Powerful Effects of Diagnostic Information on Automatic and Self-reported Evaluation: The Moderating Role of Memory Recall

Pieter Van Dessel [a*], Jeremy Cone [b], and Anne Gast [c]

[a] Department of Experimental-Clinical and Health Psychology, Ghent University, Ghent, Belgium; [b] Department of Psychology, Williams College, Williamstown, United States; [c] Department of Psychology, University of Cologne, Cologne, Germany

Author Note

The research reported in this paper was funded by the Scientific Research Foundation, Flanders under Grant FWO16/PDO/201 to PVD, and by the Deutsche Forschungsgemeinschaft under grant GA 1520/2-1 to AG. There was no potential conflict of interest to report. Correspondence concerning this article should be sent to Pieter.vanDessel@UGent.be

Abstract

We sometimes learn about certain behaviors of others that we consider diagnostic of their character (e.g., that they did immoral things). Recent research has shown that such information trumps the impact of other (less diagnostic) information both on self-reported evaluations and on more automatic evaluations as probed with indirect measures such as the Affect Misattribution Procedure (AMP). We examined whether facilitating memory recall of alternative information moderates the impact of diagnostic information on evaluation. In Experiments 1 and 2, participants learned one diagnostic positive and one diagnostic negative behavior of two unfamiliar people. Presenting a cue semantically related to this information during evaluation influenced AMP scores but not self-reported liking scores. Experiments 3 and 4 showed that elaborative rehearsal of low diagnostic information eliminated diagnosticity effects on AMP scores and reduced them on self-reported liking scores. These findings help elucidate the role of memory recall and diagnosticity in evaluation.

Keywords: information diagnosticity, memory recall, impression formation, automatic evaluation, self-reported evaluation

We often accrue information about other people that we consider highly relevant for our evaluation of that person. For instance, a person might learn that a politician has committed a felony or that their romantic partner has committed adultery, and deem this information to be important for forming an accurate impression of this person. Prior evidence has robustly established that information diagnosticity (i.e., the extent to which information is considered relevant or informative for knowing the true character of a person) plays a crucial role in impression formation (Cone & Ferguson, 2015; Skowronski & Carlston, 1987). Yet, an important outstanding question is whether effects of diagnostic information on evaluation depend on specific moderators or boundary conditions. Once people learn a diagnostic piece of information, does it exert an indelible, stable influence on evaluation or can its effects be moderated to any substantial degree?

Moderators of effects of diagnostic information on evaluation

Recent studies examining moderation of the effect of diagnosticity focused on the question of whether diagnosticity differentially influences different types of evaluation measures. These studies established that diagnostic information has a strong and immediate impact both on measures of self-reported and more automatic evaluation (see Cone et al., 2017, and De Houwer et al., 2020, for recent reviews). For instance, Cone and Ferguson (2015) showed that participants who had first learned many pieces of positive information about a person named Bob and then learned a single piece of counter-attitudinal information that they considered more diagnostic of Bob's true character (e.g., that Bob was a convicted child molester), exhibited negative evaluations of Bob as measured with self-reported ratings of liking and with measures of more automatic evaluation such as the Affect Misattribution Procedure (AMP; Payne et al., 2005). In the latter type of measures, evaluative behavior might occur under some (but not all) of the conditions of automaticity (e.g., unintentional, unconscious, efficient, or fast: Moors, 2016; see Gawronski & De Houwer, 2014,

for a review). For instance, in the AMP, participants rapidly evaluate Chinese ideographs that are preceded by the presentation of a target stimulus (e.g., Bob) and it is often found that the prime influences evaluative responses to the Chinese ideograph even though participants are asked to only evaluate the Chinese ideograph.

These prior findings of overpowering effects of diagnostic information on more automatic evaluations have challenged key assumptions of important cognitive (dual-process) theories of evaluation (e.g., Rydell & McConnell, 2006). Contrary to self-reported evaluations, more automatic evaluations were often thought to reflect the automatic activation of associations in memory as determined by the total number of pieces of positive compared to negative information that people had learned about a stimulus (Smith & DeCoster, 2000). However, studies showing that both self-reported and more automatic evaluations strongly depend on how diagnostic people consider individual pieces of valenced information for evaluation could suggest that inferential processes determine both types of evaluation (De Houwer, 2014; De Houwer et al., 2020). Specifically, both automatic and self-reported evaluations might reflect inferences about stimulus valence that take into account how relevant the available stimulus information is for evaluation (Van Dessel, Hughes, & De Houwer, 2019).

It is often argued that propositional processes in evaluation operate on the basis of rule-based logic (Rydell & McConnell, 2006; Strack & Deutsch, 2004). If propositional processes underlie both self-reported and more automatic evaluations, one could therefore argue that both types of evaluation reflect computations of diagnosticity (e.g., based on logical rules) and should therefore always correspond strongly. However, this might not fit with the extant evidence showing that different (types of) evaluation measures sometimes exhibit dissociations after learning valenced information (Gawronski & Brannon, 2019). For instance, in studies by Van Dessel et al.

(2016), there was a dissociative effect of approach-avoidance information. Participants first learned that members of one unfamiliar social group (e.g., Niffites) had positive traits and members of another group (e.g., Luupites) had negative traits and were then informed that they would later avoid the members of the former group and approach the members of the latter group. Both manipulations might evoke inferences about the valence of the social groups but the former information might be considered more diagnostic for these evaluative inferences. Interestingly, while self-reported evaluations readily reflected the (probably more diagnostic) trait information, more automatic evaluations were also impacted by the conflicting approach-avoidance information.

Though these results suggest that diagnosticity is not always overpowering in particular in determining more automatic evaluation, the evidence is only indirect (diagnosticity was not manipulated). Moreover, it is unclear when and why this might be the case and under which conditions this might also apply to self-reported evaluations. Recent studies have therefore started to investigate moderators of diagnosticity effects on automatic and self-reported evaluation (e.g., Brannon & Gawronski, 2017; Cone et al., 2019; Fourakis et al., under review; Mann & Ferguson, 2015; Van Dessel et al., 2020). The current work focuses on the role of memory recall.

Moderation of diagnosticity effects by memory recall

Drawing on an inferential account of evaluation (Van Dessel, Hughes, et al., 2019), we postulate that both self-reported and more automatic evaluation might critically depend on goal-dependent inferences. From this perspective, inferential processes constitute the generation of propositional information on the basis of momentarily entertained information and this process is controlled by current goals (i.e., wanted outcomes). In a typical evaluation task, participants typically have the goal to emit an accurate evaluative response which prompts a person to infer

stimulus valence (e.g., 'Bob is positive') and this inference is transferred into evaluative responding. Given its relevance for the goal to provide accurate evaluations, diagnostic information should be strongly integrated in evaluative responding. However, the inferential process is also considered to be biased by the availability of information in memory (Sanborn & Chater, 2016). As a result, the effect of diagnostic information should depend on the ease of retrieval of this information from memory in light of other information that is currently entertained (Van Dessel et al., 2020).

From this perspective, memory recall should play a key role in effects of diagnostic information on evaluation, in accordance with the more general idea that ease of recall of valenced information strongly impacts evaluation (Gast, 2018; see also Gawronski & Bodenhausen, 2006). Note that we refer to memory recall as a behavioural phenomenon (i.e., the likelihood that a certain piece of previously presented information is reported) to distinguish behavioural level evidence from explanations of this evidence at the mental process level (e.g., memory retrieval processes; Hughes et al., 2016)[1]. Numerous studies provide evidence that the impact of different pieces of valenced information on evaluation depends on the extent to which these pieces of information are easy to recall (e.g., Keller, 1987; Hansen & Wänke, 2008; Wänke et al., 1996; Wood, 1982) or have been recalled (Benedict et al., 2019). For instance, Gast et al. (2021) used a manipulation that is known to influence memory recall – i.e., presentation of contextual cues - and showed that this influenced evaluations that were formed based on evaluative conditioning, valenced behavioural information, and autobiographic memories (see also Richter & Gast, 2017).

[1] The term "memory recall" is often used to refer to mental level constructs (e.g., memory is a mental structure). While we define our manipulations at the behavioural level, we chose to adopt this term for the sake of consistency with (more cognitively oriented) research that described the memory recall manipulations that were tested in this manuscript (i.e., cueing and elaborative rehearsal).

Importantly, however, if both memory recall and diagnosticity determine evaluation then it is important to examine how they interact to achieve their effects. The constructs of diagnosticity and memory recall might be strongly related. For instance, information that is considered highly relevant for evaluation (i.e., diagnostic) might also be recalled easily and, vice versa, information that is easily recalled might also be considered highly relevant for evaluation. Nevertheless, it is of interest to better understand their joint and independent contributions to evaluation both for theoretical reasons (e.g., to facilitate test and development of evaluation theories) and for practical reasons (e.g., to develop interventions that target formation and change of evaluations).

