Behavioral and Physiological Evidence Challenges the Automatic Acquisition of Evaluations

Article *in* Current Directions in Psychological Science · July 2020

CITATIONS
0

READS
594

2 authors:

Olivier Corneille
Université Catholique de Louvain - UCLouvain
137 PUBLICATIONS **4,002** CITATIONS
SEE PROFILE

Gaëtan Mertens
Tilburg University
47 PUBLICATIONS **256** CITATIONS
SEE PROFILE

Behavioral and Physiological Evidence Challenges the Automatic Acquisition of Evaluations.

Olivier Corneille, UCLouvain.

Gaëtan Mertens, Tilburg University & Utrecht University.

Manuscript in press in ***Current Directions in Psychological Science***

Olivier Corneille (Contact author). Psychological Sciences Research Institute, UCLouvain, 10 Place du Cardinal Mercier, 1348 Louvain-la-Neuve, Belgium. Email: olivier.corneille@uclouvain.be

Behavioral and Physiological Evidence Challenges the Automatic Acquisition of Evaluations.

1. The Automatic Acquisition of Evaluations.

An influential view pervading psychological research (e.g., in social, consumer, health, and clinical psychology) is that evaluations can be acquired through an associative/affective learning mode that automatically registers mere stimuli co-occurrences encountered in the environment. In attitude research, associative learning is thought to allow for the unconscious, efficient, involuntary, and uncontrollable formation of evaluations and to be insensitive to the relational meaning of stimuli co-occurrences (e.g., Gawronski & Bodenhausen, 2014). Likewise, influential theories of fear learning posit that fear can be automatically acquired (i.e., without consciousness, effort, or intention, and uncontrollably), resulting in non-conscious and unqualified associative representations that automatically elicit fear when activated (e.g., LeDoux & Pine, 2016; Öhman & Mineka, 2001).

This automatic learning view implies a high susceptibility of individuals to social and environmental influences. For instance, people may fall prey to a negative political advertisement if the information it contains creates, uncontrollably or unconsciously, a negative evaluation of the targeted candidate. It also often assumes a difficulty to change automatic evaluative responses based on mere verbal information. For instance, a fear of dogs may be difficult to change using verbal instruction if the fear response reflects the automatic activation of an inaccessible or unconscious mental association between a dog and having been bitten.

These questions have been mostly investigated using conditioning procedures. In attitude research, a neutral stimulus (the conditioned stimulus, or the CS; e.g., a neutral face) is paired with a positive or negative stimulus (the unconditioned stimulus, or the US: e.g., a pleasant or unpleasant sound or picture). An evaluative conditioning effect is said to occur

when the evaluation of the CS changes after the CS-US pairing. In fear-conditioning research, typically aversive USs (e.g., an electric shock, a loud noise) are used, and subjective, behavioral, and physiological responses related to fear and arousal (e.g., distress ratings, avoidance responses, skin conductance responses) are measured (e.g., Lipp, 2006). Evaluative and fear conditioning procedures involve a simple associative procedure (i.e., pairing CSs with USs) that is considered ideally suited to the investigation of associative/affective learning.

2. Testing the Automatic Learning of Evaluations Using "Explicit", "Implicit" and Physiological Measures.

The automatic learning of evaluations has been tested using "explicit", "implicit", and physiological measures. Explicit measures are self-reported evaluations, such as good/bad judgments or direct scale ratings about an attitude object. Implicit measures are less clearly defined (see Corneille & Hütter, 2020), but generally rely on indirect behavioral tasks that reduce participants' control and deliberation during measurement. The Implicit Association Test is considered the gold standard of "implicit" evaluative measures. In this task, respondents may sort faster positive stimuli alongside category A (e.g., thin people) and negative stimuli alongside category B (e.g., obese people). The extent to which they do so (or the opposite) indicates their implicit preference for one category over the other. Measurement outcomes of this sort are thought to reflect biased mental associations, commonly referred to as "implicit biases" or "unconscious biases." Physiological measures record participants' physiological responses to stimuli. Skin conductance responses and potentiation of the startle-reflex are examples of such measures, and they capture a mix of arousal and evaluation (Lonsdorf et al., 2017). Whereas "explicit" and "implicit" evaluative measures are frequently used in social psychological, social cognition, and consumer research, physiological measures are more common in clinical and neuroscience research.

