Unawareness of Attitudes, their Environmental Causes, and their Behavioral Effects

Bertram Gawronski[1] & Olivier Corneille[2]

[1] University of Texas at Austin, USA
[2] UCLouvain, Belgium

Claims about unawareness are abundant in attitude research. The current article provides an analysis of evidence regarding three aspects of an attitude for which people may lack awareness: (1) the attitude itself, (2) its environmental causes, and (3) its behavioral effects. Our analysis reveals that, despite widespread claims of unawareness of the three aspects, strong empirical evidence for these claims is surprisingly scarce. The article concludes with a discussion of the most likely aspects of attitudes that people may be unaware of; their relation to contextual factors that might influence evaluative responses outside of awareness; open questions about the (un)awareness of attitudes, their environmental causes, and their behavioral effects; and methodological recommendations for future research that aims to provide more compelling evidence for aspects of attitudes that may evade awareness.

Keywords: attitudes; awareness; attitude-behavior relations; consciousness; implicit measures; unconscious learning

A common definition specifies *attitude* as "a psychological tendency that is expressed by evaluating a particular entity with some degree of favor or disfavor" (Eagly & Chaiken, 2007, p. 582). A major theme in attitude research pertains to the unawareness of different aspects of attitudes. Is it possible to hold an attitude without being aware of that attitude? Can environmental stimuli influence attitudes outside of awareness? And can attitudes influence behavioral responses in a manner that evades awareness? The current article provides an analysis of evidence relevant to these questions. To this end, we first describe the conceptual framework to organize our analysis and then review evidence pertaining to the three questions above. Counter to the prevalence of claims about unawareness in the attitude literature, our analysis reveals that strong empirical evidence for these claims is surprisingly scarce. We conclude our analysis with a discussion of the most likely aspects of attitudes that people may be unaware of; their relation to contextual factors that might influence evaluative responses outside of awareness; open questions about the (un)awareness of attitudes, their environmental causes, and their behavioral effects; and methodological recommendations for future research that aims to provide more compelling evidence for aspects of attitudes that may evade awareness.

Conceptual Framework

An important aspect of the above-cited definition is the distinction between *attitude* as a latent mental construct and the behavioral expression of latent attitudes in overt evaluative responses (Eagly & Chaiken, 2007).[1] This distinction stipulates that measures of evaluative responses should not be treated as direct indicators of attitudes, because variance in evaluative responses can be due to various other factors, and behavioral expressions of attitudes can be disrupted by non-attitudinal factors (see Calanchini, 2020; De Houwer et al., 2013). Thus, when studying attitudes, it is important to always consider the extent to which observed differences in evaluative responses are driven by genuine differences in underlying attitudes or by other non-attitudinal factors.

Expanding on the definition of *attitude*, it is possible to distinguish between three aspects of an attitude for which people may lack awareness (Gawronski & Bodenhausen, 2012): the attitude itself, its environmental causes, and its behavioral effects (see Figure 1). Although establishing unawareness can be a difficult methodological endeavor, a common approach to investigate unawareness of a psychological entity X is to test whether participants can report X (Nisbett & Wilson, 1977; Timmermans & Cleeremans, 2015). To the extent that there is a discrepancy between X and people's self-report of X, unawareness would provide a potential explanation for the observed discrepancy (see Gawronski & Bodenhausen, 2015). However, inferences of unawareness additionally require that there is no alternative explanation that may account for the observed discrepancy. For example, while some discrepancies between X and people's self-report of X may reflect a genuine inability to report X, other discrepancies may reflect a motivationally driven unwillingness to report X, low correspondence between measures, low measurement reliability, or low sensitivity in capturing the to-be-measured constructs. Although inferences of unawareness would seem justified in cases involving a genuine inability to report X, they would be premature and potentially

[1] Another important aspect is that attitudes can have affective, cognitive, and motivational bases. However, whether a given attitude arises from affective, cognitive, or motivational processes is an empirical question that goes beyond the definition of the attitude construct.

unwarranted in the other cases. Thus, when interpreting discrepancies between X and self-reports of X as evidence for unawareness, it is critical to always rule out alternative explanations for the observed discrepancies.

Expanding on these considerations, unawareness claims for the three aspects can be linked to specific empirical questions. Regarding the presumed unawareness of attitudes, the central question is whether there is evidence for attitudes that people are unable to report. Regarding the presumed unawareness of the environmental causes of attitudes, the central question is whether there is evidence for environmental influences on attitudes when people are unable to report the stimulus event that is responsible for their attitude or the causal impact of a stimulus event on their attitude. Finally, regarding the presumed unawareness of the behavioral effects of attitudes, the central question is whether there is evidence that attitudes can influence behavior when people are unable to report this behavior or the causal impact of their attitudes on their behavior.

Unawareness of Attitudes

A widespread assumption in the attitude literature is that people can have attitudes that they do not know they have—or put differently: people often do not know that they like or dislike something. In addition to being of great theoretical and practical interest, the question of whether people can be unaware of their attitudes constitutes the most fundamental one for the current analysis, because it is logically impossible to have accurate knowledge about the environmental causes or behavioral effects of an attitude if one is unaware of the attitude itself (Gawronski & Bodenhausen, 2012).

To empirically establish unawareness of an attitude, research requires an indicator of actual (dis)liking and a measure of people's beliefs about their (dis)liking. The most common approach is to compare responses on so-called "objective" and "subjective" measures of (dis)liking, with discrepancies between the two being interpreted as evidence for unawareness. A central assumption underlying this approach is that the objective indicator of (dis)liking can be used as a normative criterion for judging the (in)accuracy of participants' subjective reports of their (dis)liking (Kruglanski, 1989). But how do we know that objective indicators capture a person's (dis)liking of an object better than the person's subjective self-report? This question highlights the delicate issue that inferences of unawareness are often based on claims by researchers that they know better what their participants like or dislike than the participants themselves, which fundamentally depends on the validity of the so-called "objective" measure. To the extent that the validity of the "objective" measure seems questionable, inferences of unawareness would be based on a weak foundation, which can lead to inaccurate attitude theories (Gawronski & Bodenhausen, 2015) and potentially harmful consequences at the individual and the societal level (Cameron et al., 2010; Daumeyer et al., 2019).

Another important issue for the interpretation of discrepancies between objective and subjective measures is that the two measures have similarly high reliability and sensitivity (Shanks & St. John, 1994), use the same attitudinal stimuli (Gawronski, 2019), and are not confounded with other factors that are different from awareness, such as the timing of responding to the focal stimuli (Moors, 2016). Otherwise, discrepancies between objective and subjective measures may be driven by any of these factors, which undermines inferences of unawareness.

Physiological vs. Self-Report Measures

One approach based on the distinction between objective and subjective measures is to treat physiological responses to an object as objective indicators of attitudes, and responses on self-report measures as subjective indicators of participants' beliefs about their attitudes (Cunningham et al., 2009; Ito & Cacioppo, 2007). To the extent that responses on the two measures diverge, participants are often assumed to be unaware of the attitudes captured by the physiological measure.

Although this may be the case, there are several issues that undermine straightforward inferences of unawareness from dissociations between physiological and self-report measures. First, many physiological measures capture responses to stimuli that are much faster than responses on traditional self-report measures (e.g., event-related potentials). In these cases, discrepancies between physiological and self-report measures may be driven by differences in the time-window of measured responses rather than lack of awareness (Cunningham et al., 2007; Moors, 2016). Second, many physiological measures suffer from low reliability, which is less common for traditional self-report measures (Krosnick et al., 2005). This difference can lead to systematic dissociations for simple methodological reasons that have nothing to do with lack of awareness (i.e., reliability problem; see Shanks & St. John, 1994). Third, to serve as an "objective" indicator of attitudes, a physiological measure must capture responses along the valence dimension rather than responses along other dimensions (e.g., arousal). Thus, if responses on a physiological and a self-report measure do not align, a potential explanation is that the physiological measure captures responses along a dimension that is different from valence (i.e., sensitivity problem; see Shanks & St. John, 1994). Fourth, although the latter problem can be addressed via thorough validation of the physiological measure, studies on the construct validity of physiological

measures require knowledge about the valence of the utilized stimuli. Yet, this knowledge typically comes from studies using self-report measures (e.g., Lang et al., 2008), which leads to an inferential paradox for research that relies on physiological measures to study unawareness of attitudes. To ensure that a physiological measure captures responses along the valence dimension, responses on the measure must converge to those on self-report measures. Yet, to demonstrate unawareness of attitudes, responses on physiological measures need to diverge from those on self-report measures. Together, these issues create major problems for research that aims to demonstrate unawareness of attitudes via discrepancies between physiological and self-report measures. These problems may at least partly explain why recent research on unawareness of attitudes has moved away from treating physiological measures as objective indicators of attitudes.

Indirect vs. Direct Measures

A much more popular approach to study unawareness of attitudes is to compare self-reported evaluations on direct measures to responses on a particular type of indirect measures, such as the Implicit Association Test (IAT; Greenwald et al., 1998), the Evaluative Priming Task (EPT; Fazio et al., 1995), and the Affect Misattribution Procedure (AMP; Payne et al., 2005). The background assumption underlying this approach is that responses on indirect measures can be treated as objective indicators of attitudes, whereas responses on direct measures are subjective indicators of participants' beliefs about their attitudes. Thus, in line with the argument outlined at the beginning of this section, discrepancies between direct and indirect measures have been claimed to reflect unawareness of attitudes captured by indirect measures, which is reflected in descriptions of the relevant instruments as *implicit measures* and the attitudes captured by these instruments as *implicit attitudes* (Gawronski & Brannon, 2019; for a critique of the implicit terminology, see Corneille & Hütter, 2020).