The current set of studies focuses on the question whether effects of diagnostic information on evaluation are impacted by manipulations that influence memory recall of alternative information. When learning diagnostic information about the behavior of someone (e.g., an abusive co-worker, a cheating partner), does this information always stand out as particularly relevant for evaluation such that it is strongly integrated in evaluation no matter what, or might (contrasting) information with a higher likelihood of memory recall also influence evaluation?

Two recent studies examined effects of a memory recall manipulation (i.e., contextual cueing) on effects of diagnostic information on evaluation. Contextual cueing is well-known to improve recall of information and to also influence evaluation, especially more automatic evaluations (see Gawronski et al., 2015, 2018, for reviews). For instance, Rydell and Gawronski (2009) first presented participants with either positive or negative information about a target person against one colored background (e.g., a yellow screen) and then presented counter-attitudinal information against a different colored background (e.g., a blue screen). Results showed that evaluations of the target person as assessed with the AMP reflected the counter-attitudinal information only when this information was cued by presenting the target against the background

of the second learning block during evaluation. A set of studies by Brannon and Gawronski (2017) extended this investigation to the diagnosticity effects observed by Cone and Ferguson (2015). Drawing on the mental process level idea that automatic activation of associative representations underlies automatic evaluations and that this activation is highly context-dependent, they argued that both initially learned (low diagnostic) and counter-attitudinal (high diagnostic) information might be stored in contextualized (associative) representations. As a result, the extent to which non-diagnostic and diagnostic information might influence automatic evaluation should depend on whether the evaluation context is similar to the learning context. To test this, they manipulated background color when presenting initial and counter-attitudinal information. Counter to the authors' expectations, however, both automatic and self-reported evaluations reflected information diagnosticity independent of the background that was present during evaluation.

These results suggest that the manipulation of memory recall of valenced information in the context of diagnostic information does not moderate its effects on evaluation. However, an alternative explanation is that the studies lacked a memory recall manipulation that is sufficiently strong to counteract highly diagnostic information. From the perspective of the inferential model of evaluation (Van Dessel, Hughes, et al., 2019), the generation of propositional information is goal-directed. Cueing information with previously co-occurring information that is irrelevant for evaluative goals (i.e., background color), might therefore be ineffective. To counteract the powerful influence of diagnosticity, memory recall manipulations might be required that have evaluative meaning and therefore allow activation of the valenced behavioral information in accordance with the evaluation goal. To achieve this, one could, for instance, present cues that are semantically related to the valenced content of the cued information.

A recent set of experiments by Van Dessel et al. (2020) examined semantic contextual cueing in the context of diagnosticity effects. Participants first read one piece of positive and one piece of negative information about a target person (i.e., Bob) and were then presented with an image semantically related to this behavior during evaluation. Results revealed effects on AMP scores and self-reported evaluations that reflected differences in the diagnosticity of the two pieces of information but did not show evidence for cueing effects. Notably, however, this series of studies examined effects of semantically related cues in the context of other manipulations (e.g., information primacy manipulations) which might have counteracted cueing effects. Moreover, participants read only two pieces of diagnostic valenced information which could have prevented cueing effects (i.e., due to a ceiling effect) because both pieces of information were easily recalled.

Overview of studies

To systematically test whether memory recall can moderate effects of diagnostic information on evaluation, we performed four experiments which tested effects of two distinct manipulations that are known to influence memory recall (Experiment 1 and 2: semantic contextual cueing, Experiments 3 and 4: elaborative rehearsal). In each experiment, we first presented one piece of positive and one piece of negative information about the behavior of unfamiliar men (intermixed with neutral behavioral statements) and varied diagnosticity of these pieces of information. Next, we manipulated memory recall of some of these pieces of information and then tested the impact of this memory recall manipulation on automatic and self-reported evaluations of these men. For the sake of consistency with the previously mentioned key findings of diagnosticity effects (Cone & Ferguson, 2015; Rydell & Gawronski, 2009), evaluations were measured with the AMP and with self-reported ratings.

In Experiment 1, positive and negative information about two men (Bob and Jake) was matched in rated diagnosticity and the memory recall manipulation involved the presentation of a contextual cue during evaluation that was semantically related to the learned positive information of one person and the learned negative information of the other person (e.g., an image of sports shoes when Bob was said to have stolen sports shoes). Experiment 2 was a replication in which positive and negative information about the two men differed in rated diagnosticity (e.g., the negative information was less diagnostic for both men) and we probed effects of cueing the less diagnostic information of one of the targets.

Experiments 3 and 4 also used information that differed in rated diagnosticity across valence categories and examined effects of memory recall of low diagnostic information of one of the men. However, these experiments used another potent (and goal-relevant) memory recall manipulation (i.e., elaborative rehearsal). Elaborative rehearsal involves deep thinking about learned information to foster connection with other information and is typically found to be one of the most effective manipulations of memory recall (Bartsch et al., 2018; Goldstein, 2011).

For all experiments, we specified hypotheses based on the inferential theory of evaluation (Van Dessel, Hughes, et al., 2019). Specifically, we postulated that AMP scores would be influenced by the manipulations of memory recall implemented in the different experiments. We did not specify directional hypotheses for self-reported evaluations (in Experiments 1 and 2) because we assumed that, on the one hand, facilitation of memory recall of valenced information should facilitate the integration of this information in self-reported evaluation, which might allow for effects of the manipulation (Gast, 2018). On the other hand, however, these effects might be weak or absent when participants have ample opportunity and motivation to also take into account more difficult to recall information in self-reported evaluation measures where evaluation occurs

under less automatic conditions (Van Dessel et al., 2016). Prior to data-collection, these hypotheses were pre-registered together with the sampling plan, study design, data-analytic plans, and experimental hypotheses. The pre-registered plans, raw data, and experiment and analytic scripts of all experiments are available at https://osf.io/2by8c/ (pre-registrations: https://osf.io/xc9pr/) .

Experiment 1

Experiment 1 examined the effect of semantic contextual cueing of diagnostic information on evaluation. Participants learned one positive and one negative piece of information about two unfamiliar men. We examined whether presenting an image semantically related to one of the pieces of information during evaluation influences AMP scores and self-reported evaluations.

Method

Participants. A total of 250 participants were recruited from the online participant recruitment platform Prolific Academic (https://www.prolific.ac/). Only participants with English as native language were allowed to participate (because we considered it important that participants could fluently read our English stimulus materials). The sample size was determined using sequential Bayes hypothesis testing. This is a technique where evidence for how strongly the data support either the null hypothesis (BF_0) or the alternative hypothesis (BF_1) is computed with Bayes Factors and sampling continues until an a priori defined level of evidence is reached. This allows flexible sampling plans which can be important when it is difficult to correctly guess effect sizes in an a priori power analysis (Schönbrodt et al., 2017). Moreover, it prevents inflating Type I error rates under the null hypothesis significance testing (NHST) paradigm. Analyses were performed when 100 participants had completed the study and sample size was then increased by steps of 50 participants until a decisive Bayes factor (larger than 6) was obtained for the critical *t*-test comparing AMP scores for the two target persons. A maximum of 400 participants was set

(ensuring sufficient power, power>0.90, to detect a small effect of *dz*=0.20 at alpha=0.05). Bayesian analyses were performed according to the procedures outlined by Rouder et al. (2009).

Following our preregistration plan and standard procedures for data-reduction of AMP tasks (e.g., Cone & Ferguson, 2015), we excluded data from participants who (1) did not complete all measures and tasks (1 participant: 0.4%), (2) had completed AMP blocks using only one response key (10 participants: 4.0%), (3) indicated they knew Chinese and therefore recognized some of the Chinese ideographs (2 participants: 0.8%), or (4) made at least one error in the memory check questions (44 participants: 17.6%). Analyses were performed on the data of 193 participants (88 women; mean age=38.12, *SD*=12.11, range=19-66; country of residence: 72.5% UK, 23.8% US, 3.7% other). Participants received a monetary reward of 1.50 Great-British Pounds for participation.

Design. The experiment used a mixed design involving one focal within-subjects factor with two conditions: Cued Valence (positive vs. negative information was cued for this target). We also manipulated six method factors between-subjects: Target Name (positively cued target named Bob vs. Jake), Target Picture (positively cued target depicted with Picture 1 vs. Picture 2), Target Information Order (information of positively cued target presented first vs. second), Target Positive Statement (positively cued target presented with positive behavioral statement 1 vs. positive statement 2), Target Negative Statement (positively cued target presented with negative statement 1 vs. negative statement 2), and Evaluation Task Order (AMP first vs. self-report rating task first). Manipulations were counterbalanced across participants.

Materials. Two attitude targets (Bob and Jake) were represented by two pictures of White men selected from the Chicago Face Database (Ma et al., 2015) on the basis of a rating study in which 51 Prolific Academic participants rated the faces as the most evaluatively neutral (i.e.,

ratings near the mid-point of the 7-point Likert scale ranging from 1[very negative] to 7 [very positive]; Picture 1: *M*=3.96, *SD*=0.96; Picture 2: *M*=3.97, *SD*=1.00). We counterbalanced which face represented Bob and which represented Jake.