Dual-learning theories assume that less controllable and deliberate measures of evaluations (i.e., "implicit" and physiological measures) provide a better window into associative/affective learning (e.g., Gawronski & Bodenhausen, 2014; Ledoux & Pine, 2016). This is because these measures reduce the influence of deliberate and controlled post-learning processes on task performance.

3. Growing Evidence Challenges the Automatic Acquisition of Evaluations on all Categories of Measures.

Research on evaluative conditioning has largely failed to obtain conclusive evidence for automatic evaluative learning (for a comprehensive review, see Corneille & Stahl, 2019). No evaluative learning effect is observed in procedures preventing a conscious encoding of the stimuli, such as when using short-timed, parafoveal, or visually suppressed presentations. Evaluative learning is also resource demanding. It is fully disrupted when participants are engaged in a concurrent task during the stimuli presentation. Finally, research indicates that evaluative learning is sensitive to participants' processing goals. For instance, evaluative learning effects are largely weakened when participants are distracted from processing the affective quality of the information.

Of importance too, evidence that may be considered supportive of automatic evaluative learning is unrelated to "implicit" versus "explicit" measurement. The mere exposure effect, often considered the hallmark of "preferences-that-need-no-inferences", is typically demonstrated on self-reported evaluations (e.g., "Which of these two stimuli do you prefer?") Likewise, preliminary evidence for uncontrollable evaluative learning (for instance, participants presumably unable not to start liking a stimulus paired with positive information) has been best observed on "explicit" liking measures (Hütter & Sweldens, 2018). Hence, contrary to a widespread dual-process view, that an evaluation is observed on an "implicit

task" or on a self-reported measure is unrelated to how it was acquired in the first place (Corneille & Hütter, 2020).

Besides the automaticity of evaluative learning, dual-learning models of evaluations posit that associative knowledge captures raw associations between events, unqualified by the relational meaning of these associations. For instance, people may harbor negative associations about a pain-relieving medication because of its co-occurrence with pain (i.e., the raw contingency), despite the medication relieving the pain (i.e., the relational implication of the medication). Preliminary evidence supporting the possibility of such unqualified representations or processes, however, has again been found on "explicit" (e.g., Kükken, Hütter & Holland, 2019) measures, and these effects are highly sensitive to task structure when measured with "implicit" tasks (e.g., Bading, Stahl, & Rothermund, 2020).

As it appears, differences in performance on "explicit" vs. "implicit" tasks cannot be univocally interpreted in terms of dissociations in *learning* modes. Post-learning processes may be responsible for these differences. For instance, a smaller amount of information is likely retrieved from memory when completing a speeded (e.g., an Implicit Association Test) than a non-speeded (e.g., an evaluative rating) task. Hence, divergences in outcomes on "implicit" versus "explicit" tasks may reflect divergences in retrieval, despite originating in a unitary learning process that is neither unconscious nor immune to verbal information. As a matter of fact, "implicit" attitudes can be created in a snap based on mere verbal information (e.g., De Houwer, 2009). And this is also true when it comes to *changing* evaluations - for instance, merely telling participants that Gandhi prevented his wife from taking a medical treatment that could have saved her life results in a quick reversal of scores on "implicit" measures (Van Dessel, Ye, & De Houwer, 2018). This runs counter a prevalent view that sees "implicit biases" as the result of unconsciously and slowly acquired associations.

So far, we have discussed behavioral measures. As explained above, physiological measures may provide a more sensitive test for the existence of a distinct learning mode grounded in more affective systems. Here too, however, recent research questions automatic learning and supports the sensitivity of physiological indicators of evaluations to verbal instruction (e.g., Mertens et al., 2018). For instance, a recent meta-analysis has concluded in a weak fear acquisition effect, plausibly driven by publication bias and methodological artifacts, when using subliminal stimuli (see Mertens & Engelhard, 2020). And, similar to explicit and implicit evaluative measures, fear responses can also be modulated by verbal instruction (Atlas, 2019; Mertens et al., 2018). For example, conditioned physiological responses can be rapidly abolished when participants are told that no more USs will be delivered (Luck & Lipp, 2016). It is telling that these effects have also been demonstrated using the startle-reflex (Luck & Lipp, 2015; Mertens & De Houwer, 2016), a basic defensive reflex evident within 20-120 ms (Blumenthal et al., 2005) that is considered an index of amygdala activity (Hamm & Weike, 2005).