Inferences of unawareness from discrepancies between direct and indirect measures suffer from the same problems outlined for physiological measures. First, responses on many indirect measures (e.g., IAT, EPT) are much faster than responses on direct measures. Hence, discrepancies between the two kinds of measures may be driven by differences in the time-window of measured responses rather than lack of awareness (Cunningham et al., 2007; Ranganath et al., 2008). Second, many indirect measures suffer from low reliability, the only exception being the IAT and the AMP (Gawronski & De Houwer, 2014; Greenwald & Lai, 2020). Thus, for indirect measures with low reliability, dissociations to direct measures may be driven by differences in their reliability rather than lack of awareness (Cunningham et al., 2001). Third, like physiological measures, the validity of most indirect measures has been established via stimuli of known valence, and this knowledge came from studies using self-report measures (e.g., Fazio et al., 1986; Greenwald et al., 1998; Payne et al., 2005). This issue creates the same inferential paradox described for physiological measures. On the one hand, responses on direct and indirect measures must converge to confirm the construct validity of the indirect measure. On the other hand, responses on the two kinds of measures must diverge to demonstrate unawareness of attitudes. Moreover, when convergence is demonstrated for some stimuli (e.g., flowers and insects) and divergence is found for other stimuli (e.g., Black and White faces), the observed differences across content domains open the door for alternative explanations that do not involve claims of unawareness. In line with this concern, some have argued that discrepancies between direct and indirect measures reflect unwillingness rather than inability to report one's personal attitudes (e.g., Dunton & Fazio, 1997; Nier, 2005; M. A. Olson et al., 2007), although claims that indirect measures are immune to strategic control have been disputed (e.g., Calanchini, 2020; Corneille & Lush, 2023; Gawronski et al., 2007). Fourth, direct and indirect measures tend to differ in terms of various structural features, rendering dissociations between the two kinds of measures conceptually ambiguous (Payne et al., 2008). The significance of this concern is supported by studies showing that at least some dissociations between direct and indirect measures disappear when confounds with structural task characteristics are eliminated (e.g., Béna et al., 2022).

Another concern is that deliberate evaluations on direct measures are influenced by various response-related factors that do not affect spontaneous evaluations on indirect measures to the same extent (Fazio, 2007; Gawronski & Bodenhausen, 2006). Thus, dissociations between the two kinds of measures may be driven by any of these response-related factors rather than lack of awareness. Hahn at al. (2014) aimed to address this ambiguity by asking participants to predict their preferences for different social groups on several IATs before they completed those IATs. Participants showed high accuracy in predicting their IAT scores regardless of their prior experience with the IAT, regardless of how much information they received about the IAT, and regardless of whether the IAT was introduced as a measure of true beliefs or cultural associations (see Gawronski et al., 2008; M. A. Olson et al., 2009). Moreover, participants showed high accuracy in predicting their IAT scores although self-reported evaluations on direct measures showed the same small correlations with IAT scores found in prior research (for meta-analyses, see Cameron et al., 2012; Hofmann et al., 2005). Together, these results pose a

challenge to the idea that indirect measures such as the IAT capture attitudes that people do not know they have.

A noteworthy aspect of Hahn et al.'s (2014) findings is that participants showed high accuracy in predicting their personal patterns of IAT scores at a within-subjects level, but they were much less accurate in predicting their scores on a particular IAT at a between-subjects level. Put differently, although participants were highly accurate in predicting their personal rank order of preferences in the completed IATs (e.g., that their preference for White over Black people was stronger than their preference for White over Hispanic people), they were less accurate in predicting how their preference on a given IAT compares to that of the other participants in the sample (e.g., that their preference for White over Black people is stronger compared to the majority of the other participants in the sample). This difference is important, because it illustrates an inherent problem of between-subjects approaches in research on unawareness of attitudes. To obtain high convergence between an objective and a subjective measure at a between-subjects level, participants not only need to know their attitude toward the focal object (e.g., how much they like apples); they also need to know how their attitude compares to the attitudes of the other participants in the sample (e.g., how their liking of apples compares to the liking of apples among the other participants). Thus, low convergence in between-subjects designs may not necessarily reflect unawareness of the attitude; it may also reflect limited knowledge about the attitudes of the other participants (see Goffin & Olson, 2011; J. M. Olson et al., 2007). This situation is different in within-subjects designs that focus on attitudes toward multiple objects among individual participants (e.g., Hahn et al., 2014; Hahn & Gawronski, 2019). To obtain high convergence between objective and subjective measures for multiple objects at a within-subjects level, participants must know how their attitude toward one object compares to their attitude toward other objects (e.g., how much they like apples compared to oranges, bananas, mangos, etc.), but they do not have to know anything about the other participants in the sample. These issues have to be considered when interpreting findings of studies that used between-subjects designs to establish unawareness of attitudes (see Hahn & Goedderz, 2020).

The high level of accuracy in the prediction of IAT scores appears to conflict with evidence that participants tend to be rather surprised when they receive feedback about their performance (e.g., when they learn that they have a strong preference for White over Black people; see Goedderz & Hahn, 2022). Anecdotes of such surprise reactions have been interpreted as evidence that people are unaware of the attitudes captured by the IAT (e.g., Banaji, 2011; Krickel, 2018), which seems difficult to reconcile with the conclusion that high accuracy in the prediction of IAT scores demonstrates awareness. To resolve this apparent contradiction, it is worth noting that surprise reactions in response to IAT feedback merely reflect a discrepancy between participants' verbal quantification of their subjective preference (e.g., *moderate preference for White over Black people*) and the experimenters' verbal quantification of the obtained measurement score (e.g., *strong preference for White over Black people*). Hence, participants may be surprised about their IAT feedback, not because they are unaware of their attitudes, but because the metric underlying their verbal quantification does not match the metric underlying the verbal quantification in the feedback they receive (Gawronski, 2019). Consistent with this argument, Hahn et al. (2014) found that, although participants were highly accurate in predicting their IAT scores, the metric underlying their verbal quantifications "stretched" the metric commonly used to convert numeric IAT scores into verbal feedback. Because labeling conventions for what should be considered a "weak," "moderate," or "strong" bias are entirely arbitrary in the sense that there is no objective basis to treat one metric as "correct" and another one as "incorrect" (Kruglanski, 1989), interpretations of surprise reactions as evidence for unawareness are based on a questionable normative premise (Gawronski et al., 2022a). These concerns are further supported by evidence that the standard algorithm to calculate IAT scores dramatically inflates their size (Wolsiefer et al., 2017), suggesting that the mismatch in verbal quantifications underlying surprise responses is rooted in a systematic distortion of IAT feedback, not unawareness of attitudes.

Revealed vs. Stated Preferences

An alternative to using indirect measures as "objective" indicators of attitudes is to compare evaluative responses on self-report measures to other behavioral expressions of attitudes that do not involve self-report (e.g., consumption behavior, purchasing decisions). This approach is captured by the distinction between stated and revealed preferences (De Corte et al., 2021). For example, stated and revealed preferences for popcorn would be discrepant if self-reported liking of popcorn is unrelated to actual popcorn consumption (e.g., Neal et al., 2011). Although the distinction is more common in research on consumer behavior than research on attitudes, discrepancies between stated and revealed preferences have been interpreted as evidence that people sometimes do not know what they like or dislike (for examples in research on romantic attraction, see Eastwick & Finkel, 2008). Although this may the case, such inferences must be treated with caution in the absence of further evidence. A widely accepted notion in research on attitude-behavior relations is that

attitudes do not influence behavior in a direct, unconditional manner and that behavior is influenced by various other factors beyond attitudes (Ajzen & Kruglanski, 2019; Fazio 1990). Although it is possible that discrepancies between stated and revealed preferences reflect people's unawareness of their attitudes, the known complexity of attitude-behavior relations renders such inferences premature without additional data that rule out other factors as the cause of the observed discrepancies (e.g., social norms, perceived behavioral control). Thus, inferences of unawareness from discrepancies between stated and revealed preferences have to be evaluated in the context of extant theories of attitude-behavior relations and the proposed factors that moderate attitude-behavior relations. To our knowledge, there is no empirical work that has systematically addressed these issues.

Discrepant Self-Reports

The preceding sections illustrate the problems of treating physiological measures, indirect measures such as the IAT, and measures of revealed preferences as objective indicators of attitudes for inferences of unawareness. An alternative to comparing responses across measures that do versus do not involve self-reports is to compare responses across two self-report measures. One example involves potential discrepancies between self-reported evaluations of types and tokens (see Ledgerwood et al., 2018, 2020). Whereas types are classes of objects, tokens are individual instances of a class of objects. Operationally, the distinction is captured by the difference between measures involving evaluations of an abstract category and measures involving evaluations of individual exemplars of that category.

Discrepancies in self-reported evaluations of types versus tokens may arise for several reasons. First, types and tokens may be considered conceptually distinct attitude objects, in that an abstract category is not the same as the aggregate of multiple individual exemplars of that category. From this perspective, discrepant evaluations of a category and individual exemplars of that category would reflect a simple lack of measurement correspondence. Second, even when types and tokens are deemed conceptually equivalent (e.g., Ajzen & Fishbein, 1977), discrepancies may arise when a given factor differentially affects responses toward types versus tokens. For example, self-reported evaluations of Black people as a social category may differ from self-reported evaluations of individual Black exemplars because people may be more concerned about expressing negative evaluations of Black people as a social category than about expressing negative evaluations of individual Black exemplars (Dovidio & Gaertner, 2004).

Finally, and most relevant for the current analysis, discrepancies between self-reported evaluations of types versus tokens may reflect lack of awareness. For example, self-reported evaluations of an abstract category (e.g., self-reported liking of Pinot Noir as a type of red wine) may differ from self-reported evaluations of exemplars of that category (e.g., self-reported liking of specific tokens of Pinot Noirs in a blind tasting) because people have genuinely inaccurate beliefs about what they like and dislike (e.g., a person may think they do not like Pinot Noirs, but they actually do). In this case, self-reported evaluations of an abstract category would reflect a person's beliefs about their (dis)liking of exemplars of that category, whereas self-reported evaluations of individual exemplars of the focal category would reflect the person's actual (dis)liking of exemplars of that category (Ledgerwood et al., 2020). Unawareness of this kind may occur when evaluations of abstract categories are based on inductive inferences from concrete experiences with individual exemplars (Alcser-Isais et al., 2022; Da Silva Frost et al., 2023; Woiczyk & Le Mens, 2021) and these inferences are distorted by sampling error or biases in inductive reasoning (see Fiedler & Plessner, 2009). In such cases, people may draw conclusions about their liking of a category that does not accurately reflect their liking of individual exemplars of that category. Although empirical work along this line is still scarce, the conceptual idea underlying these arguments raises interesting questions about how people draw inferences about their (dis)liking of an abstract category from their (dis)liking of individual exemplars of that category, the conditions under which these inferences can produce beliefs about the (dis)liking of the category that do not align with one's actual (dis)liking of exemplars of that category, and what such discrepancies can tell us about people's (un)awareness of their own attitudes (see Ledgerwood et al., 2018).