We selected two positive behavioral statements (PS1 and PS2) and two negative behavioral statements (NS1 and NS2) based on a rating study in which 100 Prolific Academic participants rated 90 behavioral statements presented in compound with a picture that is related to the behavior on four characteristics: (1) valence of the behavior (on a scale ranging from 1 [highly negative] to 7 [highly positive], (2) valence of the picture presented with the statement, (3) how well they could remember the statement after cueing with the picture, and (4) how diagnostic of the person's true character they considered the statement to be (on a scale ranging from 1 [not at all diagnostic] to 5 [extremely diagnostic]). Importantly, we included statements that were rated high, but not extreme in diagnosticity (scores 3-4; *Ms*=3.23-3.69). This was done to ensure that statements could be matched on valence and that diagnosticity was not confounded with valence extremity (see Cone & Ferguson, 2015). Four diagnostic statements were selected, two of positive and two of negative valence, that were (1) matched on valence extremity (highly negative: *M*=1.69/1.50 or highly positive: *M*=6.31/6.25), (2) easy to remember (i.e., ratings>3 for the question "How well do you remember the statement that went together with the following image?" measured on a scale ranging from 1 [Not at all] to 5 [I remember it perfectly]), and (3) represented by a picture of neutral valence (i.e., ratings>3 and<5 on the 7-point evaluative scale). We counterbalanced which positive and which negative behavioral statement was presented with Bob and which was presented with Jake (Table 1). Additionally, six statements were selected that were rated as low in diagnosticity (*M*=[1.20-1.36], *SD*=[0.63-0.80]) and evaluatively neutral. These statements were randomly assigned to Bob and Jake. The statements used in Experiments 1-3 are provided in Appendix.

Procedure. In line with recommendations by Zhou and Fishbach (2016) to prevent selective attrition, participants were first (1) informed about the duration of the experiment and (2) warned that dropping out could affect the quality of data and that it is essential for scientific advance to have good data. Next, participants provided informed consent and answered demographic questions regarding their age, gender, and country of residence.

Participants were shown images of Bob and Jake and informed that they would learn information about the two people by going through several trials in which they would see their picture together with information about a behavior they engaged in. Participants were asked to read the information carefully and to try and remember it. Participants then received 5 evaluative learning task trials about Bob during which a picture of Bob was continuously presented at the bottom of the screen. In each trial, a behavioral statement was presented in the middle of the screen and the picture related to the behavior was presented above the statement (Figure 1). After 8 seconds, a prompt appeared indicating that participants could now progress to the next piece of information by pushing the space bar. Next, the evaluative learning was repeated with different behavioral statements for Jake. The two valenced statements (one positive and one negative) were always presented on the second and fourth trials whereas the neutral statements were presented on the odd-numbered trials. Whether the positive or the negative statement was presented first was counterbalanced across participants (but remained identical for both targets) as well as whether the information for Bob or Jake was presented first.

Participants next completed an AMP measuring relatively automatic evaluations of Bob and Jake and a self-report rating task (order counterbalanced). We deployed commonly used versions of both tasks with the crucial exception that, for the full duration of both tasks, two pictures were always presented on top of the screen: one picture related to the positive behavior of Bob and

one picture related to the negative behavior of Jake (or vice versa) (Figure 2). This was done to manipulate memory recall of the cued behavioral statements. For the AMP, participants received instructions explaining that they would complete a judgment task in which each trial involved (1) the brief presentation of a fixation point, (2) an image of a specific person and (3) a Chinese character in the middle of the screen that would eventually be covered by (4) a noisy image. Participants were told that they would need to indicate on each trial whether the Chinese character was less pleasant (E key) or more pleasant than average (I key) and they were instructed not to be influenced by the images of the specific people. In line with standard procedures (Payne et al., 2005), during each trial, participants saw a prime stimulus for 75ms consisting of the image of Bob or Jake, a blank screen for 125ms, and a Chinese ideograph for 100ms. Finally, a mask image was presented (a black and white pattern) until the participant made a response by using the E or I key of their computer keyboard. In this task, it is often found that people's evaluation of the prime influences their evaluative responses to the Chinese ideograph even though they are asked to only evaluate the Chinese ideograph. This evaluation effect is often considered to be automatic in some ways (e.g., in the sense of fast or unintentional: Mann et al., 2019). The AMP consisted of 60 trials, half with the face of Bob as prime and half with the face of Jake as prime. The memory cueing pictures were presented at the top of the screen for the full duration of the task.

The self-report rating task consisted of four questions asking participants to rate their liking of Bob and Jake: "How pleasant or unpleasant do you find Bob/Jake?" and "To what extent do you have warm feelings for Bob/Jake?". Participants gave their evaluative ratings by selecting an option on a 7-point Likert scale (1=Extremely unpleasant/cold; 7=Extremely pleasant/warm). The order of the questions was randomized. The memory cueing pictures were presented at the top of the screen also for the full duration of this task.

Next, participants were shown the image of Bob or Jake as well as a list of eight valenced behavioral statements in which the previously presented valenced statements were interspersed. Participants were asked to indicate which of the behaviors the depicted person had engaged in by selecting the behavioral statement from a set of statements. Finally, participants indicated whether they knew Cantonese or Mandarin and therefore recognized some of the Chinese characters they saw in the AMP and were probed for demand compliant responding by asking to what extent they had faked their responses to accord with what they thought the experimenter wanted them to do. They were then thanked for their participation.

Results

AMP scores. We calculated AMP scores as the percentages of 'pleasant' responses for Bob and Jake. These scores were recoded based on the cues that were present during the evaluation task such that there were two AMP scores, one for the person for whom the positive information was cued and one for the person for whom the negative information was cued. The AMP scores were subjected to analyses with item-based linear mixed effects (lme) models as implemented in R package lme4 (Bates et al., 2015). We tested a model that included (1) the within-subjects factor Cued Valence (positive or negative), (2) the three between-subjects factors Target Name (counterbalancing factor: assignment of positively cued target to Bob vs. Jake), Target Picture (counterbalancing factor: assignment of positively cued target to Picture 1 vs. Picture 2), and Evaluation Task Order (counterbalancing factor: AMP first vs. self-report rating task first) as fixed factors, (3) all interactions, and (4) random intercepts of Participant, Target Positive Statement (PS1 vs. PS2) and Cued Negative Statement (NS1 vs. NS2). The reported *p*-values for the fixed effects are based on a Type-III ANOVA using a χ^2-distribution as implemented in the R package 'car' (Fox & Weisberg, 2011). The three random effects explained 39.06% of the total variance.

In line with our hypotheses, we observed (only) the main effect of Cued Valence, $\chi^2(1)=8.28$, $p=.004$. The planned paired *t*-test revealed that participants exhibited more positive evaluations of the positively cued person ($M=0.59$, $SD=0.20$) than the negatively cued person ($M=0.54$, $SD=0.21$), $t(192)=3.07$, $p=.001$, $BF_1=14.82$, $d_z= 0.22$.

Self-reported rating scores. Liking and warmth ratings of Bob and Jake were aggregated into a single score for each person by averaging the respective scores (mean Cronbach's Alpha=.88, $SD=0.02$). Scores were recoded such that they indicated liking of the positively cued and negatively cued person. The mean correlation of self-reported rating scores and AMP scores for the positively cued and negatively cued person was moderate, at respectively $r(191)=.30$, and $r(191)=.22$, ps<.001. The self-reported rating scores were subjected to the same lme model as we used for AMP scores. The three random effects explained 9.87% of the total variance.

We observed no main effect of Cued Valence, $\chi^2(1)=0.59$, $p=.44$. The planned paired *t*-test showed no significant difference between the positively cued person ($M=4.04$, $SD=1.16$) and the negatively cued person ($M=3.92$, $SD=1.02$), $t(192)=1.12$, $p=.13$, $BF_0=3.86$, $d_z= 0.08$. We did observe two interaction effects: an interaction of Target Name and Target Picture, $\chi^2(1)=5.63$, $p=.018$, indicating that, if the positively cued target was named Bob, participants gave more positive overall ratings when this target was assigned to Picture 1 than to Picture 2, $p= .037$, but not when the positively cued target was named Jake, $p=.46$. More importantly, we also observed an interaction of Cued Valence, Target Name and Task Order, $\chi^2(1)=11.48$, $p<.001$, revealing that, only when the positively cued target was named Bob and participants first completed the explicit rating task, participants exhibited a significant preference for the positively cued person ($M=4.32$, $SD=1.18$) over the negatively cued person ($M=3.91$, $SD=1.01$), $t(60)=2.12$, $p=.038$. However, evidence for this effect was weak, $BF_1=2.18$, $d_z= 0.27$.