4. Implications for theory and practice.

Implications of these findings for psychological theories are straightforward. There is less evidence for and more evidence against the automatic learning of evaluations than one may think. There is also more evidence than is commonly assumed for the role of verbal information in the acquisition and change of evaluations. Alternative single learning models should, therefore, be considered a more parsimonious alternative to dual-learning models. Examples of such models are propositional models (De Houwer, 2009), goal-directed models (e.g., Boddez, Moors, Mertens, & De Houwer, 2020) or retrieval-based models (e.g., Stahl & Aust, 2018). Perhaps even more important are the practical implications of these recent findings, which we now briefly discuss in the social, consumer, and health and clinical psychology domains.

Social psychology: An influential social-psychological view that found its way to the general public, and that is currently inspiring large-scale social interventions, is that "mental associations" automatically formed about social groups (coined "unconscious bias" or "implicit bias") influence people's judgment and behavior. Drawing on automatic evaluative learning may facilitate people's acceptance of their prejudice. Unfortunately, this may also normalize prejudice (Eberhardt & Banks, 2019) and undermine accountability for social discrimination (Daumeyer, Onyeador, Brown, & Richeson, 2019). That is, if "mental associations" about social groups build up unconsciously and uncontrollably, one may consider that people are not so much responsible for holding them.

As discussed here, however, the evidence is scarce that evaluations may be acquired automatically. Likewise, it is unclear whether "implicit" measures assess unconscious evaluations (e.g., Gawronski, 2019). As provocative as this may seem, even assuming "implicit" tasks do assess unconscious and automatically acquired associations, we do not even know whether they have any behavioral impact at all. Because outcomes on "implicit" measures can be neither coherently nor distinctly induced, their causal role remains unestablished (for a recent discussion, see Corneille & Hütter, 2020).

Again, this conclusion runs against a popular view holding that "unconscious bias infiltrates every arena of life" and "can be just as devastating as the harm done by explicit racism, sexism or homophobia » (Eberhardt & Banks, 2019). Admittedly, "unconscious bias" may be understood here in terms of automatically enacted behaviors as opposed to automatically acquired and unconsciously held "mental associations" presumably underlying these behaviors (e.g., Gawronski, Ledgerwood & Eastwick, in press). However, this alternative understanding is uncommon, and it would be highly problematic if the "unconscious bias" construct were applied both to the cause and to its consequences.

Similar conclusions apply to personality research that has theorized discrepancies in so-called "implicit" and "explicit" self-esteem. If "implicit" measures of the self are not diagnostic of distinct learning processes and representations (see also Schimmack, 2019), one is left to wonder how discrepancies in these measures should be interpreted, as well as how and why "implicit" self-esteem should be changed at all.

Problematic theorization on the "implicit" construct is widespread in social psychological research and this can be consequential (for a comprehensive discussion, see Corneille & Hütter, 2020). During the Covid-19 outbreak, the *Society for Personality and Social Psychology* featured research inviting people to indulge in comfort food to feed "primitive and implicit feelings of being cared for and loved" (SPSP, 2020). This surprising recommendation (literally: eat high-caloric food to please your unconscious inner self) may contribute to another epidemic: the obesity one.

Consumer psychology: Consumers may be upset about the use of subliminal conditioning (such as when briefly pairing Al Gore with "Rats" in a political advertisement). In all likelihood, however, subliminal conditioning is largely ineffective for changing people's evaluations (see Corneille & Stahl, 2019). Ironically, people generally feel at ease with blatant social influence techniques (e.g., a presidential candidate featured against the American flag) that are much more influential. This is not to say that automatic processes, including unconscious ones, play no role in consumer or social behavior. They certainly do. Yet, as we have already pointed out, many of these effects may occur at a post-learning stage.