Interim Conclusions

Several conceptual issues undermine inferences of unawareness from discrepancies between responses on physiological measures and responses on self-report measures. Moreover, counter to a dominant narrative in the literature, there is no evidence supporting the idea that indirect measures such as the IAT, the EPT, and the AMP would capture attitudes that people are unaware of. If anything, the available evidence suggests the opposite. Although it is possible that discrepancies between stated and revealed preferences are driven by unawareness of one's attitudes, such interpretations must be evaluated in the context of extant theories about attitude-behavior relations. An alternative to comparing responses across measures that do versus do not involve self-reports is to infer unawareness from discrepancies between two self-reports, one example being discrepancies between self-reported evaluations of a type and tokens of that type. However, such discrepancies can also be driven by alternative factors

and compelling evidence for unawareness as a driving force is still lacking.

Unawareness of Environmental Causes

Even when people are aware that they (dis)like an object, they may not be aware why they (dis)like it. There are several obvious ways in which people may be unaware of the causes of their attitudes. First, there is no uncaused cause, and distant ones often escape understanding. Second, causal influences occur at various levels (e.g., synaptic activity involved in attitude learning; cultural determinants of systems of preferences), and no individual can possibly hold comprehensive knowledge about all of them. Third, relatedly, people may be unable to introspect on the various mechanisms (including the various psychological mechanisms) driving their phenomenological experiences and behavior. In this very broad sense, people remain necessarily unaware of the complex set of causes and mechanisms influencing their attitudes.

In this section, we discuss how psychological research has informed two more specific questions about environmental causes of attitudes: (1) Can an attitude be created when people are unable to report the stimulus event that is responsible for it? (2) Provided people are aware of the stimulus event, are there cases where people are nevertheless unable to report that the stimulus event influenced their attitude? Answering these questions requires tight control over the stimulus event that is responsible for the attitude, reliable measures of evaluation, adequate measures of awareness, and sensitive analytic procedures, all of which raise significant methodological challenges (see Newell & Shanks, 2014; Shanks et al., 2021).

Our analysis in this section focuses on evaluative conditioning (EC) and mere exposure (ME), which are frequently considered strong cases for attitude learning without awareness. To address our first question regarding unawareness of the stimulus event, we organize our discussion around studies using procedures that (1) weakened the strength of stimuli, (2) weakened top-down attention to stimuli, and (3) linked evaluative responses to measures of recollective memory. Expanding on this analysis, we discuss our second question regarding unawareness of the influence of stimulus events. Although space constraints do not permit elaborate discussions of measurement and analytic issues, we regularly touch on them and provide references to more comprehensive treatments of these issues.

Unawareness of Stimulus Event

Evaluative Conditioning

EC procedures involve pairing a neutral stimulus (the conditioned stimulus, or CS) with a stimulus of positive or negative valence (the unconditioned stimulus, or US). Following this CS-US pairing, the CS is typically evaluated in line with the valence of the US, a phenomenon known as *EC effect* (for a review, see Moran et al., 2023). For example, pairing the logo of an unfamiliar brand (i.e., neutral CS) with the picture of a beautiful scenery (i.e., positive US) may result in more positive evaluations of the CS. Conversely, pairing the logo with a picture showing dental decay (i.e., negative US) may elicit more negative evaluations of the CS. EC effects have been found for various types of CSs (e.g., human faces, consumer products, kanji symbols, abstract visual patterns, or meaningless letter strings). They are frequently claimed to be driven by low-level processes operating without awareness of the CS-US pairings (e.g., Gawronski & Bodenhausen, 2006; Petty et al., 2019). As we discuss here, however, evidence has accumulated that is at odds with this conclusion.

Weak stimulus strength. Early studies claimed successful EC effects when using briefly presented stimuli, often called "subliminal" stimuli (e.g., Krosnick et al., 1992). These procedures are commonly assumed to prevent participants from accurately reporting the stimulus event. However, early demonstrations with "subliminal" presentations have been criticized for using flawed designs, inadequate awareness checks, or no awareness checks at all (for a review, see Sweldens et al., 2014). Recent studies relying on stronger designs and adequate measures of awareness have generally failed to support "subliminal" EC. The most complete and controlled investigation of short exposure effects was realized by Stahl et al. (2016) in a series of 6 experiments involving 27 experimental conditions. These authors found EC effects only for CS exposures associated with high CS identification performance and high attention. Overall, there is little to no evidence for EC effects when using low-strength stimuli (for a detailed review, see Corneille & Stahl, 2019).

Weak top-down attention. Although studies with "subliminal" stimuli have long been considered the strongest case for unconscious influences, such studies have been criticized for their lack of ecological validity. As Bargh (2022) pointed out, "subliminal stimuli are a creation of 20th century technology, [and] the human mind could not possibly have evolved to process them" (p. 90). One solution to this problem is to use high strength stimuli combined with low attention. Several EC studies did so by asking participants to perform a concurrent attention-demanding task while processing CS-US pairs displayed on a computer screen (e.g., Mierop et al., 2017; Pleyers et al., 2009). EC effects under attentional-load conditions are compared to EC effects in a control condition in which participants do not perform an attention-demanding task. Overall, studies using this approach consistently found that EC effects vanish to non-significance under attentional

load (for a detailed review, see Corneille & Stahl, 2019).

A limitation of studies using manipulations of attentional load is that they confound awareness of CS-US pairings with processing goals (e.g., processing numeric values vs. listening to music). To address this issue, Dedonder et al. (2014) compared effects of foveal and parafoveal presentations. These authors presented the USs in participants' foveal eye region and paired them with either foveal or parafoveal CSs. An EC effect was found only for foveal but not parafoveal CSs, the latter of which were less likely to enter awareness.

Expanding on concerns that brief-exposure studies confound awareness with stimulus duration and that attentional-load studies confound awareness with processing goals, Hödgen et al. (2018) pointed out that studies comparing effects of foveal and parafoveal presentations confound awareness with spatial proximity. In the latter type of studies, CS-US pairs are presented closer to each other when both stimuli are presented in the foveal region than when one of the two stimuli is presented parafoveally. To address this issue, Hödgen et al. relied on continuous flash suppression to present the CSs outside of awareness. Here, participants were presented with different visual information in their left and right eye. One eye received "high-energy" US information (i.e., the continuous flashing of a sequence of US photos and colored pixel masks) while the other eye received "low-energy" CS information (i.e., a stationary grey shape of low visual contrast). In these procedures, the high-energy information is dominant and impairs awareness of the low-energy information. A series of four experiments consistently failed to obtain EC effects for suppressed CSs on both direct and indirect measures.

Olson and Fazio (2001) relied on yet another rationale: incidental learning. In a simulated surveillance task, two Pokémon characters (CSs) were incidentally paired with either positive or negative stimuli (USs) on distractor trials that were irrelevant for participants' goal in the task (i.e., press the space bar whenever they see a particular image). This way, participants' attention was pulled away from the intentional processing of the CS-US pairs. Although the procedure has been found to be effective in producing significant EC effects, a high-powered replication of Olson and Fazio's (2001) original study found that EC effects in the surveillance task are driven by a subset of consciously encoded CS-US pairings (Kurdi et al., 2022; Moran et al., 2021). Again, earlier conclusions of unawareness could not be supported.

Lack of conscious recollection. Several studies relied on high-strength stimuli and high attention conditions and tested whether EC effects are observed in the absence of conscious recollection of the CS-US pairings. Although measures of recollective memory are ambiguous about the role of (un)awareness during exposure to CS-US pairings (Gawronski & Walther, 2012), research using such measures are still relevant for the current question of whether people can (dis)like something without being able to report the stimulus event that is responsible for their (dis)liking. Depending on the procedure and analytic approach, some studies found evidence for memory-independent EC (e.g., Hütter & Sweldens, 2013; Jurchiș et al., 2020; Walther & Nagengast, 2006; Waroquier et al., 2020) while other studies did not (e.g., Kurdi et al., 2022; Mierop et al., 2017; Pleyers et al., 2007; Stahl et al., 2009).

To reconcile the mixed findings and to address methodological limitations of prior studies, Stahl and colleagues (2023) developed a new procedure for examining EC effects in the absence of feelings of remembering the US valence. Following exposure to CS-US pairs, participants were asked to use two buttons sets, labeled *SET 1* and *SET 2*. If they felt they could remember the valence of the US paired with a given CS, they were asked to use the buttons from SET 1 and to press *pleasant* (vs. *unpleasant*) for reporting their recollection of a positive (vs. negative) US. If, however, they felt they could not remember the US valence, they were asked to use the buttons from SET 2 and to press *pleasant* (vs. *unpleasant*) to report liking (vs. disliking) the CS. Compared to the approaches used in prior work, two advantages of this procedure are that it (1) relates evaluations to subjective memory states at the within-person-within-item level (information criterion) and (2) measures evaluative and memory judgments closely in time (immediacy criterion). When validating and using this new procedure, the authors found no evidence for EC effects in the absence of feelings of remembering.

Mere Exposure

In ME studies, neutral stimuli typically void of meaning (e.g., unfamiliar shapes of polygons) are evaluated more positively when they have been presented before than when they have not been presented before, a phenomenon known as *ME effect* (Zajonc, 1968). This effect is often considered another compelling case for attitude learning without awareness.

Weak stimulus strength. A recent meta-analysis found a significant linear (and quadratic) ME effect at durations of < 15 ms exposure (Montoya et al., 2017). This result suggests that ME effects can be established with low-strength stimuli. However, it is unclear whether these effects were established without awareness of the stimulus event. A minority of the relevant effects came from unpublished raw data files for which procedural information is lacking. All remaining effects came from two articles that relied on a unique "subliminal" procedure by Förster (2009). This procedure reportedly presented sandwich-masked stimuli for 10 ms or 14 ms in the center of computer

screens. Yet, based on the provided information, it remains unclear whether the software and computer monitors in these studies guaranteed such fast and precise exposure durations. Most critically, as is often the case in "subliminal" research, awareness measures were either lacking in these studies or did not meet reliability, immediacy, and sensitivity criteria (see Newell & Shanks, 2014). Besides widespread procedural concerns of this sort, ME effects at short exposures also seem to depend on moderators that are yet to be better understood. For example, Kawakami and Yoshida (2019) did not find a significant ME effect at 10 ms durations on a direct measure, but found one on indirect measures (GNAT, IAT). Newell and Shanks (2007) compared ME effects for short (40 ms) versus long (400 ms) exposure durations and found a significant effect on a direct measure only for the longer exposure duration and when recognition performance was best. Notably, when observed, dissociations between evaluative judgments and memory judgments in "subliminal" ME studies may also arise from different decision strategies for the two kinds of judgments. When these decision strategies are swapped, the dissociation has been found to be reversed, with above-chance recognition memory and at-chance liking (Whittlesea & Price, 2001). Such findings further complicate inferences of unawareness.

