mauldin, M. A., & Laughery, K. R. (1981). Composite production effects on subsequent facial recognition. *Journal of Applied Psychology, 66*, 351–357. https://doi.org/10.1037/0021-9010.66.3.351.

McClure, K. A., & Shaw, J. S. (2002). Participants’ free-hand drawings of a target face can influence recognition accuracy and the confidence–accuracy correlation. Applied Cognitive Psychology,* 16, 387–405. doi:https://doi.org/10.1002/acp.802.

Meissner, C. A., & Brigham, J. C. (2001). A meta-analysis of the verbal overshadowing effect in face identification. *Applied Cognitive Psychology, 15*, 603–616. https://doi.org/10.1002/acp.728.

Miller, J. (1996). The sampling distribution of *d*. *Perception & Psychophysics, 58*, 65–72. https://doi.org/10.3758/bf03205476.

Orwin, R. G., & Vevea, J. L. (2009). Evaluating coding decisions. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 177–205). New York: Russell Sage.

Pigott, T. D. (2012). *Advances in meta-analysis*. New York: Springer. https://doi.org/10.1007/978-1-4614-2278-5.

*Pike, G. E., Brace, N. A., Turner, J., & Vredeveldt, A. (2018a). Comparing the effects of feature-based and holistic facial composite systems on eyewitness identification accuracy. *Manuscript under review*.

*Pike, G. E., Brace, N. A., Turner, J., & Vredeveldt, A. (2018b). The effect of facial composite construction on eyewitness identification accuracy in an ecologically valid paradigm. Criminal Justice and Behavior, Advance online publication. doi:https://doi.org/10.1177/0093854818811376.

R Core Team (2018). R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Schmidt, H. C., & Tredoux, C. G. (2006). Utilisation and usefulness of face composites in the South African Police Service - an evaluation study. *South African Journal of Criminal Justice, 19*, 303–314.

Skelton, F. C., Hancock, P. J. B., Jones, H. S., Battersby, K., Fodarella, C., Logan, K., Jones, B. C., & Frowd, C. D. (in press). Constructing identifiable composite faces: the importance of cognitive alignment of interview and construction procedure. *Journal of Experimental Psychology: Applied.*

Sporer, S. L. (1996). Experimentally induced person mix–ups through media exposure and ways to avoid them. In G. M. Davies, S. Lloyd-Bostock, M. McMurran, & C. Wilson (Eds.), *Psychology and law: advances in research* (pp. 64–73). Berlin: De Gruyter.

Sporer, S. L., Penrod, S., Read, D., & Cutler, B. (1995). Choosing, confidence, and accuracy: a meta-analysis of the confidence–accuracy relation in eyewitness identification studies. *Psychological Bulletin, 118*, 315–327. https://doi.org/10.1037/0033-2909.118.3.315.

Sporer, S. L., & Cohn, L. D. (2011). Meta-analysis. In B. D. Rosenfeld & S. D. Penrod (Eds.), *Research methods in forensic psychology* (pp. 43–62). New York: Wiley.

*Sporer, S. L., Tredoux, C. G., Schöppl, J., Schäfer, A., Nortje, A., & Kempen, K. (2016). Eyewitness identification and face composite construction. Unpublished manuscript, Universities of Giessen, and Cape Town.

State v. Henderson (2011). 27 A.3d 872, 918–19 (N.J. 2011).

Steblay, N. K., Wells, G. L., & Bradfield-Douglass, A. (2014). The eyewitness post identification feedback effect 15 years later: theoretical and policy implications. *Psychology, Public Policy, and Law, 20*, 1–18. https://doi.org/10.1037/law0000001.

Suero, M., Privado, J., & Botella, J. (2017). Methods to estimate the variance of some indices of the signal detection theory: a simulation study. *Psicologica: International Journal of Methodology and Experimental Psychology, 38*(1), 149–175.

Topp-Manriquez, L. D., McQuiston, D., & Malpass, R. S. (2016). Facial composites and the misinformation effect: how composites distort memory. *Legal and Criminological Psychology, 21*, 372–389. https://doi.org/10.1111/lcrp.12054.

Tousignant, J. P., Hall, D., & Loftus, E. F. (1986). Discrepancy detection and vulnerability to misleading postevent information. *Memory & Cognition, 14*(4), 329–338. https://doi.org/10.3758/bf03202511.