Consistent with associative/affective learning views, companies have relied on "implicit" measures to assess "unconscious" preferences in their consumers. The company Pizza Hut, for instance, has used eye-tracking technology to assess unconscious pizza topping preferences in consumers. There is, however, no indication that eye-tracking is relevant to the study of "unconsciously" acquired or "unconsciously" held pizza topping preferences (not to

speak of whether eye-tracking provides a valid measure of preferences at all). And, even if it was, there is no evidence either that this measure of preference would allow for tastier experiences. It is largely assumed, when using "implicit" measures, that testers know more about respondents than respondents know about themselves. This, however, is a philosophically and empirically tricky assumption. And, if it is incorrect, it may have ethically questionable consequences (e.g. feeding consumers with pizza topping they don't like, persuading individuals that they hold conflicting conscious and unconscious evaluations about themselves and about others).

Health and clinical psychology: That fear responding is acquired and may be changed on the basis of mere verbal instructions could be inspirational for new clinical interventions. For instance, in therapy, patients can be asked to formulate their expectations (e.g., if I walk in the park, a dog will attack me) and asked how these expectations may be tested (e.g., take a walk in the park during a busy afternoon). This approach can help identify patients' idiosyncratic expectations and allow for more targeted exposure therapy. Interventions of this sort have been successfully developed in cognitive-behavioral therapies (Hofmann, 2008). Surprisingly enough, research on these types of interventions remains surprisingly scarce in fundamental research on fear conditioning (see Carpenter, Pinaire, & Hofmann, 2019). A broader appreciation for the role of consciously accessible representations could lead to a more widespread application of such cognitive interventions to challenge dysfunctional beliefs about contingencies in the world.

5. Clarification and limitations.

We want to emphasize that we fully endorse the view that automatic processes influence people's judgments, and behaviors. Clearly, for instance, people may show efficient behavioral and physiological responses when exposed to disliked stimuli. Likewise, it would be unreasonable to assume that people are constantly aware of the determinants of their

perceptions, judgments and behaviors. Instead, we are pointing to a growing body of empirical evidence challenging the view that these responses are automatically *learned* and are insensitive to verbal information. We also question the view that some measures (e.g., "implicit" and physiological measures) would best indicate an associative/affective learning mode or qualitatively different representations. Although our analysis was mainly based on conditioning procedures (for the reasons we have explained), similar conclusions have been reached in related paradigms, such as approach-avoidance and the mere exposure effect (for an in-depth discussion, see Corneille & Stahl, 2019).

Our analysis did not include studies with brain imaging techniques and brain-lesioned populations. With regard to studies on brain-lesions and process-specific impairments, some studies have shown that region-specific lesions (e.g., amygdala damage) result in process-specific impairments (e.g., successful acquisition of fear conditioning with skin conductance responses, but no acquisition of declarative knowledge), while the reverse being true for damage in another region (i.e., hippocampal damage and unsuccessful fear conditioning, but the successful acquisition of declarative knowledge; Bechara et al., 1995).

Such findings may be seen as providing strong evidence for dual-learning theories of evaluations. However, other researchers have pointed out that double dissociations of this sort do not necessarily need to be interpreted as evidence for modularity and independent processes (e.g., Plaut, 1995). Furthermore, conflicting findings have been reported in the literature. For instance, Coppens et al. (2009) reported that patients with unilateral amygdala damage can in fact acquire conditioned fear responses after acquiring explicit knowledge of CS-US contingencies.

With regard to brain imaging studies, we believe that brain activation should not be considered a direct measure of attitude or fear. Instead, brain imaging data reflect a neural implementation level of analysis and can be logically consistent with different models of how

evaluations are acquired, retrieved, and expressed. Additionally, it can be noted that associative mental representations are not necessarily neuro-anatomically, -chemically, or –biologically more plausible than non-associative (e.g., propositional) representations (e.g., Langille & Gallistel, 2020). Given their different levels of analysis, however, psychological and neurocognitive evidence can be mutually informative to approximate a more accurate psychological and neurobiological model of how evaluations and fears are acquired and can be changed (e.g., Amodio & Ratner, 2011).