Weak top-down attention. In our discussion of EC effects, we noted that the continuous-flash-suppression procedure resolves several confounds in the study of awareness. We also emphasized the importance of including adequate awareness measures instead of merely assuming that the procedure precludes awareness. A ME study by de Zilva et al. (2013) addressed both issues. Combining continuous flash suppression with online identification measures, these authors unexpectedly found that 36% (Experiment 1) and 64% (Experiment 2) of their samples were aware of the supposedly suppressed stimuli. This finding is remarkable, because "none of these participants would have been excluded on the basis of a traditional post-exposure recognition test" (de Zilva et al., 2013, p. 6). Moreover, a significant ME effect emerged only for unsuppressed stimuli and for "suppressed" stimuli that had entered awareness. However, if we relax the stringency of the unawareness test, some studies lend support for ME effects under conditions of weakened attention, such as when attending to high-strength stimuli presented as distractors (e.g., Hansen & Wänke, 2009) or when attending to visually suppressed high-strength stimuli (Huang & Hsieh, 2013). In sum, there is some evidence for a ME effect under conditions of weakened attention, but it is not clear if those conditions prevented awareness of the stimulus event at the time of exposure.

Lack of recollection. Hansen and Wänke (2009) used a process-dissociation procedure to quantify the respective contributions of familiarity and conscious recollection to the ME effect. To do so, they compared memory judgments for previously presented names of unknown products in two experimental conditions where familiarity and conscious recollection of these names can be assumed to have converging versus diverging effects on memory performance (see Jacoby, 1991). They found that repeated high-strength exposure to the product names increased participants' liking of these names, and that this ME effect was associated with feelings of familiarity but not with conscious recollection. A noteworthy feature of Hansen and Wänke's study is that the authors experimentally validated the functional independence of the recollection and familiarity estimates provided by the process-dissociation procedure. While the conscious recollection estimate (but not the familiarity estimate) was influenced by a manipulation of attention at encoding, the familiarity estimate (but not the conscious recollection estimate) was influenced by a manipulation of figure-ground contrast. Together, these findings suggest that repetition-induced feelings of familiarity can influence one's liking of a stimulus in the absence of conscious recollection of being previously exposed to this stimulus.

Unawareness of Causal Influence

The study by Hansen and Wänke (2009) represents an interesting case for introducing our second question regarding unawareness of the influence of stimulus events. In that study, prior exposure to stimuli increased liking of the stimuli without conscious recollection of their prior presentation and, by implication, of the influence of the stimulus event. Because people are constantly exposed to a large number of events, it is unlikely that they can consciously recollect them all, and consciously weigh how much these events collectively influenced their evaluation. At this point, however, it is important to specify what the causally effective event is. For example, participants may not recollect their prior exposure to a stimulus, but they may be perfectly aware that its processing is fluent. In turn, they may rely on this meta-cognitive cue to draw inferences about their liking of the stimulus (Greifeneder & Schwarz, 2014). A good illustration is provided by the ease-of-retrieval effect, whereby mental contents and the subjective ease of their retrieval can have opposite effects on evaluative judgments (Schwarz et al., 1991). For example, asking participants to recall five arguments (difficult experience) rather than two arguments (easy experience) in favor of a surgery fee can result in unfavorable evaluations of this fee when people infer an unfavorable evaluation from the difficulty of generating favorable arguments (Greifeneder & Bless, 2007). For such effects to occur,

participants should not question the diagnostic value of their feelings of difficulty. Hence, they should not attribute it to the experimental manipulation. However, they need to be aware of the feeling of difficulty itself. As a case in point, ease-of-retrieval effects are typically found when participants are asked to report their subjective feelings before rather than after the judgment at hand (Kühnen, 2010). Furthermore, participants also need a naïve meta-cognitive theory for relating their subjective experiences to this judgment (Schwarz, 2004). The question now becomes whether people are aware of drawing causal inferences from their meta-cognitive feelings when forming evaluations, which may qualify as a meta-meta-cognitive question.

This begs the question of people's knowledge about stimulus-response relations. Over the past seven decades, instruction-based replication studies have shown that participants can accurately predict, simulate, or produce many attitudinal phenomena based on procedural information delivered in the original studies (for a discussion, see Corneille & Béna, 2023). A famous case is a study by Bem (1967), in which observers were informed about the procedures used in the classic cognitive dissonance study by Festinger and Carlsmith (1959). Bem (1967) found that, based on this procedural information alone, observers could accurately estimate how the participants had completed the evaluative ratings in the original study. Because the observers in Bem's study presumably did not experience a state of cognitive dissonance, this instruction-based replication study questioned the role of arousal in cognitive dissonance effects. More recently, instruction-based replications have been reported on direct and indirect evaluative measures for EC (e.g., De Houwer, 2006), ME (e.g., Van Dessel et al., 2017), approach-avoidance (i.e., liking stimuli better when they were approached rather than avoided; e.g., Van Dessel et al., 2015), and vicarious evaluative learning effects (i.e., liking stimuli better when they were seen to elicit a positive rather than a negative reaction in another person; e.g., Kasran et al., 2022).

These results suggest that participants hold causal knowledge relating stimulus events (e.g., CS-US pairings, stimulus repetitions) to evaluative responses. A question worth examining in future research is whether participants can verbally report this causal knowledge (i.e., if they are aware of it) and, if so, whether they are using it (and are aware of using it) to inform their evaluations. Finally, it would be important to know if participants use it (and are aware of using it) because of compliance with experimental demands. Indeed, whenever participants are aware of how experimental procedures relate to responses, experimental demand effects cannot be easily ruled out (for discussions, see Corneille & Béna, 2023; Corneille & Lush, 2022).

Given space limitations, we limited our discussion to EC and ME. We chose to do so because these procedures are typically considered low-level attitude learning procedures (e.g., Petty et al., 2019). Other paradigms may offer different conclusions, but they face similar challenges. To illustrate, consider the spreading-of-alternatives effect in studies using the free-choice paradigm (Brehm, 1956). Here, participants typically evaluate a chosen option more favorably than a rejected option after making a choice, even when they evaluated the two options similarly before making a choice. It seems reasonable to assume that participants in these studies are aware of the stimulus event (i.e., the options they selected and rejected). Yet, when asked to report the reason for their post-choice evaluations, participants may fail to mention the influence of their choice. However, in contrast to interpretations of the latter finding as indicating lack of awareness, several studies suggest that a spreading-of-alternatives effect can be observed even when participants do not make a choice (e.g., Chen & Risen, 2010; Gawronski et al., 2007). These findings suggest that, when participants "fail" to report the influence of their choice on their evaluations, it may be because the choice itself had no causal impact. This conclusion also explains why a spreading-of-alternatives effect can be found among participants with amnesia who cannot remember their choice (Lieberman et al., 2011). Similar concerns apply to inferences of unawareness in research on decision-making more broadly, which have been discussed extensively by Newell and Shanks (2014; see also Shanks et al., 2021).

Interim Conclusions

Regarding our first question about unawareness of stimulus events, the case for "unaware EC" is generally unsupported for low-strength stimuli, and for high-strength stimuli combined with weak attention. More research is needed to reconcile the mixed evidence for effects of high-strength stimuli combined with high-attention but no conscious recollection. The case for "unaware ME" is weaker than frequently stated for low-strength stimuli, moderate for high-strength stimuli combined with weak attention, and comparatively strong for high-strength stimuli combined with high-attention but no conscious recollection. Regarding our second question about unawareness of the influence of stimulus events, ME research suggests that people can be unaware that a stimulus event influenced their attitudes. This typically applies to situations where the stimulus event(s) cannot be recollected at the evaluation stage. However, when this is the case, a more proximal "event" (e.g., a repetition-driven feeling of familiarity) may be consciously used as a meta-cognitive cue informing participants' evaluations. Finally, studies on instruction-based EC and instruction-based ME suggest that people have causal knowledge relating stimulus

pairings and stimulus repetitions to evaluative responses, but it remains unclear whether this knowledge can be reported. More generally, although people may sometimes report reasons for their attitudes that are discrepant with those posited by the experimenter, knowing who is mistaken in these cases is often much less straightforward than experimenters would like to believe (Corneille & Lush, 2022; Cotton, 1980; Kruglanski, 1989; Newell & Shanks, 2023).

Unawareness of Behavioral Effects

Even when people are aware of an attitude and the environmental causes of that attitude, they may not be aware of its behavioral effects. The general notion underlying this idea is that attitudes may sometimes influence behavior in a manner that evades awareness. A frequently cited example of such effects are biases in social behavior that arise from unrecognized influences of intergroup attitudes (Fazio et al., in press). In general, behavioral effects of attitudes can be deemed as being outside of awareness either (1) when people are unaware that they are engaging in the relevant behavior or (2) when people are aware that they are engaging in the relevant behavior, but they are unaware that the behavior is influenced by their attitudes.

Unawareness of Behavioral Response

The first case involves instances where people are unaware that they are engaging in the behavior that is being influenced by their attitudes. Logically, a person cannot be aware of the impact of their attitudes on a given behavior if the person is unaware that they are engaging in that behavior. Unawareness of this type is likely limited to low-level, unintentional reactions and less common for high-level, intentional actions. For example, people may often be unaware of their nonverbal expressions (e.g., Dovidio et al., 2002) and visual attention (e.g., Roskos-Ewoldsen & Fazio, 1992) in social interactions, but they are generally aware of what they are doing when they hire a job candidate or call the police on a suspicious person (Gawronski et al., 2022b). Although there is evidence for attitudinal influences on both low-level, unintentional reactions and high-level, intentional actions (for a review, see Fazio, 1990), a major problem for inferences of unawareness in studies on low-level, unintentional reactions is the lack of awareness checks in these studies. We are not aware of any empirical work in this area that included measures to probe participants' awareness of the relevant behavior. Moreover, if one were to include awareness checks in such studies, asking participants about a specific behavioral reaction can increase awareness of this reaction even when it remains unrecognized in the absence of awareness checks (Kouider & Dehaene, 2007; see also Fox et al., 2011). These issues create a methodological problem for studies that aim to provide empirical evidence for the idea that attitudes can influence low-level, unintentional behaviors that people do not know they engage in. Thus, despite the intuitive plausibility of this idea, there is currently no direct empirical evidence for it. Another problem in this line of work pertains to the underlying hypothesis that an observed behavioral response is driven by attitudes rather than other non-evaluative representations (e.g., semantic beliefs or stereotypes). We will discuss this issue in more detail after the following section on causal influences of attitudes on behaviors that people know they are engaging in.