Discussion

The results of Experiment 1 suggest that manipulation of memory recall on the basis of relevant semantic cues can moderate the effect of diagnostic information on evaluation. When participants had learned one positive and one negative piece of information that were of matched diagnosticity, both about Bob and about Jake, a preference for one of the two targets could be induced by presenting a semantic contextual cue for the positive information of one of the targets and for the negative information of the other target. Notably, this effect was only observed on AMP scores and not on self-reported evaluations. The absence of the latter effect, however, might be due to a lack of power given that the Bayes Factor for the absence of an effect was unconvincing (the stopping rule for data collection was based on AMP score analyses) and given (weak) evidence for an effect for participants who started with the explicit rating task (depending on the target name).

Experiment 2

The aim of Experiment 2 was to replicate and extend the observed effect of semantic contextual cueing on AMP scores to situations where the available valenced information is not matched in diagnosticity. It is in this situation that previous studies have typically examined diagnosticity effects, showing evidence for overpowering effects of diagnosticity on automatic evaluation (Cone & Ferguson, 2015) and for the context-independentness of this effect (Brannon & Gawronski, 2017). An important question is therefore whether memory cueing of valenced information with a relevant (i.e., semantically related) cue can moderate the effects of this information on evaluation when (1) the cued information is more diagnostic than the information that is not cued and (2) the cued information is less diagnostic than the information that is not cued. To test this, we provided two separate pieces of positive and negative information about Bob, one that was high and one that was moderate in diagnosticity, and presented matched statements for

Jake (for each participant, the two high diagnostic statements thus had the same valence). We examined whether cueing one of these four pieces of information (about Bob OR Jake) during evaluation influenced automatic and self-reported preferences for Bob or Jake. Note that, in contrast to Experiment 1, we cued only one piece of information (for one target) rather than one piece of information for each target. This was done to allow testing whether the effect depends on the type of information that was cued (e.g., high versus low diagnostic information and positive versus negative information). Note that we had no specific predictions related to these factors (exploratory analyses). As in Experiment 1, we only predicted a main effect of semantic contextual cueing on AMP scores.

Method

Participants. A total of 150 Prolific Academic participants were recruited. The sampling plan was identical to Experiment 1 with the Bayesian stopping rule dependent on the *t*-test comparing AMP scores for the two target persons. We excluded data from participants who (a) indicated that they had recognized some of the Chinese ideographs (3 participants; i.e., 0.02%), or (b) made at least one error on the questions that probed memory for the pieces of valenced information (29 participants; i.e., 19.33%). There were no participants who did not fully complete all questions and tasks or who responded either "positive" or "negative" to all AMP trials. Analyses were performed on the data of 118 participants (71 women, mean age=35.21 years, *SD*=9.02).

Design. The experiment also used a mixed design involving one focal within-subjects factor with two conditions: Cued Target (the target that was cued relatively more positive or negative). Note that, in contrast to Experiment 1, there was only one cued piece of information and the levels of Cued Target therefore refers to the relative valence of the cued information about the target person. That is, one person was r*elatively* more positively cued (either the positive information was

cued [and no information was cued for the other person] or no information was cued [and negative information was cued for the other person]) and one person was *relatively* more negatively cued (either no information or negative information was cued). There was also one important between-subjects factor: Cued Information Diagnosticity (more vs. less diagnostic information was cued). We also manipulated the following non-focal method factors: Cued Information Order (cued information presented first vs. second in the learning phase), Valence of Cued Information (positive or negative information cued), and Evaluation Task Order. Manipulations were counterbalanced across participants. Because the factors Target Name and Target Picture did not influence evaluations in Experiment 1, these were not manipulated in Experiment 2.

Procedure. The procedure was identical to Experiment 1 with two exceptions (Figure 2). First, we used other valenced behavioral statements. Specifically, we used four statements rated as strongly negative, two of which were rated as high in diagnosticity: “Person X stole money from a church's charity fund”, “Person X snatched a purse that an elderly woman set down on the bus”, and two were rated as moderate in diagnosticity: “Person X pretended he did not hear a woman's request for his help in lifting a baby carriage over a high curb”, “Person X cheated on a take-home exam from the university”. We also used four statements rated as strongly positive, two of which were rated as high in diagnosticity: “Person X anonymously donated money to develop a new wing of a hospital”, “Person X took in and cared for an old woman from his church when her husband died”, and two were rated as lower in diagnosticity: “Person X offered to share an umbrella with a stranger during a downpour”, “Person X donated blood.” For each combination of valence (positive, negative) and diagnosticity (high, low), there were thus two different statements. These two matched statements we refer to as “sets”. Of the four different sets, we used two for each participant, either highly diagnostic positive (HDP) and less diagnostic negative (LDN) or highly

diagnostic negative (HDN) and less diagnostic positive (LDP) sets (it was counterbalanced whether participants had the HDP and LDN sets or the HDN and LDP sets). For each participant, the two targets were assigned to matched statements (either both targets HDP and LDN or both targets HDN and LDP) (Table 1).

A second modification was that, during both automatic and self-reported evaluation tasks, a semantic contextual cue was provided for only one (rather than two) of the four pieces of valenced information. As mentioned above, it was manipulated whether the cued information was high or low in diagnosticity and whether it was positive or negative. It was counterbalanced with the other method factors whether the cued information referred to Bob or Jake.

Results

AMP scores. As in Experiment 1, we calculated one AMP score as the percentage of 'pleasant' responses for each person prime (Bob and Jake) and scores were recoded such that there was one score for the person who was relatively more positively cued and one score for the person who was relatively more negatively cued. We tested an lme model that included Participant as random factor and the within-subjects factor Cued Target (target cued relatively more positive or negative). The random effect explained 25.67% of the total variance.

We observed the expected main effect of Cued Target, $\chi^2(1)=7.05$, $p=.008$, $BF_1=9.29$. Participants exhibited more positive automatic evaluations of the relatively more positively cued person ($M=0.61$, $SD=0.19$) than of the relatively more negatively cued person ($M=0.55$, $SD=0.21$), $t(117)=2.66$, $p=.005$, $BF_1=10.11$, $d_z= 0.24$. Importantly, the inclusion of the between-subjects factor Cued Information Diagnosticity (high or low diagnostic) did not improve model fit, $\chi^2(2)=3.50$, $p=.17$, nor did the inclusion of Target Information Order (presented first or second),

Valence of Cued Information (positive or negative information cued), Evaluation Task Order, or any of the interactions, *ps*>.12.

Self-reported rating scores. Self-reported rating scores were calculated for the relatively more positively and more negatively cued person (mean Cronbach's Alpha=.91, *SD*=.05). The mean correlation of self-reported rating scores and AMP scores for the two people was low at $r(116)=.07$ and $r(116)=.11$, *p*s>.25. The self-reported rating scores were subjected to an lme model that included the within-subjects factor Cued Target and the between-subjects factor Valence of Cued Information as fixed factors, and Participant as a random factor. Inclusion of the variables Target Information Diagnosticity, Target Information Order, and Evaluation Task Order as fixed factors did not improve model fit, χ^2s<4.70, *p*s>.095, so they were not included in the analyses. The random effect explained 49.96% of the total variance.

We did not observe main or interaction effects that included the factor Cued Target, χ^2s<2.46, *p*s>.11, BF_{1s}<1.47. Participants did not provide significantly more positive ratings for the relatively more positively cued person (*M*=4.02, *SD*=1.37) compared to the relatively more negatively cued person (*M*=3.83, *SD*=1.37), $t(117)=1.56$, $p=.061$, $BF_1=1.50$, $d_z= 0.14$. We did observe a main effect of Valence of Cued Information, $\chi^2(1)=5.02$, $p=.025$, indicating more positive overall ratings when the cued information was positive than when it was negative.

Discussion

The results of Experiment 2 corroborate and extend the results of Experiment 1, indicating that semantic contextual cueing of learned information can influence evaluations of target people even when there is a clear imbalance in the diagnosticity of learned information. Similar to Experiment 1, cueing effects were only observed on AMP scores and not on self-reported liking ratings. It is noteworthy that we did not observe the cueing effect on AMP scores to depend on

whether the high or low diagnostic information was cued, which might suggest that memory recall moderates effects of valenced information independent of the diagnosticity of this information. However, caution is warranted when interpreting these findings given that (1) we did not power the experiment for observing this interaction effect and (2) exploratory analyses (reported in Appendix) show that the cueing effect was significant (and represented strong evidence) when the high diagnostic information was cued but non-significant (and represented weak evidence) when the low diagnostic information was cued.

Experiment 3

In Experiments 1 and 2, we observed effects of semantic contextual cueing on AMP scores but not on self-reported evaluation. Notably, even the effects on AMP scores were small overall (ds=0.22-0.24). These results might suggest that memory recall manipulations can only have a weak effect on (automatic) evaluations in the context of diagnostic information. Importantly, however, the previous experiments relied on only one method for manipulating memory recall: semantic contextual cueing. However, there are many other methods that are known to improve recall of information from memory. Perhaps the method that has been found to influence memory recall most strongly, is elaborative rehearsal (i.e., rehearsing information while thinking about the meaning of this information: Goldstein, 2011). Also from the perspective of the inferential model of evaluation, this could be a particularly good memory recall manipulation because it might facilitate integration of information in the network of beliefs that supports evaluative inferences.