Tredoux, C. (2002). A direct measure of facial similarity and its relation to human similarity perceptions. *Journal of Experimental Psychology: Applied, 8*(3), 180–193. https://doi.org/10.1037/1076-898x.8.3.180.

Tredoux, C. G., Nunez, D. T., Oxtoby, O., & Prag, B. (2006). An evaluation of ID : an eigenface based construction system. *South African Computer Journal, 37*, 90–97.

Valentine, T., & Mesout, J. (2009). Eyewitness identification under stress in the London Dungeon. *Applied Cognitive Psychology, 23*, 151–161. https://doi.org/10.1002/acp.1463.

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

Viechtbauer, W., & Cheung, M. W.-L. (2010). Outlier and influence diagnostics for meta-analysis. *Research Synthesis Methods, 1*(2), 112–125. https://doi.org/10.1002/jrsm.11.

Vredeveldt, A., Groen, R. N., Ampt, J. E., & van Koppen, P. J. (2017). When discussion between eyewitnesses helps memory. *Legal and Criminological Psychology, 22*, 242–259. https://doi.org/10.1111/lcrp.12097.

Vredeveldt, A., Tredoux, C. G., Nortje, A., Kempen, K., Puljević, C., & Labuschagne, G. N. (2015). A field evaluation of the Eye-Closure Interview with witnesses of serious crimes. *Law and Human Behavior, 39*, 189–197. https://doi.org/10.1037/lhb0000113.

Wells, G. L., Charman, S. D., & Olson, E. A. (2005). Building face composites can harm lineup identification performance. Journal of Experimental Psychology: Applied,* 11, 147–156. doi:https://doi.org/10.1037/1076-898x.11.3.147.

Wilson, D. B. (2009). Systematic coding. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 159–176). New York: Russell Sage.

*Wogalter, M. S., Laughery, K. R., & Thompson, B. G. (1989). Eyewitness identification: effects of composite construction on subsequent recognition performance. Unpublished manuscript, Rensselaer Polytechnic Institute, Troy, New York.

*Yu, C. J., & Geiselman, R. E. (1993). Effects of constructing identi-kit composites on photospread identification performance. Criminal Justice and Behavior, 20, 280–292. doi:https://doi.org/10.1177/0093854893020003005.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Colin Tredoux, Ph.D. is a Professor in the Department of Psychology at the University of Cape Town, South Africa. He teaches courses in social psychology, psychology and law, and statistics. His research in psychology and law focuses on eyewitness identification. He is the 2016 recipient of the Chair d'Attractivité at the University of Toulouse, France. His publications have appeared in peer-reviewed journals such as American Psychologist, Psychological Science, Journal of Experimental Psychology: Applied, Developmental Science, and Law and Human Behaviour.

Siegfried Ludwig Sporer, Ph.D. is (retired) Professor of Social Psychology and Psychology and Law at the University of Giessen, Germany. His research has focused on eyewitness testimony, facial recognition and person identification, and eyewitness meta-memory as well as on deception and its detection. In recent years, he has specialized on meta-analyses of various aspects of eyewitness testimony and deception.

Annelies Vredeveldt is Associate Professor of legal psychology at the Department of Criminal Law and Criminology at VU University Amsterdam. Her research focuses on memory in legal settings. She has secured research funding from various prestigious sources, including the High-Value Detainee Interrogation Group, the Society in Science and the European Research Council. She coordinates Project Reasonable Doubt in The Netherlands, regularly serves as an expert witness in criminal cases, serves on advisory committees of the Netherlands Register of Court Experts and is Governing Board member of the Society for Applied Research in Memory and Cognition.

Kate Kempen is a Doctoral Candidate in the Department of Psychology at the University of Cape Town. Her research is predominantly focused on eyewitness memory and contamination, facial processing and recognition, and the construction of facial composites by eyewitnesses of crimes. Part of her Doctoral work involves field-work composite construction interviews with victims of crimes and police officers from the Facial Identification Unit of the South African Police Services.

Alicia Nortje is a Postdoctoral Research Fellow in the Department of Psychology at the University of Cape Town, South Africa. For her postdoctoral work, she is looking at novel ways to analyse witness statements using natural language processing techniques. Her other research interests include associative memory, upper limits of memory for faces, computational models of memory, and memory for multiple perpetrators (which was the topic of her PhD thesis).