Unawareness of Causal Influence

A second case involves instances where people are aware that they are engaging in a specific behavior, but they are unaware that the behavior is influenced by their attitudes. The idea underlying the hypothesis of unawareness in this case is that people are often fully aware of what they are doing, but they may nevertheless be unaware of the causal influence of their attitudes on their actions. An example of such influences is the impact of social category cues on action decisions (Gawronski et al., 2022a). For example, in research on gender bias in hiring decisions, participants are generally aware that they are making a hiring decision, but they may not be aware that their hiring decision is influenced by their gender attitudes. Similarly, in the real-world cases described under the hashtag *#LivingWhileBlack* (see Griggs, 2018), people were presumably aware of what they were doing when they called the police on a Black person, but they may not have been aware that their decision to call the police was influenced by their racial attitudes.

Extant research suggests two potential mechanisms by which attitudes may influence high-level, intentional actions outside of awareness. First, attitudes may influence high-level, intentional actions by influencing the weighting of available information (Gawronski et al., 2022a). For example, in a hiring scenario involving a highly qualified man with superior credentials in terms of a Criterion A and highly qualified woman with superior credentials in terms of another Criterion B, gender attitudes may lead a decision-maker to give more weight to Criterion A than Criterion B, leading them to hire the man and not the woman. Yet, in a scenario where the credentials of the two candidates are reversed, gender attitudes may lead the decision-maker to give more weight to Criterion B than Criterion A, leading to the same hiring decision regardless of who is superior in terms of the two criteria (e.g., Norton et al., 2004; Uhlmann & Cohen, 2005). Thus, to the extent that attitudes can influence the weighting of information outside of awareness, relevant evidence would support the idea that people can be aware that they are engaging in a high-level, intentional action

without being aware that the action is influenced by their attitudes.

Second, attitudes may influence high-level, intentional actions by influencing the interpretation of ambiguous information (Gawronski et al., 2022a). For example, if a White and a Black target person show the same ambiguous behavior, racial attitudes may lead perceivers to interpret the ambiguous behavior as threatening when the target is Black, but as harmless when the target is White (e.g., Duncan, 1976; Hugenberg & Bodenhausen, 2003; Kunda & Sherman-Williams, 1993; Sagar & Schofield, 1980). Moreover, based on their differential interpretations of the ambiguous behavior, perceivers may call the police on the Black target, but not the White target. Thus, to the extent that attitudes can influence the interpretation of ambiguous information outside of awareness, relevant evidence would support the idea that people can be aware that they are engaging in a high-level, intentional action without being aware that the action is influenced by their attitudes.

Although there is considerable evidence for biased weighting of mixed information and biased interpretation of ambiguous information (Gawronski et al., 2022a), there is very limited evidence that attitudes can influence actions via the two mechanisms outside of awareness. One potential reason for this lack of evidence might be the difficulty of demonstrating unawareness in these cases. Similar to the difficulties of demonstrating unawareness of a low-level, unintentional behavior, probing awareness of attitudinal influences on high-level, intentional actions can increase awareness of these influences even when they remain unrecognized in the absence of awareness checks (Kouider & Dehaene, 2007). Moreover, in areas involving socially sensitive attitudes (e.g., gender attitudes, racial attitudes), potential discrepancies between actual and acknowledged influences may reflect unwillingness rather than inability to report attitudinal influences. For example, in cases where gender attitudes influence hiring decisions via biased weighting of mixed information, a person may be unwilling to admit that they deliberately weighted the candidates' credentials in manner that rationalizes their preference for a man over a woman. Similarly, in cases where racial attitudes influence decisions to call the police via biased interpretation of ambiguous information, a person may be unwilling to admit that they deliberately relied on the target's race to disambiguate the target's behavior. In both cases, it would be unwarranted to infer unawareness from discrepancies between actual and acknowledged influences of attitudes.

An alternative approach that avoids these issues is to investigate people's control over attitudinal influences under conditions of high motivation and high ability to control (Gawronski et al., 2022b). To the extent that an attitudinal effect on behavior remains uncontrolled under such conditions and statistical power for the detection of a significant moderation is sufficiently large, a plausible interpretation of the obtained null effect is that the attitudinal effect remained uncontrolled because participants were unaware of it (see Strack & Hannover, 1996; Wegener & Petty, 1997; Wilson & Brekke, 1994). To our knowledge, there is only one study that has utilized this approach to probe for unawareness of attitudinal effects. In a study by Gawronski et al. (2003), German participants were presented with ambiguous descriptions of either a German-looking or a Turkish-looking young man and asked to rate the target's behavior along multiple evaluative dimensions. In addition to the impression formation task, the study included measures of ethnic attitudes toward Germans and Turks and a measure of motivation to control prejudiced reactions. The results showed that participants rated the target's behavior more negatively for the Turkish-looking than the German-looking target, and the size of this difference increased as a function of participants' attitudinal preference for Germans over Turks. Interestingly, this pattern was not moderated by motivation to control prejudiced reactions, suggesting that ethnic attitudes influenced the interpretation of ambiguous information even for participants who were highly motivated to control prejudiced reactions. Because the relevant behavior (i.e., responses on a rating scale) was relatively easy to control, these results are consistent with the idea that attitudes biased participants' interpretations of ambiguous behavior outside of awareness. However, limitations of the study design leave the obtained findings open to alternative interpretations, one being that biased interpretations influenced ethnic attitudes rather than vice versa (because ethnic attitudes were measured after the task to measure biased interpretations). Thus, although Gawronski et al.'s (2003) findings are consistent with the hypothesis that attitudes can influence behavior outside of awareness, compelling evidence for this hypothesis is still lacking.

Another obstacle in research on this question pertains to the critical background assumption that an observed behavioral response is driven by attitudes rather than other non-evaluative representations (e.g., semantic beliefs or stereotypes). Because this issue applies to both unawareness of behavioral responses and unawareness of causal influences, we discuss it in more detail in our interim conclusions for this section.

Interim Conclusions

Although it seems intuitively plausible that attitudes can influence behavior outside of awareness, stringent tests of this idea require thorough awareness checks. Yet, probing for awareness of attitudinal influences can

raise awareness of the focal influence, which can make it difficult to provide evidence for unawareness (Kouider & Dehaene, 2007). Moreover, even when such evidence can be provided, it is critical to also establish the hypothesized effect of attitudes on the focal behavior and to rule out alternative interpretations in terms of non-evaluative representations that tend to be confounded with attitudes (e.g., semantic beliefs or stereotypes). For example, when studying effects of racial attitudes, it seems important to not only provide positive evidence for the proposed role of racial attitudes, but to also rule out alternative interpretations in terms of racial stereotypes that tend to be confounded with racial attitudes (see Amodio & Devine, 2006; Phills et al., 2020). In correlational studies, such confounds can lead to spurious associations between racial attitudes and a focal behavior even when the focal behavior is causally influenced by racial stereotypes and not by racial attitudes.

The significance of this issue can be illustrated with the presumed role of gender attitudes in hiring decisions. Research on gender attitudes typically shows a preference for women over men (Eagly & Mladinic, 1989), which conflicts with the commonly observed preference for men over women in hiring decisions (e.g., Moss-Racusin et al., 2012). Considering the mismatching patterns of preferences in attitudes and hiring decisions, a more plausible interpretation of the latter is that the observed bias in favor of men over women is driven by other non-attitudinal factors. In line with this concern, some research suggests that gender bias in hiring decisions arises from the (mis)fit between stereotypical beliefs about men and women and semantic beliefs about different occupations (Glick et al., 1988). In contrast, the commonly observed attitudinal preference for women over men seems to have little impact on hiring decisions (Eagly & Mladinic, 1994). Although attitudes can sometimes be rooted in semantic beliefs (like prejudice can be rooted in stereotypes), the two are conceptually and empirically distinct (like prejudice and stereotypes are conceptually and empirically distinct). Thus, when studying whether attitudes can influence behavior outside of awareness, it seems important to rule out alternative interpretations in terms of non-attitudinal representations that tend to be confounded with attitudes. One possibility to do this is to test effects across multiple target stimuli that are similar in valence but different in terms of other non-evaluative properties (e.g., Roskos-Ewoldsen & Fazio, 1992). Another possibility is to investigate effects of newly created attitudes that are not confounded with semantic beliefs or non-evaluative stereotypes.

Conclusions

Our analysis suggests that, despite widespread claims of unawareness in the attitude literature, strong empirical evidence for these claims is surprisingly scarce. One potential case of attitudes that evade awareness involves discrepant self-reported evaluations of a type and tokens of that type, but research on this question is still very limited and alternative interpretations of the observed discrepancies have not yet been ruled out. A large body of findings speaks against the idea that indirect measures such as the IAT capture attitudes that people do not know they have. Inferences of unawareness from discrepancies between physiological and self-report measures and between revealed and stated preferences are undermined by conceptual, methodological, and theoretical issues.

Regarding the environmental causes of attitudes, the available evidence suggests that ME effects can occur without conscious recollection of the stimulus event, although such effects may depend on decision strategies applied at the judgment stage. Evidence is mixed for ME effects arising from brief exposures or longer exposures combined with low attention. Counter to the common assumption that CS-US pairings can influence attitudes without awareness of the pairings during encoding, there is no compelling evidence for EC effects under such conditions. The available evidence for EC effects in the absence of subjective feelings of remembering the CS-US pairings is mixed, at best.

Regarding the behavioral effects of attitudes, inferences of unawareness are undermined by the lack of awareness measures in prior research. Because including such measures can raise awareness of behavioral effects that may remain unrecognized in the absence of awareness checks, alternative approaches are needed to provide evidence for claims of unawareness. To establish the postulated causal role of attitudes, future research on this question should also rule out effects of non-attitudinal factors such as stereotypes and semantic beliefs.[2]

It is worth noting that our review did not cover research on contextual influences on evaluative responses. Examples include an object's position in a series of options (Nisbett & Wilson, 1977), incidental mood states (Schwarz & Clore, 2003), and primed mental concepts (Loersch & Payne, 2011), all of which have been claimed to influence evaluative responses without awareness. However, similar to the conscious

[2] A potential objection is that our conclusions are based on a treatment of (un)awareness as a categorical property rather than a continuous dimension. In response to this concern, it is worth noting that categorical treatments tend to favor conclusions of unawareness (e.g., responses on trials may be categorized as "unaware" on a dichotomous indicator, while the same responses would be classified as "aware" on a continuous indicator; see Newell & Shanks, 2023). Thus, if anything, a continuous treatment would suggest even greater concerns about the scarcity of evidence for unawareness.

reliance on processing fluency in ME effects, the relevant *proximal* factor in these studies may be a conscious reliance on the response elicited by the context (e.g., positive mood aroused by a sunny day) when generating an evaluative response to a target object (e.g., conscious reliance on mood when judging one's life satisfaction). More generally, unawareness claims for various contextual influences have been questioned on methodological and empirical grounds (e.g., Adair & Spinner, 1981; Cotton, 1980; Hughes et al., 2023; Newell & Shanks, 2023; White, 1980).