Experiment 3 probed effects of elaborative rehearsal of valenced information on evaluation. Similar to Experiment 2, participants first received one piece of negative information and one piece of positive information that differed in diagnosticity about each of two target persons. Different from Experiment 2, however, we did not match information for the two targets. If high diagnostic

positive (HDP) and low diagnostic negative information (LDN) was given for Bob, then low diagnostic positive (LDP) and high diagnostic negative information (HDN) was given for Jake. We then measured evaluations on AMP and self-report rated evaluation tasks that did not include recall cues. Next, we instructed participants to rehearse and elaborate on one of the pieces of learned information for both Bob and Jake, specifically the low diagnostic information (LDN and LDP). We then measured evaluations a second time.

Note that we only tested effects of elaborative rehearsal of the low diagnostic information. We considered this condition most important (and therefore maximized statistical power in it) because effects of a manipulation that goes against effects of diagnosticity is probably the most counter-intuitive and yet the inferential account argues that facilitating ease of memory recall of less diagnostic information can reduce effects of diagnosticity (on automatic evaluation: see Van Dessel et al., 2016; Van Dessel & De Houwer, 2019). Moreover, exploratory analyses in Experiment 2 provided the weakest evidence for an effect of memory cueing of low (rather than high) diagnostic information on automatic evaluation. To further maximize statistical power, we also deviated from Experiments 1 and 2 by measuring evaluations before (Time 1) and after the memory recall manipulation (Time 2).

As for the previous experiments, we hypothesized that our memory recall manipulation (in this case, elaborative rehearsal) should influence AMP scores. However, we also specified two additional hypotheses. First, we predicted a typical diagnosticity effect at Time 1, such that AMP scores and self-report rating scores were predicted to reveal a preference for the person for whom high diagnostic positive (and low diagnostic negative) information was provided over the person for whom high diagnostic negative (and low diagnostic positive) information was provided. Second, based on the results of Experiments 1 and 2 which only showed effects of the memory

recall manipulation on AMP scores, we hypothesized that, at Time 2, this diagnosticity effect would be either strongly reduced or absent on AMP scores but not on self-reported rating scores.

Method

Participants. A total of 150 Prolific Academic participants were recruited with a similar sampling plan as in Experiments 1 and 2 with the Bayesian stopping rule dependent on the *t*-test comparing AMP scores for the two target persons. Data were excluded from participants who (a) did not fully complete all questions and tasks (1 participant:0.67%), (b) had completed AMP blocks using only one response key (12 participants: 8.00%), (c) indicated they had recognized some of the Chinese ideographs (6 participants: 4.00%), or (d) made at least one error on the questions that probed memory for the pieces of valenced information (19 participants: 12.67%). Analyses were performed on the data of 112 participants (70 women, mean age=34.89 years, *SD*=11.96).

Design. The experiment used a mixed design involving two focal within-subjects factors with two conditions: Diagnosticity Valence (target with high diagnostic positive [and low diagnostic negative] vs. high diagnostic negative [and low diagnostic positive] information) and Time of Evaluation (before vs. after rehearsal). We also manipulated the following method factors between-subjects: Information Order (information of high diagnostic positive target presented first vs. second), Rehearsal Order (rehearsal of high diagnostic positive target first vs. second), Diagnostic Information Order (high diagnostic information [for both targets] presented first vs. second), and Evaluation Task Order. Manipulations were counterbalanced across participants.

Procedure. Similar to Experiment 1 and 2, the experiment started with the evaluative learning task in which participants learned about Bob and Jake. Again, there were four valenced statements (one of each valence for both targets) (Figure 2). However, Experiment 3 used four different behavioral statements, one that was high diagnostic positive (HDP), one that was low

diagnostic positive (LDP), one that was high diagnostic negative (HDN) and one that was low diagnostic negative (LDN). As such, there was an imbalance in the information participants received about Bob and Jake such that for one person the negative information was more diagnostic and for the other person the positive information was more diagnostic (Table 1). In contrast to Experiments 1 and 2, participants did not see images related to the statements during learning.

After this phase, participants completed the AMP and self-reported rating task (Time 1). Importantly, both tasks did not include presentation of memory recall cues. Participants also completed the AMP and self-reported rating task a second time (Time 2), but in between these two evaluation phases, participants saw the low diagnostic information they received about Bob and Jake again and were instructed to elaborate on this information. Specifically, participants received instructions to look at the information presented about one of the target persons and think about it for a while, visualizing it by forming an interactive image of the information and elaborating on it any way they can. Below these instructions, participants saw an image of Bob or Jake, the behavioral statement, and an image related to the statement (to further facilitate memory recall: Carney & Levin, 2002). Participants had to press the space bar to indicate that they read the instructions and were ready to start focusing on the information. One minute after pressing the space bar, participants saw a prompt that they could now proceed to the next phase.

After completing the Time 2 evaluation measures, we probed memory of the statements, knowledge of Chinese and demand compliant responding and asked participants to indicate to what extent they had actually formed a mental image of the low diagnostic information (on a scale from 1: not at all to 7: very much).

Results

AMP scores. We calculated AMP diagnosticity scores by subtracting the standardized mean proportion of "pleasant" responses on trials in which the prime was the person for whom low diagnostic positive information was provided from the standardized proportion of "pleasant" responses on trials in which the prime was the person for whom high diagnostic positive information was provided. AMP diagnosticity scores were significantly reduced at Time 2 compared to Time 1, $t(111)=2.49$, $p=.007$, $BF_1=7.21$, $d_z= 0.24$. At Time 1, participants exhibited a diagnosticity effect, that is, a preference for the person for whom high diagnostic positive information was provided ($M=0.38$, $SD=1.43$), $t(111)=2.82$, $p=.003$, $BF_1=14.36$, $d_z= 0.27$. However, participants did not exhibit a significant diagnosticity effect at Time 2 ($M=-0.06$, $SD=1.39$), $t(111)=-0.44$, $p=.67$, $BF_0=4.10$, $d_z= -0.04$ (Figure 3).

Self-reported rating scores. Self-reported rating scores that indicate a diagnosticity effect were computed by subtracting the standardized mean self-reported rating on warmth and liking rating scales (collapsed) for the person for whom low diagnostic positive information was provided from the standardized mean self-reported rating on warmth and liking rating scales (collapsed) for the person for whom high diagnostic positive information was provided. The mean correlation of self-reported rating and AMP diagnosticity scores was moderate, at respectively $r(110)=.27$, for Time 1 scores, and $r(110)=.43$, for Time 2 scores, ps$<.005$. The diagnosticity effect on self-reported rating scores was significantly reduced at Time 2 compared to Time 1, $t(111)=6.46$, $p<.001$, $BF_1>1000$, $d_z= 0.61$. At Time 1, participants exhibited a strong diagnosticity effect ($M=1.41$, $SD=1.13$), $t(111)=13.20$, $p<.001$, $BF_1>1000$, $d_z= 1.25$. Participants also exhibited a diagnosticity effect at Time 2 but it was strongly reduced ($M=0.59$, $SD=1.66$), $t(111)=3.78$, $p<.001$, $BF_1=177.71$, $d_z= 0.36$ (Figure 3).

Additional (exploratory) analyses. We also performed an ANOVA on standardized diagnosticity scores together (AMP and self-reported rating scores) which revealed a stronger diagnosticity effect for self-reported rating scores but no interaction with Time of Evaluation. We also examined results when excluding ten participants who indicated they did not engage strongly in rehearsal (score<4) and five participants who indicated that they might have been demand compliant. The overall pattern of results, however, did not change depending on these exclusions.

Discussion

Results of Experiment 3 indicate that the presentation of elaborative rehearsal instructions for low diagnostic valenced information strongly moderates diagnosticity effects on evaluation, providing further evidence that manipulation of memory recall influences effects of valenced information on evaluation even in the context of diagnostic information. When participants received instructions to engage in elaborative rehearsal of relatively low diagnostic valenced information about one of two target people, this eliminated the preference for the person who participants learned more diagnostic positive information about. Notably, and in contrast to Experiments 1 and 2, the memory recall manipulation also influenced self-reported evaluations. Interestingly, however, participants still indicated a (reduced) preference in line with the diagnosticity manipulation on self-reported but not automatic evaluations after elaborative rehearsal of the low diagnostic information.

It is important to note, however, that the observed reduction in the diagnosticity effect at Time 2 is not necessarily the result of the memory recall manipulation (i.e., of elaborative rehearsal). Specifically, it is possible that this reduction is the result of passage of time, or due to more frequent or more recent exposure to the elaborated information.