6. Conclusion

A growing body of research challenges the existence of an associative/affective formation of attitudes and fears when this learning mode is defined as automatic and impervious to verbal information. As discussed in this article, the acquisition of evaluations and fear is much less automatic than often assumed, and both can be modulated using verbal instruction, even when using "implicit" and physiological measures. Acknowledging more broadly this recent evidence is important for both theory and practice. The unconscious and resource-free learning of attitudes and fears should be questioned. In addition, the potential of verbal instructions for changing "deep-rooted" evaluations offers a promising avenue for social and clinical interventions.

References:

Amodio, D. M., & Ratner, K. G. (2011). A memory systems model of implicit social cognition. *Current Directions in Psychological Science*, *20*, 143-148.

Atlas, L. Y. (2019). How instructions shape aversive learning: higher order knowledge, reversal learning, and the role of the amygdala. *Current Opinion in Behavioral Sciences*, *26*, 121-129.

Bading, K., Stahl, C., & Rothermund, K. (2020). Why a standard IAT effect cannot provide evidence for association formation: The role of similarity construction. *Cognition and Emotion, 34*(1), 128-143. doi: 10.1080/02699931.2019.1604322.

Bechara, A., Tranel, D., Damasio, H., Adolphs, R., Rockland, C., & Damasio, A. R. (1995). Double dissociation of conditioning and declarative knowledge relative to the amygdala and hippocampus in humans. *Science*, *269*(5227), 1115-1118.

Blumenthal, T. D., Cuthbert, B. N., Filion, D. L., Hackley, S., Lipp, O. V., & Van Boxtel, A. (2005). Committee report: Guidelines for human startle eyeblink electromyographic studies. *Psychophysiology*, *42*(1), 1-15.

Boddez, Y., Moors, A., Mertens, G., & De Houwer, J. (2020). Tackling fear: Beyond associative memory activation as the only determinant of fear responding. *Neuroscience & Biobehavioral Reviews*.

Carpenter, J. K., Pinaire, M., & Hofmann, S. G. (2019). from extinction learning to anxiety treatment: Mind the gap. *Brain sciences*, *9*(7), 164.

Coppens, E., Spruyt, A., Vandenbulcke, M., Van Paesschen, W., & Vansteenwegen, D. (2009). Classically conditioned fear responses are preserved following unilateral temporal lobectomy in humans when concurrent US-expectancy ratings are used. *Neuropsychologia*, *47*(12), 2496-2503.

Corneille, O., & Hütter, M. (2020). Implicit? What do you mean? A comprehensive review of the implicitness construct in attitude research. *Personality and Social Psychology Review.*

Corneille, O., & Stahl, C. (2019). Associative attitude learning: A closer look at evidence and how it relates to attitude models. *Personality and Social Psychology Review*, *23*(2), 161-189.

Daumeyer, N. M., Onyeador, I. N., Brown, X., & Richeson, J. A. (2019). Consequences of attributing discrimination to implicit vs. explicit bias. *Journal of Experimental Social Psychology*, *84*, 103812.

De Houwer, J. (2009). The propositional approach to associative learning as an alternative for association formation models. *Learning & Behavior*, *37*(1), 1-20.

Eberhardt, J.L, Banks, S. (2019; July 12). Implicit bias puts lives in jeopardy. Can mandatory training reduce the risk? *Los Angeles Times*. Retrieved from https://www.latimes.com

Gawronski, B. (2019). Six lessons for a cogent science of implicit bias and its criticism. *Perspectives on Psychological Science, 14,* 574-595.

Gawronski, B., & Bodenhausen, G. V. (2014). The associative-propositional evaluation model: Operating principles and operating conditions of evaluation. *Dual-process theories of the social mind*, 188-203.

Gawronski, B., Ledgerwood, A., & Esatwick, P. (in press). Implicit bias and anti-discrimination policy. *Policy Insights from the Behavioral and Brain Sciences*.

Hamm, A. O., & Weike, A. I. (2005). The neuropsychology of fear learning and fear regulation. *International Journal of Psychophysiology*, *57*(1), 5-14.

Hofmann, S. G. (2008). Cognitive processes during fear acquisition and extinction in animals and humans: Implications for exposure therapy of anxiety disorders. *Clinical Psychology Review*, *28*(2), 199-210.

Hütter, M., & Sweldens, S. (2018). Dissociating controllable and uncontrollable effects of affective stimuli on attitudes and consumption. *Journal of Consumer Research, 45,* 320-349.