A notable limitation of our analysis is that it focused on general effects without considering the possibility of individual differences in (un)awareness. Although this focus is consistent with the modal approach in this area, it is possible that some people have better self-insight than others (Schriber & Robins, 2012). For example, although participants in Hahn et al.'s (2014) studies were, on average, highly accurate in predicting their IAT scores, there was considerable variability across participants, in that some showed perfect accuracy while others showed extremely poor performance in predicting their IAT scores. Future research extending the dominant focus on general effects to include individual differences may help to gain deeper insights into aspects of attitudes that may evade awareness.

Another question that we have not addressed yet is how people gain awareness of an attitude, its environmental causes, and its behavioral effects. The reviewed evidence speaks only to the question of *whether* people hold accurate beliefs about the three aspects, but it does not address *how* people form such beliefs (see Morris & Kurdi, 2023). Although introspection is widely regarded as an important mechanism to gain self-insight, it is not the only mechanism and the extent to which people have introspective access to psychological processes has been disputed (Nisbett & Wilson, 1977; Wilson & Brekke, 1994). An alternative way to gain knowledge about one's attitude toward an object, originally proposed in Bem's (1972) self-perception theory, is to observe one's evaluative responses to that object under different situational conditions. Because awareness of an attitude is a necessary precondition for awareness of its environmental causes and its behavior effects, the mechanisms by which people form beliefs about their attitudes are central for all three aspects of attitudes. Yet, although self-perception theory has been developed more than half century ago, research on this question is still scarce. Thus, based on the current conclusion that people seem to have much more insight into the three aspects of attitudes than commonly assumed in the literature, an important question for future research is how people gain accurate knowledge of an attitude, its environmental causes, and its behavioral effects.

Going beyond studies that selectively focus on one of the three aspects, another interesting direction for future research involves investigations of their interplay. For example, a recent line of work has started to examine inductive generalizations from tokens to types following exposure to CS-US pairings (e.g., Glaser & Kuchenbrandt, 2017; Hütter et al., 2014; Reichmann et al., 2023). To the extent that generalizations from specific CSs to abstract types of CSs are prone to biases in inductive reasoning (e.g., illusory correlations; see Fiedler & Plessner, 2009), combining a focus on environmental causes with a focus on discrepant evaluations of a type and tokens of that type may provide valuable insights into when and why people are (un)aware of an attitude and its environmental causes. Similarly valuable insights may be gained from other potential combinations of research foci on the three aspects of attitudes. Although the current review suggests that empirical support for claims of unawareness is much weaker than commonly assumed, we do not rule out that novel integrative approaches could provide more compelling evidence. Likewise, the current analysis is agnostic about the possibility of unrecognized influences in attitude paradigms that are not typically seen as involving unawareness and were therefore not discussed in this contribution. To support future research along these lines, Table 1 provides a list of methodological recommendations for sound inferences of (un)awareness, summarizing key points raised throughout this article. We hope that the current analysis provides valuable directions for future studies on the intriguing questions of whether it possible to hold an attitude without being aware of that attitude; whether environmental stimuli can influence attitudes outside of awareness; and whether attitudes can influence behavioral responses in a manner that evades awareness.

References

Adair, J. G., & Spinner, B. (1981). Subjects' access to cognitive processes: Demand characteristics and verbal report. *Journal for the Theory of Social Behaviour*, *11*, 31-52.

Ajzen, I., & Fishbein, M. (1977). Attitude-behavior relations: A theoretical analysis and review of empirical research. *Psychological Bulletin, 84*, 888-918.

Ajzen, I., & Kruglanski, A. W. (2019). Reasoned action in the service of goal pursuit. *Psychological Review, 126*, 774–786.

Alcser-Isais, A. N., Smith, L. K., & Eastwick, P. W. (2022). Inferring one's own attitude toward an unknown attribute: The moderating role of complexity in juice tasting. *Journal of Consumer Behavior*, *21*, 1378-1389.

Amodio, D. M., & Devine, P. G. (2006). Stereotyping and evaluation in implicit race bias: Evidence for independent constructs and unique effects on behavior. *Journal of Personality and Social Psychology, 91*, 652–661.

Banaji, M. R. (2011). A vehicle for large-scale education about the human mind. In J. Brockman (Ed.), *How is the internet changing the way you think?* (pp. 392-395). New York: Harper Collins.

Bargh, J. A. (2022). The cognitive unconscious in everyday life. In A. S. Reber & R. Allen (Eds). *The cognitive unconscious: The first half century* (pp. 89-113). New York: Oxford University Press.

Bem, D. J. (1967). Self-perception: An alternative interpretation of cognitive dissonance phenomena. *Psychological Review*, *74*, 183-200.

Bem, D. J. (1972). Self-perception theory. *Advances in Experimental Social Psychology, 6,* 1-62.

Béna, J., Melnikoff, D. E., Mierop, A., & Corneille, O. (2022). Revisiting dissociation hypotheses with a structural fit approach: The case of the prepared reflex framework. *Journal of Experimental Social Psychology*, *100*, Article 104297.

Brehm, J. W. (1956). Postdecision changes in the desirability of alternatives. *Journal of Abnormal and Social Psychology, 52*, 384-389.

Calanchini, J. (2020). How multinomial processing trees have advanced, and can continue to advance, research using implicit measures. *Social Cognition*, *38*, s165-s186.

Cameron, C. D., Brown-Iannuzzi, J., & Payne, B. K. (2012). Sequential priming measures of implicit social cognition: A meta-analysis of associations with behaviors and explicit attitudes. Personality and Social Psychology Review, 16, 330-350.

Cameron, C. D., Payne, B. K., & Knobe, J. (2010). Do theories of implicit race bias change moral judgments? *Social Justice Research*, *23*, 272-289.

Chen, M. K., & Risen, J. L. (2010). How choice affects and reflects preferences: Revisiting the free-choice paradigm. *Journal of Personality and Social Psychology, 99*, 573-594.

Corneille, O., & Béna, J. (2023). Instruction-based replication studies raise challenging questions for psychological science. *Collabra, 9,* Article 82234.

Corneille, O., & Hütter, M. (2020). Implicit? What do you mean? A comprehensive review of the delusive implicitness construct in attitude research. *Personality and Social Psychology Review, 24,* 212-232.

Corneille, O., & Lush, P. (2023). Sixty years after Orne's American Psychologist article: A conceptual framework for subjective experiences elicited by demand characteristics. *Personality and Social Psychology Review, 27,* 83-101.

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*, 161-189.

Cotton, J. L. (1980). Verbal reports on mental processes: Ignoring data for the sake of the theory? *Personality and Social Psychology Bulletin*, *6*, 278-281.

Cunningham, W. A., Packer, D. J., Kesek, A., & Van Bavel, J. J. (2009). Implicit measurement of attitudes: A physiological approach. In R. E. Petty, R. H. Fazio, & P. Briñol (Eds.), *Attitudes: Insights from the new implicit measures* (pp. 485-512). New York: Psychology Press.

Cunningham, W. A., Preacher, K. J., & Banaji, M. R. (2001). Implicit attitude measurement: Consistency, stability, and convergent validity. *Psychological Science, 12,* 163-170.

Cunningham, W. A., Zelazo, P. D., Packer, D. J., & Van Bavel, J. J. (2007). The iterative reprocessing model: A multilevel framework for attitudes and evaluation. *Social Cognition*, *25*, 736-760.

Da Silva Frost, A., Wang, Y. A., Eastwick, P. W., & Ledgerwood, A. (2023). Summarized attribute preferences have unique antecedents and consequences. *Journal of Experimental Psychology: General.* Advance Online Publication.

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,* Article 103812

De Corte, K., Cairns, J., & Grieve, R. (2021). Stated versus revealed preferences: An approach to reduce bias. *Health Economics*, *30*, 1095-1123.

De Houwer, J. (2006). Using the Implicit Association Test does not rule out an impact of conscious propositional knowledge on evaluative conditioning. *Learning and Motivation, 37,* 176-187.

De Houwer, J., Gawronski, B., & Barnes-Holmes, D. (2013). A functional-cognitive framework for attitude research. *European Review of Social Psychology, 24,* 252-287.

De Zilva, D., Vu, L., Newell, B. R., & Pearson, J. (2013). Exposure is not enough: Suppressing stimuli from awareness can abolish the mere exposure effect. *PloS ONE*, *8*(10), e77726.

Dedonder, J., Corneille, O., Bertinchamps, D., & Yzerbyt, V. (2014). Overcoming correlational pitfalls: Experimental evidence suggests that evaluative conditioning occurs for explicit but not implicit encoding of CS–US pairings. *Social Psychological and Personality Science*, *5*, 250-257.

Dovidio, J. F., & Gaertner, S. L. (2004). Aversive racism. *Advances in Experimental Social Psychology, 36,* 1-52.

Dovidio, J. F., Kawakami, K., & Gaertner, S. L. (2002). Implicit and explicit prejudice and interracial interaction. *Journal of Personality and Social Psychology, 82*, 62-68.

Duncan, B. L. (1976). Differential perception and attribution of intergroup violence: Testing the lower limits of stereotyping of Blacks. *Journal of Personality and Social Psychology, 34,* 590-598.

Dunton, B. C., & Fazio, R. H. (1997). An individual difference measure of motivation to control prejudiced reactions. *Personality and Social Psychology Bulletin, 23,* 316-326.

Eagly, A. H., & Chaiken, S. (2007). The advantages of an inclusive definition of attitude. *Social Cognition, 25,* 582-602.

Eagly, A. H., & Mladinic, A. (1989). Gender stereotypes and attitudes toward women and men. *Personality and Social Psychology Bulletin*, *15*, 543-558.

Eagly, A. H., & Mladinic, A. (1994). Are people prejudiced against women? Some answers from research on attitudes, gender stereotypes, and judgments of competence. *European Review of Social Psychology*, *5*, 1-35.

Eastwick, P. W., & Finkel, E. J. (2008). Speed-dating: A powerful and flexible paradigm for studying romantic relationship initiation. In S. Sprecher, A. Wenzel, & J. Harvey (Eds.), *Handbook of relationship initiation* (pp. 217-234). New York: Psychology Press.

Fazio, R. H. (1990). Multiple processes by which attitudes guide behavior: The MODE model as an integrative framework. *Advances in Experimental Social Psychology, 23,* 75-109.