Experiment 4

To deal with the limitations of Experiment 3, we performed another experiment that investigated effects of elaborative rehearsal of low diagnostic information on evaluation. To rule out alternative explanations in terms of passage of time, recency, and frequency of exposure, we used only a single measurement and asked participants to elaborate on one of the person-behavior combinations immediately after their presentation. Therefore, we presented information about three different targets, with one target acting as a no-elaboration control for the elaboration target. Specifically, participants first learned 6 pieces of valenced information: (a) one low diagnostic positive (or negative) and one high diagnostic negative (or positive) behavior of one person (elaboration target), (b) one low diagnostic positive (or negative) and one high diagnostic negative (or positive) behavior of another person (control target), matched to the behavioral information about the elaboration target, and (c) one low diagnostic negative (or positive) and one high diagnostic positive (or negative) behavior about a third person (contrast target). Immediately after participants see the low diagnostic information about the elaboration target, they receive instructions to elaborate on this information. After learning, evaluations of the three targets were measured. This design allows us to compare the diagnosticity effect without elaboration (no-elaboration control versus contrast target) to the diagnosticity effect with elaboration (elaboration versus contrast target), which provides an estimate of the effect of elaboration without confounding influences such as passage of time, recency of information, or mere exposure.

There were also four other novel aspects of Experiment 4. First, we included a different set of stimuli (i.e., new statements, images, and names) to examine generalizability of results (Yarkoni, 2019). Second, we included a different demand compliance question to allow a better test of whether the effect of elaboration specifically might result from processes related to participants changing their evaluations to comply with researcher demand. Third, in contrast to Experiment 3,

for half the participants, elaboration was on positive low diagnostic information whereas the other participants elaborated on negative low diagnostic information. Fourth, participants had to write down their elaboration to test whether participants actually engaged in elaborative rehearsal.

We pre-registered the hypotheses that, in accordance with Experiment 3, elaborative rehearsal would influence both AMP and self-reported liking scores. We also hypothesized that the relative impact of elaborative rehearsal compared to diagnosticity would be bigger on AMP scores such that the diagnosticity effect would be robustly significant on self-reported ratings after elaborative rehearsal but not on AMP scores.

Method

Participants. A total of 250 Prolific Academic participants were recruited with a similar sampling plan as Experiments 1-3 with the Bayesian stopping rule dependent on two *t*-tests testing the effect of elaborative rehearsal on AMP and self-reported liking scores. Data were excluded from participants who (a) did not fully complete all questions and tasks (10 participants: 4.00%), (b) had completed AMP blocks using only one response key (21 participants: 8.40%), (c) indicated they had recognized some of the Chinese ideographs (4 participants: 1.60%), (d) made at least one error on the questions that probed memory for the elaborated valenced information (34 participants: 13.60%), or (e) did not provide elaboration about the correct information or noted issues when performing the experiment (2 participants: 0.80%). Analyses were performed on the data of 179 participants (105 women, mean age=33.78 years, *SD*=14.05).

Design. The experiment used a mixed design involving one focal within-subjects factor with three conditions: Target Person (elaboration vs. contrast vs. no-elaboration control target). We also counterbalanced the following method factors between-subjects: Information Order (the order in which people receive information about the three target persons), Target Name and Picture

(assignment of name and picture to the three target persons), Target Information (assignment of the sets of information to the target persons), Elaboration Valence (elaboration on positive vs. negative information), and Evaluation Task Order.

Materials. Three attitude targets (Mike, Paul, and Brian) were represented by three pictures of White men selected as most evaluatively neutral in the rating study referred to above. We also selected two high diagnostic positive (HDP), two low diagnostic positive (LDP), two high diagnostic negative (HDN) and two low diagnostic negative (LDN) behavioral statements based on the diagnosticity and valence ratings in the statement rating study (see Appendix).

Procedure. The experiment started with the evaluative learning task in which participants learned about Mike, Paul, and Brian with four statements (two neutral and one of each valence for all targets). During this phase and immediately after reading the low diagnostic statement about the elaboration target, participants were asked to think about the last piece of behavioral information they just read for a while, to visualize the behavior and elaborate on the information (writing down the elaboration).

After the learning task, participants completed the AMP and self-reported rating task (Figure 2). We then probed memory of the statements, knowledge of Chinese and demand compliant responding and asked participants to indicate to what extent they had actually formed a mental image of the elaboration information. Note that demand compliance was now probed with two questions both for AMP and self-reported ratings. They specifically probed whether responses were based on what they thought the researchers expected and whether participants changed their responses to more strongly take into account the elaboration information because they thought that the researcher wanted them to do so.

Results

AMP scores. For participants in the negative information elaboration condition, we calculated AMP diagnosticity scores with elaboration by subtracting the standardized mean proportion of "pleasant" responses on trials in which the prime is the contrast target from the standardized mean proportion of "pleasant" responses on trials in which the prime is the elaboration target. AMP diagnosticity scores without elaboration were computed by subtracting the standardized mean proportion of "pleasant" responses on trials in which the prime is the contrast target from the standardized mean proportion of "pleasant" responses on trials in which the prime is the no-elaboration control target. Scores were reversed for participants in the positive elaboration condition.

AMP diagnosticity scores with elaboration were significantly reduced compared to AMP diagnosticity scores without elaboration, $t(178)=2.43$, $p=.008$, $BF_1=6.21$, $d_z= 0.16$. AMP diagnosticity scores without elaboration revealed a significant preference for the person for whom high diagnostic positive (or low diagnostic negative) information was provided ($M=0.37$, $SD=1.32$), $t(178)=3.75$, $p<.001$, $BF_1=198.16$, $d_z= 0.28$, whereas AMP diagnosticity scores with elaboration did not reveal this preference ($M=0.17$, $SD=1.27$), $t(178)=1.79$, $p=.075$, $BF_0=2.51$, $d_z= 0.13$ (Figure 3).

Self-reported rating scores. Self-reported diagnosticity scores with and without elaboration were computed by subtracting the standardized mean explicit rating on warmth and liking rating scales (collapsed) for the contrast target from the standardized mean explicit rating on warmth and liking rating scales (collapsed) for the elaboration target or the control target. Scores were reversed for participants in the positive elaboration condition. The mean correlation of self-reported rating and AMP diagnosticity scores was moderate, at respectively $r(177)=.28$, for scores with elaboration, and $r(177)=.43$, for scores without elaboration, $ps<.001$.

Self-reported diagnosticity scores with elaboration were significantly reduced compared to diagnosticity scores without elaboration, $t(178)=4.13$, $p<.001$, $BF_1=693.03$, $d_z= 0.25$. Both AMP diagnosticity scores with and without elaboration revealed a significant preference for the person for whom high diagnostic positive (or low diagnostic negative) information was provided (with elaboration: $M=0.72$, $SD=1.34$, $t[178]=7.14$, $p<.001$, $BF_1=693.03$, $d_z= 0.53$, without elaboration: $M=1.06$, $SD=1.38$, $t[178]=10.30$, $p<.001$, $BF_1>1000$, $d_z= 0.77$) (Figure 3).

Additional (exploratory) analyses. We also performed an ANOVA on standardized diagnosticity scores of both measures which revealed a stronger diagnosticity effect for self-reported rating scores but no interaction with elaboration. We also examined results when excluding 18 participants who indicated that they might have been demand compliant on any of the demand compliance questions. The overall pattern of results, however, did not change depending on this exclusion.

Discussion

Results of Experiment 4 replicate and extend the results of Experiment 3, showing that the presentation of elaborative rehearsal instructions for low diagnostic valenced information moderates diagnosticity effects on more automatic and self-reported evaluations. This effect was not the result of the confounding influences of time or recency and was not moderated by whether elaboration is on positive or negative information. Similar to Experiment 3, we also observed that, after elaboration on low diagnostic information, participants still showed a significant diagnosticity effect on self-reported ratings but not on AMP scores.

General Discussion

We examined the role of memory recall in effects of diagnostic valenced information on evaluation. In Experiments 1 and 2, participants learned two separate pieces of positive and

negative information about target people that were either matched in diagnosticity (Experiment 1) or one piece was more diagnostic than the other piece (Experiment 2). Both experiments showed that the presentation of a semantically related cue during evaluation influenced more automatic evaluations as probed with the AMP but not self-reported evaluations. In Experiments 3 and 4, participants learned high diagnostic positive and low diagnostic negative information about one (or two) target(s) and high diagnostic negative and low diagnostic positive information about another (or two) target(s). Results showed that a different manipulation of memory recall that involved providing instructions to engage in elaborative rehearsal about one of the pieces of low diagnostic information influenced both more automatic and self-reported evaluations in the direction of this information. This manipulation eliminated diagnosticity effects on AMP scores and reduced them on self-reported liking scores.