Kukken, N., Hütter, M., & Holland, R. W. (2019). Are there two independent evaluative conditioning effects in relational paradigms? Dissociating the effects of CS-US pairings and their meaning. *Cognition and Emotion,* 34, 1-20.

Langille, J., & Gallistel, C. (2020). Locating the engram: Should we look for plastic synapses or information-storing molecules? *Neurobiology of Learning and Memory*, *169*.

LeDoux, J. E., & Pine, D. S. (2016). Using neuroscience to help understand fear and anxiety: a two-system framework. *American Journal of Psychiatry*, *173*, 1083-1093.

Lipp, O. V. (2006). Human Fear Learning: Contemporary Procedures and Measurement.

Lonsdorf, T. B., Menz, M. M., Andreatta, M., Fullana, M. A., Golkar, A., Haaker, J., ... & Drexler, S. M. (2017). Don't fear 'fear conditioning': Methodological considerations for

the design and analysis of studies on human fear acquisition, extinction, and return of fear. *Neuroscience & Biobehavioral Reviews*, *77*, 247-285.

Luck, C. C., & Lipp, O. V. (2015). A potential pathway to the relapse of fear? Conditioned negative stimulus evaluation (but not physiological responding) resists instructed extinction. *Behaviour Research and Therapy*, *66*, 18-31.

Luck, C. C., & Lipp, O. V. (2016). Instructed extinction in human fear conditioning: History, recent developments, and future directions. *Australian Journal of Psychology*, *68*(3), 209-227.

Mertens, G., & De Houwer, J. (2016). Potentiation of the startle reflex is in line with contingency reversal instructions rather than the conditioning history. *Biological Psychology*, *113*, 91-99.

Mertens, G., Boddez, Y., Sevenster, D., Engelhard, I. M., & De Houwer, J. (2018). A review on the effects of verbal instructions in human fear conditioning: Empirical findings, theoretical considerations, and future directions. *Biological Psychology*, *137*, 49-64.

Mertens, G., & Engelhard, I. M. (2020). A systematic review and meta-analysis of the evidence for unaware fear conditioning. *Neuroscience & Biobehavioral Reviews*, *108*, 254-266.

Öhman, A., & Mineka, S. (2001). Fears, phobias, and preparedness: toward an evolved module of fear and fear learning. *Psychological Review*, *108*(3), 483.

Plaut, D. C. (1995). Double dissociation without modularity: Evidence from connectionist neuropsychology. *Journal of Clinical and Experimental Neuropsychology*, *17*(2), 291-321.

Schimmack, U. (2019). The implicit association test: a method in search of a construct. *Perspectives on Psychological Science*, 1745691619863798.

Stahl, C., & Aust, F. (2018). Evaluative conditioning as memory-based judgment. *Social Psychological Bulletin*, *13*(2), 1-30.

SPSP (2020, March 25)."Preparing and eating [comfort] foods activates a primitive and implicit feeling of being cared for and loved". [Tweet] https://twitter.com/SPSPnews/status/1242837564156936194?s=20

Van Dessel, P., Ye, Y., & De Houwer, J. (2018). Changing Deep-Rooted Implicit Evaluation in the Blink of an Eye: Negative Verbal Information Shifts Automatic Liking of Gandhi. *Social Psychological and Personality Science*, *10*, 266-273.

Recommended readings:

Corneille, O., & Hütter, M. (2020). Implicit? What do you mean? A comprehensive review of the implicitness construct in attitude research. *Personality and Social Psychology Review.*

Corneille, O., & Stahl, C. (2019). Associative attitude learning: A closer look at evidence and how it relates to attitude models. *Personality and Social Psychology Review*, *23*(2), 161-189.

De Houwer, J. (2009). The propositional approach to associative learning as an alternative for association formation models. *Learning & Behavior*, *37*(1), 1-20.

Gawronski, B. (2019). Six lessons for a cogent science of implicit bias and its criticism. *Perspectives on Psychological Science, 14,* 574-595.

Mertens, G., & Engelhard, I. M. (2020). A systematic review and meta-analysis of the evidence for unaware fear conditioning. *Neuroscience & Biobehavioral Reviews*, 108, 254-266.