Fazio, R. H. (2007). Attitudes as object–evaluation associations of varying strength. *Social Cognition*, *25*, 603-637.

Fazio, R. H., Jackson, J. R., Dunton, B. C., & Williams, C. J. (1995). Variability in automatic activation as an unobtrusive measure of racial attitudes: A bona fide pipeline? *Journal of Personality and Social Psychology, 69,* 1013-1027.

Fazio, R. H., Samayoa, J. A. G., Boggs, S. T., & Ladanyi, J. (in press). What is implicit bias? In J. A., Krosnick, T. H., Stark, & A. L. Scott (Eds.), *The Cambridge handbook of implicit bias and racism*. Cambridge: Cambridge University Press.

Fazio, R. H., Sanbonmatsu, D. M., Powell, M. C., & Kardes, F. R. (1986). On the automatic activation of attitudes. *Journal of Personality and Social Psychology, 50*, 229–238.

Festinger, L., & Carlsmith, J. M. (1959). Cognitive consequences of forced compliance. *Journal of Abnormal and Social Psychology*, *58*, 203-210.

Fiedler, K., & Plessner, H. (2009). Induction: From simple categorization to higher-order inference problems. In F. Strack & J. Förster (Eds.), *Social cognition: The basis of human interaction* (pp. 93-120). New York: Psychology Press.

Förster, J. (2009). Cognitive consequences of novelty and familiarity: How mere exposure influences level of construal. *Journal of Experimental Social Psychology*, *45*, 444-447.

Fox, M. C., Ericsson, K. A., & Best, R. (2011). Do procedures for verbal reporting of thinking have to be reactive? A meta-analysis and recommendations for best reporting methods. *Psychological Bulletin*, *137*, 316-344.

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

Gawronski, B., & Bodenhausen, G. V. (2012). Self-insight from a dual-process perspective. In S. Vazire & T. D. Wilson (Eds.), *Handbook of self-knowledge* (pp. 22-38). New York: Guilford Press.

Gawronski, B., & Bodenhausen, G. V. (2015). Social-cognitive theories. In B. Gawronski, & G. V. Bodenhausen (Eds.), *Theory and explanation in social psychology* (pp. 65-83). New York: Guilford Press.

Gawronski, B., Bodenhausen, G. V., & Becker, A. P. (2007). I like it, because I like myself: Associative self-anchoring and post-decisional change of implicit evaluations. *Journal of Experimental Social Psychology, 43*, 221-232.

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: Routledge.

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., Geschke, D., & Banse, R. (2003). Implicit bias in impression formation: Associations influence the construal of individuating information. *European Journal of Social Psychology, 33,* 573-589.

Gawronski, B., LeBel, E. P., & Peters, K. R. (2007). What do implicit measures tell us? Scrutinizing the validity of three common assumptions. *Perspectives on Psychological Science, 2*, 181-193.

Gawronski, B., Ledgerwood, A., & Eastwick, P. W. (2022a). Implicit bias ≠ bias on implicit measures. *Psychological Inquiry*, *33*, 139-155.

Gawronski, B., Ledgerwood, A., & Eastwick, P. W. (2022b). Reflections on the difference between implicit bias and bias on implicit measures. *Psychological Inquiry*, *33*, 219-231.

Gawronski, B., Peters, K. R., & LeBel, E. P. (2008). What makes mental associations personal or extra-personal? Conceptual issues in the methodological debate about implicit attitude measures. *Social and Personality Psychology Compass, 2,* 1002-1023.

Gawronski, B., & Walther, E. (2012). What do memory data tell us about the role of contingency awareness in evaluative conditioning? *Journal of Experimental Social Psychology, 48,* 617-623.

Glaser, T., & Kuchenbrandt, D. (2017). Generalization effects in evaluative conditioning: Evidence for attitude transfer effects from single exemplars to social categories. *Frontiers in Psychology*, *8*, Article 103.

Glick, P., Zion, C., & Nelson, C. (1988). What mediates sex discrimination in hiring decisions? *Journal of Personality and Social Psychology, 55*, 178-186.

Goedderz, A., & Hahn, A. (2022). Biases left unattended: People are surprised at racial bias feedback until they pay attention to their biased reactions. *Journal of Experimental Social Psychology*, *102*, Article 104374.

Goffin, R. D., & Olson, J. M. (2011). Is it all relative? Comparative judgments and the possible improvement of self-ratings and ratings of others. *Perspectives on Psychological Science*, *6*, 48-60.

Greenwald, A. G., & Lai, C. K. (2020). Implicit social cognition. *Annual Review of Psychology, 71,* 419–445.

Greenwald, A. G., McGhee, D. E., & Schwartz, J. K. L. (1998). Measuring individual differences in implicit cognition: The Implicit Association Test. *Journal of Personality and Social Psychology, 74,* 1464-1480.

Greifeneder, R., & Bless, H. (2007). Relying on accessible content versus accessibility experiences: The case of processing capacity. *Social Cognition, 25,* 853-881.

Greifeneder, R., & Schwarz, N. (2014). Metacognitive processes and subjective experiences. In J. W. Sherman, B. Gawronski, & Y. Trope (Eds.), *Dual-process theories of the social mind* (pp. 314-327). New York: Guilford Press.

Griggs, B. (2018, December 28). Living while black: Here are all the routine activities for which police were called on African-Americans this year. *CNN*. Retrieved from https://www.cnn.com/2018/12/20/us/living-while-black-police-calls-trnd/index.html (July 26, 2023).

Hahn, A., & Gawronski, B. (2019). Facing one's implicit biases: From awareness to acknowledgment. *Journal of Personality and Social Psychology*, *116,* 769-794.

Hahn, A., & Goedderz, A. (2020). Trait-unconsciousness, state-unconsciousness, preconsciousness, and social miscalibration in the context of implicit evaluations. *Social Cognition, 38,* s115-s134.

Hahn, A., Judd, C. M., Hirsh, H. K., & Blair, I. V. (2014). Awareness of implicit attitudes. *Journal of Experimental Psychology: General*, *143*, 1369-1392.

Hansen, J., & Wänke, M. (2009). Liking what's familiar: The importance of unconscious familiarity in the mere-exposure effect. *Social Cognition*, *27*, 161-182.

Högden, F., Hütter, M., & Unkelbach, C. (2018). Does evaluative conditioning depend on awareness? Evidence from a continuous flash suppression paradigm. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *44*, 1641-1657.

Hofmann, W., Gawronski, B., Gschwendner, T., Le, H., & Schmitt, M. (2005). A meta-analysis on the correlation between the Implicit Association Test and explicit self-report measures. *Personality and Social Psychology Bulletin, 31*, 1369-1385.

Huang, Y. F., & Hsieh, P. J. (2013). The mere exposure effect is modulated by selective attention but not visual awareness. *Vision Research*, *91*, 56-61.

Hughes, S., Cummins, J., & Hussey, I. (2023). Effects on the Affect Misattribution Procedure are strongly moderated by influence awareness. *Behavior Research Methods*, *55*, 1558-1586.

Hütter, M., Kutzner, F., & Fiedler, K. (2014). What is learned from repeated pairings? On the scope and generalizability of evaluative conditioning. *Journal of Experimental Psychology: General, 143*, 631-643.

Hütter, M., & Sweldens, S. (2013). Implicit misattribution of evaluative responses: Contingency-unaware evaluative conditioning requires simultaneous stimulus presentations. *Journal of Experimental Psychology: General, 142*, 638–643.

Hugenberg, K., & Bodenhausen, G. V. (2003). Facing prejudice: Implicit prejudice and the perception of facial threat. *Psychological Science, 14,* 640-643.

Ito, T. A., & Cacioppo, J. T. (2007). Attitudes as mental and neural states of readiness: Using physiological measures to study implicit attitudes. In B. Wittenbrink & N. Schwarz (Eds.), *Implicit measures of attitudes* (pp. 125-158). New York: Guilford Press.

Jacoby, L. L. (1991). A process dissociation framework: Separating automatic from intentional uses of memory. *Journal of Memory and Language*, *30*, 513-541.

Jurchiș, R., Costea, A., Dienes, Z., Miclea, M., & Opre, A. (2020). Evaluative conditioning of artificial grammars: Evidence that subjectively-unconscious structures bias affective evaluations of novel stimuli. *Journal of Experimental Psychology: General*, *149*, 1800-1809.

Kasran, S., Hughes, S., & De Houwer, J. (2022). Learning via instructions about observations: Exploring similarities and differences with learning via actual observations. *Royal Society Open Science*, *9*, Article 220059.

Kawakami, N., & Yoshida, F. (2019). Subliminal versus supraliminal mere exposure effects: Comparing explicit and implicit attitudes. *Psychology of Consciousness: Theory, Research, and Practice, 6*, 279-291.

Kouider, S., & Dehaene, S. (2007). Levels of processing during non-conscious perception: a critical review of visual masking. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *362*, 857-875.

Krickel, B. (2018). Are the states underlying implicit biases unconscious? A neo-Freudian answer. *Philosophical Psychology, 31,* 1007-1026.

Krosnick, J. A., Betz, A. L., Jussim, L. J., & Lynn, A. R. (1992). Subliminal conditioning of attitudes. *Personality and Social Psychology Bulletin*, *18*, 152-162.

Krosnick, J. A., Judd, C. M., & Wittenbrink, B. (2005). The measurement of attitudes. In D. Albarracín, B. T. Johnson, & M. P. Zanna (Eds.), *The handbook of attitudes* (pp. 21-76). Mahwah, NJ: Erlbaum.

Kruglanski, A. W. (1989). The psychology of being "right": The problem of accuracy in social perception and cognition. *Psychological Bulletin, 106*, 395–409

Kühnen, U. (2010). Manipulation checks as manipulation: Another look at the ease-of-retrieval heuristic. *Personality and Social Psychology Bulletin, 36,* 47-58.

Kunda, Z., & Sherman-Williams, B. (1993). Stereotypes and the construal of individuating information. *Personality and Social Psychology Bulletin, 19,* 90-99.

Kurdi, B., Hussey, I., Stahl, C., Hughes, S., Unkelbach, C., Ferguson, M., & Corneille, O. (2022). Unaware attitude formation in the surveillance task? Revisiting the findings of Moran et al. (2021). *International Review of Social Psychology*, 35, Article 6.

Lang, P. J., Bradley, M. M., & Cuthbert, B. N. (2008). *International affective picture system (IAPS): Affective ratings of pictures and instruction manual. Technical Report A-7*. University of Florida, Gainesville, FL.