Theoretical implications

At the theoretical level, the current results provide further support for the idea that processes that take into account the diagnosticity of information strongly determine (automatic) evaluation (Cone et al., 2017; De Houwer et al., 2020). However, our results also provide evidence for an important moderator of the effects of the diagnosticity of valenced information on evaluation. Specifically, manipulations of the ease of memory recall of valenced information can bias evaluations in-line with but also against diagnosticity effects. For instance, in Experiments 3 and 4, the effect of diagnosticity on evaluations was reduced when memory recall for low diagnostic information of opposite valence was facilitated. To accommodate these results, evaluation theories should specify how both memory recall and diagnosticity influence evaluation.

Interestingly, we also found evidence for possible dissociations between effects of diagnosticity and memory recall manipulations on more automatic and self-reported evaluations.

First, the observed effect of diagnosticity in Experiments 3 and 4 was bigger on self-reported evaluations. This accords with other findings (e.g., Van Dessel et al., 2020) and might relate to the fact that indirect measures such as the AMP have worse psychometric properties and might be more sensitive to some confounding influences (e.g., processes unrelated to evaluation of the prime). On the other hand, it is also possible that self-report measures are more sensitive to other confounding influences (e.g., demand compliance processes). Second, compared to diagnosticity, memory recall manipulations seemed to influence AMP scores to a relatively stronger extent. In Experiments 1-2, effects of semantic contextual cueing were only observed on AMP scores. In Experiments 3-4, the effect of elaborative rehearsal eliminated the effect of diagnosticity only on AMP scores. This might suggest that the differential integration of diagnosticity and memory recall influences could be a key factor in observed dissociations between evaluations measured under different (automaticity) conditions (Gawronski & Brannon, 2019).

The observed effects of memory recall manipulations on AMP scores were predicted based on an inferential account which postulates that evaluations depend on evaluative inferences (Van Dessel, Hughes, et al., 2019). These inferences are thought to take into account the relevance of valenced information for current (evaluative) goals, explaining the generally strong diagnosticity effects. From this perspective, however, aiding memory recall of valenced information on the basis of relevant semantic cues and elaborative rehearsal of valenced information also facilitates use of this information for evaluation. We predicted this effect specifically for evaluations measured with an AMP because in this task the opportunity or motivation of participants to carefully integrate difficult to retrieve information in evaluations might be reduced such that diagnosticity is less predominant. We argued that, opportunity and motivation for careful weighing of information is higher in self-reported tasks which might give rise to dissociations (see Van Dessel et al., 2016).

The fact that results support the prediction that memory cueing and elaborative rehearsal influence AMP scores provides support for this framework but not all our results were predicted a priori (e.g., the dissociation in Experiments 1 and 2). These findings help to constrain further development of this (and other) inferential evaluation model(s), an endeavor that has only recently been kickstarted yet could be of high theoretical and practical use (see Van Dessel et al., 2018). For instance, results might suggest that elaborative rehearsal is more potent than semantic cueing in facilitating evaluative inferences (although one should also take into account the time between manipulation and learning: see Zanon et al., 2014). Future research might try to more precisely elucidate evaluative inferences and the factors that determine them. For instance, in accordance with predictive coding theories, context-dependence of evaluative effects could be modelled computationally on the basis of new data (Sanborn & Chater, 2016).

Of course, our results cannot distinguish between broad classes of models such as inferential and dual-process models (in fact, this is impossible on the basis of any set of data). Effects of memory recall manipulations can for instance also be accommodated by dual-process models such as the Associative-Propositional Evaluation (APE) Model (Gawronski & Bodenhausen, 2006). From the APE model perspective, inferential processes might determine diagnosticity effects on evaluation whereas associative processes (i.e., associative spreading of activation) might determine effects of memory recall manipulations. Notably, however, Brannon and Gawronski (2018) did not find effects of contextual cueing of diagnostic information that were predicted on the basis of a dual-process framework (Gawronski & Cesario, 2013). As noted above, an inferential theory could explain this dissociation with the current findings based on the idea that the current experiments used cues that were more relevant for information recall in the context of evaluation (because it was related to the content of the behavioural statements). Still, it is at least

possible that dual-process models which postulate that both associative and inferential processes contribute to both memory recall and diagnosticity effects can accommodate these findings. It seems more difficult to explain this dissociation with other dual-process theories that relate cueing effects to the activation of associations that are learned on the basis of pairings (e.g., Rydell & Gawronski, 2009) or dual-process models which claim that automatic evaluations are difficult to change (e.g., Rydell & McConnell, 2006).

Practical implications

At the practical level, the current results are also informative as they suggest that both memory recall and diagnosticity-related influences determine evaluation. This information can be of practical use when one wants to influence evaluations. Imagine a situation in which a person has learned diagnostic valenced information about someone and wants to strengthen or reduce effects of this information on their evaluative responses towards this person. For instance, after positive interactions with a colleague, one might learn about diagnostic negative (e.g., aggressive or harassing) behavior of this person. To facilitate lasting positive interaction, one may want to reduce effects of this information on evaluation. Alternatively, to prevent negative events (e.g., harassment) one may want to make sure that this diagnostic information has maximum impact. Our results suggest that a technique that focuses on the recall of the diagnostic or alternative information from memory could then prove useful. Moreover, they suggest that the installation of semantically related cues in the environment might only have weak effects whereas elaborative rehearsal might be more effective. Note that such techniques might (already) be used for more malicious purposes (e.g., people might facilitate recall of or repeat and elaborate on more positive but less diagnostic information to bury important negative information about a product or a politician). We hope our results can improve awareness of such techniques to arm people against their influences.

Limitations and future directions

One important limitation of the current findings is that effects were probed with two evaluation measures that are very different (i.e., AMP and self-reported liking questions). As a result, caution is warranted when interpreting observed dissociations because they might be due to a number of reasons. For instance, because self-reported liking ratings might be influenced less by non-evaluative influences than AMP scores and might be more reliable indices of evaluation, they can sometimes reveal much stronger effects of manipulations even if the relative impact of the manipulation is weaker (see also Van Dessel et al., 2020). Importantly, one should also not assume that dissociations reflect differences in the general construct of 'automatic or implicit attitudes' and 'self-report or explicit attitudes'. For instance, automatic evaluations measured with the AMP (as well as with other indirect evaluation measures) are not entirely automatic (i.e., they also occur under conditions of non-automaticity, e.g., awareness: Cummins et al., 2019). In general, it seems more useful not to make strong distinctions between evaluations measured with indirect vs. self-reported measures but rather to carefully examine dissociations in terms of differences in the specific measurement conditions (Van Dessel, Cummins, et al., 2020). We contrasted the AMP and self-reported ratings given prior results suggesting reduced opportunity or motivation to integrate the diagnosticity of valenced information in the former measure (Gawronski & Rydell, 2009). It is entirely possible that other self-report and automatic evaluation measures show different effects of memory recall manipulations (under certain conditions).

A second limitation of the current set of studies is that we did not assess the impact of diagnosticity and memory recall manipulations entirely independent from one another. In fact, manipulation of diagnosticity without manipulation of memory recall might be very difficult (e.g., memory recall could influence perceived diagnosticity and vice versa). For theoretical reasons, we

focused our investigation on the effect that memory recall manipulations have on the influence of diagnostic information on evaluation. In Experiments 3 and 4, the design allowed better assessment of the distinct influences (because the manipulation of diagnosticity took place at different times or included a control), but even here both manipulations might still have some effect on the other construct. An interesting avenue for future research might be to examine the precise relation between these two distinct constructs (e.g., by measuring memory recall and diagnosticity in relation to the manipulations).

Another potential confounding influence in the current studies is demand compliance. It is possible that participants use cued or elaborated valenced information more not because this information was more easy to recall but rather because participants thought this was what the experimenter wanted and they complied with this demand. Note, however, that few participants indicated demand compliance when asked on what basis they had emitted their evaluative responses (also in Experiment 4 which directly asked about demand related to the memory recall manipulation) and exclusion of identified participants did not significantly influence results. Of course, it is possible that demand compliance still contributed to effects and that participants just did not accurately report their demand compliance (e.g., because they believed that not reporting demand compliance would please the experimenter).

Note that this is only one of the first investigations of the role of memory recall in relation to effects of (diagnostic) information on evaluation. We are convinced that continuing this line of research will prove useful both for theoretical and practical purposes (e.g., by running studies that include real-life stimuli in real-life contexts). This continued research might benefit from examining effects of other memory recall manipulations (a distinction can be made between manipulations that focus on encoding, storage, or retrieval of information from memory: Gast,

2018), as well as from testing the mediation of diagnosticity effects by changes in memory recall. Furthermore, future studies might examine how exactly computation of the relevance of evaluative information might occur (e.g., see Dalege et al., 2018, for a relevant example of such an approach) and how this is influenced both by information diagnosticity and memory recall in determining (more automatic and more controlled) evaluation.

References

Bartsch, L.M., Singmann, H., & Oberauer, K. (2018). The effects of refreshing and elaboration on working memory performance, and their contributions to long-term memory formation. *Memory & Cognition, 46*, 796-808.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1).

Benedict, T., Richter, J., & Gast, A. (2019). The influence of misinformation manipulations on evaluative conditioning. *Acta Psychologica, 194*, 28-36.

Brannon, S. M., & Gawronski, B. (2017). A second chance for first impressions? Exploring the context (in)dependent updating of implicit evaluations. *Social Psychological and Personality Science, 3*, 275-283.

Carney, R. N., & Levin, J. R. (2002). Pictorial illustrations still improve students' learning from text. *Educational Psychology Review, 14*, 5-26.

Cone, J., & Ferguson, M. J. (2015). He Did what?: The Role of Diagnosticity in Revising Implicit evaluations. *Journal of Personality and Social Psychology, 108*, 37-57.

Cone, J., Flaharty, K., & Ferguson, M. J. (2019). Believability of evidence matters for correcting social impressions. *Proceedings of the National Academy of Sciences of the United States of America*. doi:10.1073/pnas.1903222116

Cone, J., Mann, T. C., & Ferguson, M. J. (2017). Changing our implicit minds: How, when, and why implicit evaluations can be rapidly revised. In J. M. Olson (Ed.), *Advances in experimental social psychology: Vol. 56. Advances in experimental social psychology* (pp. 131-199). San Diego, CA, US: Elsevier Academic Press.

Cummins, J., Hussey, I., & Hughes, S. (2019). The AMPeror's New Clothes: Performance on the Affect Misattribution Procedure is Mainly Driven by Awareness of Influence of the Primes.

Dalege, J., Borsboom, D., van Harreveld, F., & van der Maas, H. L. J. (2018). The Attitudinal Entropy (AE) Framework as a General Theory of Individual Attitudes. *Psychological Inquiry, 29*, 175-193.

De Houwer, J. (2014). A Propositional Model of Implicit Evaluation. *Social and Personality Psychology Compass, 8*, 342-353.

De Houwer, J., Van Dessel, P., & Moran, T. (2020). Attitudes Beyond Associations: On the Role of Propositional Representations in Stimulus Evaluation. *Advances in Experimental Social Psychology*. (pp. 127– 183). Amsterdam: Elsevier.

Fox, J., & Weisberg, S. (2011). An R Companion to Applied Regression (2nd ed.): Thousand Oaks, CA: Sage

Gast, A. (2018). A declarative memory model of evaluative conditioning. *Social Psychological Bulletin,* 13(3), e28590.

Gast, A., Richter, J., Benedict, T., & Ruszpel, B. (2021). *Memory Activation in Attitude Formation. Unpublished Manuscript.*

Gawronski, B., & Bodenhausen, G. V. (2006). Associative and propositional processes in evaluation: An integrative review of implicit and explicit attitude change. *Psychological Bulletin, 132*, 692-731.

Gawronski, B., & Brannon, S. M. (2019). Attitudes and the implicit-explicit dualism. In D. Albarracín & B. T. Johnson (Eds.), *The handbook of attitudes. Volume 1: Basic principles* (2nd edition, pp. 158-196). New York, NY: Routledge.

Gawronski, B., & Cesario, J. (2013). Of mice and men: What animal research can tell us about context effects on automatic responses in humans. *Personality and Social Psychology Review, 17,* 187-215.

Gawronski, B., & De Houwer, J. (2014). Implicit measures in social and personality psychology. In H. T. Reis, & C. M. Judd (Eds.), *Handbook of research methods in social and personality psychology* (2nd edition, pp. 283–310). New York: Cambridge University Press.

Gawronski, B., Hu, X., Rydell, R. J., Vervliet, B., & De Houwer, J. (2015). Generalization and contextualization in automatic evaluation revisited: A meta-analysis of successful and failed replications. *Journal of Experimental Psychology: General, 144*, e50–e64.

Gawronski, B., Rydell, R. J., De Houwer, J., Brannon, S. M., Ye, Y., Vervliet, B., & Hu, X. (2018). Contextualized attitude change. *Advances in Experimental Social Psychology, 57,* 1-52

Goldstein, B. (2011). *Cognitive Psychology: Connecting Mind, Research, and Everyday Experience* (3rd ed.). Belmont, CA: Wadsworth.

Hansen, J., & Wänke, M. (2008). It's the Difference that Counts: Expectancy/Experience Discrepancy Moderates the Use of Ease of Retrieval in Attitude Judgments. *Social Cognition, 4,* 447-468.

Hughes, S., De Houwer, J., & Perugini, M. (2016). The functional-cognitive framework for psychological research: Controversies and resolutions. *International Journal of Psychology, 51,* 4-14.

Keller, K. L. (1987), Memory Factors in Advertising: The Effects of Advertising Retrieval Cues on Brand Evaluations. *Journal of Consumer Research, 14,* 316-333.

Ma, D. S., Correll, J., & Wittenbrink, B. (2015). The Chicago face database: A free stimulus set of faces and norming data. *Behavior research methods, 47(4*),1122-1134.

Mann, T., Cone, J., Heggeseth, B., & Ferguson, M. (2019). Updating implicit impressions: New evidence on intentionality and the Affect Misattribution Procedure. *Journal of Personality and Social Psychology, 116(3),* 349-374.

Mann, T. C., & Ferguson, M. J. (2015). Can we undo our first impressions? The role of reinterpretation in reversing implicit evaluations. *Journal of Personality and Social Psychology, 108,* 823-849.

Mann, T. C., & Ferguson, M. J. (2017). Reversing implicit first impressions through reinterpretation after a two-day delay. *Journal of Experimental Social Psychology, 68,* 122–127.

Payne, B. K., Cheng, C. M., Govorun, O., & Stewart, B. D. (2005). An inkblot for attitudes: Affect misattribution as implicit measurement. *Journal of Personality and Social Psychology, 89,* 277-293.

Richter, J., & Gast, A. (2017). Distributed practice can boost evaluative conditioning by increasing memory for the stimulus pairs. *Acta Psychologica, 179,* 1-13.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review, 16,* 225–237. doi: 10.3758/PBR.16.2.225

Rydell, R. J., & Gawronski, B. (2009). I like you, I like you not: Understanding the formation of context-dependent automatic attitudes. *Cognition and Emotion, 23,* 1118–1152.

Rydell, R. J., & McConnell, A. R. (2006). Understanding implicit and explicit attitude change: A systems of reasoning analysis. *Journal of Personality and Social Psychology, 91*, 995-1008.

Sanborn A. N., & Chater, N. (2016) Bayesian brains without probabilities. *Trends in Cognitive Sciences, 20,* 883–893.

Schönbrodt, F. D., Wagenmakers, E.-J., Zehetleitner, M., & Perugini, M. (2017). Sequential hypothesis testing with Bayes factors: Efficiently testing mean differences. *Psychological Methods, 22,* 322-339.

Skowronski J.J., & Carlston, D.E. (1987). Social Judgment and Social Memory: The Role of Cue Diagnosticity in Negativity, Positivity, and Extremity Biases. *Journal of Personality and Social Psychology*, *52*, 689-699.

Van Dessel, P., Cone, J., Gast, A., & De Houwer, J. (2020). The Impact of Valenced Verbal Information on Implicit and Explicit Evaluation: The Role of Information Diagnosticity, Primacy, and Memory Cueing. *Cognition & Emotion*, *34*, 74-85.

Van Dessel, P., & De Houwer, J. (2019). Hypnotic Suggestions Can Induce Rapid Change in Implicit Attitudes. *Psychological Science, 30*, 1362-1370.

Van Dessel, P., De Houwer, J., Gast, A., Smith, C. T., & De Schryver, M. (2016). Instructing implicit processes: When instructions to approach or avoid influence implicit but not explicit evaluation. *Journal of Experimental Social Psychology, 63*, 1-9.

Van Dessel, P., Hughes, S., & De Houwer, J. (2019). How Do Actions Influence Attitudes? An Inferential Account of the Impact of Action Performance on Stimulus Evaluation. *Personality and Social Psychology Review, 23*, 267-284.

Wänke, M., Bless, H., & Biller, B. (1996). Subjective experience versus content of information in the construction of attitude judgments. *Personality and Social Psychology Bulletin, 22*, 1150-1113.

Wood, W. (1982). Retrieval of attitude-relevant information from memory: Effects on susceptibility to persuasion and on intrinsic motivation. *Journal of Personality and Social Psychology. 42*, 798–810.

Yarkoni, T. (2019). The generalizability crisis. https://doi.org/10.31234/osf.io/jqw35

Zanon, R., De Houwer, J., Gast, A., & Smith, C. T. (2014). When does relational information influence evaluative conditioning? *The Quarterly Journal of Experimental Psychology, 67*, 2105-2122. Doi: 10.1080/17470218.2014.907324

Zhou, H., & Fishbach, A. (2016). The Pitfall of Experimenting on the Web: How Unattended Selective Attrition Leads to Surprising (Yet False) Research Conclusions. *Journal of Personality and Social Psychology, 11*, 493-504.