Ledgerwood, A., Eastwick, P. W., & Gawronski, B. (2020). Experiences of liking versus ideas about liking. *Behavioral and Brain Sciences*, *43*, Article e136.

Ledgerwood, A., Eastwick, P. W., & Smith, L. K. (2018). Toward an integrative framework for studying human evaluation: Attitudes toward objects and attributes. *Personality and Social Psychology Review*, *22*, 378-398.

Lieberman, M. D., Ochsner, K. N., Gilbert, D. T., & Schacter, D. L. (2001). Do amnesics exhibit cognitive dissonance reduction? The role of explicit memory and attention in attitude change. *Psychological Science, 12,* 135-140.

Loersch, C., & Payne, B. K. (2011). The situated inference model: An integrative account of the effects of primes on perception, behavior, and motivation. *Perspectives on Psychological Science, 6,* 234-252.

Mierop, A., Hütter, M., & Corneille, O. (2017). Resource availability and explicit memory largely determine evaluative conditioning effects in a paradigm claimed to be conducive to implicit attitude acquisition. *Social Psychological and Personality Science, 8*, 758-767.

Montoya, R. M., Horton, R. S., Vevea, J. L., Citkowicz, M., & Lauber, E. A. (2017). A re-examination of the mere exposure effect: The influence of repeated exposure on recognition, familiarity, and liking. *Psychological Bulletin, 143*, 459-498.

Moors, A. (2016). Automaticity: Componential, causal, and mechanistic explanations. *Annual Review of Psychology*, *67*, 263-287.

Moran, T., Hughes, S., Hussey, I., Vadillo, M. A., Olson, M. A., Aust, F., … De Houwer, J. (2021). Incidental attitude formation via the surveillance task: A preregistered replication of the Olson and Fazio (2001) study. *Psychological Science*, *32*, 120-131.

Moran, T., Nudler, Y., & Bar-Anan, Y. (2023). Evaluative conditioning: Past, present, and future. *Annual Review of Psychology*, *74*, 245-269.

Morris, A., & Kurdi, B. (2023). Awareness of implicit attitudes: Large-scale investigations of mechanism and scope. *Journal of Experimental Psychology: General*. Advance Online Publication.

Moss-Racusin, C. A., Dovidio, J. F., Brescoll, V. L., Graham, M. J., & Handelsman, J. (2012). Science faculty's subtle gender biases favor male students. *Proceedings of the National Academy of Sciences, 109,* 16474-16479.

Neal, D. T., Wood, W., Wu, M., & Kurlander, D. (2011). The pull of the past: When do habits persist despite conflict with motives? *Personality and Social Psychology Bulletin, 37,* 1428-1437.

Newell, B. R., & Shanks, D. R. (2007). Recognising what you like: Examining the relation between the mere-exposure effect and recognition. *European Journal of Cognitive Psychology*, *19*, 103-118.

Newell, B., & Shanks, D. (2014). Unconscious influences on decision making: A critical review. *Behavioral and Brain Sciences, 37*, 1-19.

Newell, B. R., & Shanks, D. R. (2023). *Open minded: Searching for truth about the unconscious mind*. Cambridge, MA: MIT Press.

Nier, J. A. (2005). How dissociated are implicit and explicit racial attitudes? A bogus pipeline approach. *Group Processes and Intergroup Relations*, *8*, 39-52.

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review, 84*, 231-259.

Norton, M. I., Vandello, J. A., & Darley, J. M. (2004). Casuistry and social category bias. *Journal of Personality and Social Psychology, 87,* 817-831.

Olson, J. M., Goffin, R. D., & Haynes, G. A. (2007). Relative versus absolute measures of explicit attitudes: Implications for predicting diverse attitude-relevant criteria. *Journal of Personality and Social Psychology, 93*, 907–926.

Olson, M. A., & Fazio, R. H. (2001). Implicit attitude formation through classical conditioning. *Psychological Science*, *12*, 413-417.

Olson, M. A., Fazio, R. H., & Han, H. A. (2009). Conceptualizing personal and extrapersonal associations. *Social and Personality Psychology Compass, 3*, 152–170.

Olson, M. A., Fazio, R. H., & Hermann, A. D., Sr. (2007). Reporting tendencies underlie discrepancies between implicit and explicit measures of self-esteem. *Psychological Science, 18*, 287-291.

Payne, B. K., Burkley, M. A., & Stokes, M. B. (2008). Why do implicit and explicit attitude tests diverge? The role of structural fit. *Journal of Personality and Social Psychology, 94*, 16-31.

Payne, B. K., Cheng, S. 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.

Petty, R. E., Briñol, P., Fabrigar, L. R., & Wegener, D. T. (2019). Attitude structure and change. In R. F. Baumeister & E. J. Finkel (Eds.), *Advanced social psychology* (2nd Edition, pp. 117-156). Oxford: Oxford University Press.

Phills, C. E., Hahn, A., & Gawronski, B. (2020). The bidirectional causal relation between implicit stereotypes and implicit prejudice. *Personality and Social Psychology Bulletin*, *46,* 1318-1330.

Pleyers, G., Corneille, O., Luminet, O., & Yzerbyt, V. (2007). Aware and (dis)liking: Item-based analyses reveal that valence acquisition via evaluative conditioning emerges only when there is contingency awareness. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *33*, 130-144.

Pleyers, G., Corneille, O., Yzerbyt, V., & Luminet, O. (2009). Evaluative conditioning may incur attentional costs. *Journal of Experimental Psychology: Animal Behavior Processes, 35*, 279-285.

Ranganath, K. A., Smith, C. T., & Nosek, B. A. (2008). Distinguishing automatic and controlled components of attitudes from direct and indirect measurement methods. *Journal of Experimental Social Psychology*, *44*, 386-396

Reichmann, K., Hütter, M., Kaup, B., & Ramscar, M. (2023). Variability and abstraction in evaluative conditioning: Consequences for the generalization of likes and dislikes. *Journal of Experimental Social Psychology*, *108*, Article 104478.

Roskos-Ewoldsen, D. R., & Fazio, R. H. (1992). On the orienting value of attitudes: Attitude accessibility as a determinant of an object's attraction of visual attention. *Journal of Personality and Social Psychology, 63*, 198–211

Sagar, H. A., & Schofield, J. W. (1980). Racial and behavioral cues in black and white children's perceptions of ambiguously aggressive acts. *Journal of Personality and Social Psychology, 39,* 590-598.

Schriber, R. A., & Robins, R. W. (2012). Self-knowledge: An individual-differences perspective. In S. Vazire & T. D. Wilson (Eds.), *Handbook of self-knowledge* (pp. 105-127). New York: Guilford Press.

Schwarz, N. (2004). Metacognitive experiences in consumer judgment and decision making. *Journal of Consumer Psychology, 14,* 332-348.

Schwarz, N., Bless, H., Strack, F., Klumpp, G., Rittenauer-Schatka, H., & Simons, A. (1991). Ease of retrieval as information: Another look at the availability heuristic. *Journal of Personality and Social Psychology, 61*, 195-202.

Schwarz, N., & Clore, G. L. (2003). Mood as information: 20 years later. *Psychological Inquiry*, *14*, 296-303.

Shanks, D. R., Malejka, S., & Vadillo, M. A. (2021). The challenge of inferring unconscious mental processes. *Experimental Psychology, 68*, 113-129.

Shanks, D. R., & St. John, M. F. (1994). Characteristics of dissociable human learning systems. *Behavioral and Brain Sciences*, *17*, 367-395.

Stahl, C., Béna, J., Aust, F., Mierop, A., & Corneille, O. (2023). A conditional judgment procedure for probing evaluative conditioning effects in the absence of feelings of remembering. *Behavior Research Methods*. Advance Online Publication.

Stahl, C., Haaf, J., & Corneille, O. (2016). Subliminal evaluative conditioning? Above-chance CS identification may be necessary and insufficient for attitude learning. *Journal of Experimental Psychology: General*, *145*, 1107-1131.

Stahl, C., Unkelbach, C., & Corneille, O. (2009). On the respective contributions of awareness of unconditioned stimulus valence and unconditioned stimulus identity in attitude formation through evaluative conditioning. *Journal of Personality and Social Psychology*, *97*, 404-420.

Strack, F., & Hannover, B. (1996). Awareness of the influence as a precondition for implementing correctional goals. In P. M. Gollwitzer & J. A. Bargh

(Eds.), *The psychology of action: Linking cognition and motivation to behavior* (pp. 579-596). New York: Guilford Press.

Sweldens, S., Corneille, O., & Yzerbyt, V. (2014). The role of awareness in attitude formation through evaluative conditioning. *Personality and Social Psychology Review*, *18*, 187-209.

Timmermans, B., & Cleeremans, A. (2015). How can we measure awareness? An overview of current methods. In M. Overgaard (Ed.), *Behavioral methods in consciousness research* (pp. 21-46). Oxford, UK: Oxford University Press.

Uhlmann, E. L., & Cohen, G. L. (2005). Constructed criteria: Redefining merit to justify discrimination. *Psychological Science, 16,* 474-480.

Van Dessel, P., De Houwer, J., Gast, A., & Tucker Smith, C. (2015). Instruction-based approach-avoidance effects: Changing stimulus evaluation via the mere instruction to approach or avoid stimuli. *Experimental Psychology*, *62*, 161-169.

Van Dessel, P., Mertens, G., Smith, C. T., & De Houwer, J. (2017). The mere exposure instruction effect. *Experimental Psychology, 64,* 299-314.

Walther, E., & Nagengast, B. (2006). Evaluative conditioning and the awareness issue: Assessing contingency awareness with the Four-Picture Recognition Test. *Journal of Experimental Psychology: Animal Behavior Processes, 32*, 454-459.

Waroquier, L., Abadie, M., & Dienes, Z. (2020). Distinguishing the role of conscious and unconscious knowledge in evaluative conditioning. *Cognition*, *205*, Article 104460.

Wegener, D. T., & Petty, R. E. (1997). The flexible correction model: The role of naive theories of bias in bias correction. *Advances in Experimental Social Psychology, 29,* 141-208.

White, P. (1980). Limitations on verbal reports of internal events: A refutation of Nisbett and Wilson and of Bem. *Psychological Review, 87*, 105-112.

Whittlesea, B. W. A., & Price, J. R. (2001). Implicit/explicit memory versus analytic/nonanalytic processing: Rethinking the mere exposure effect. *Memory & Cognition, 29,* 234-246.

Wilson, T. D., & Brekke, N. (1994). Mental contamination and mental correction: Unwanted influences on judgments and evaluations. *Psychological Bulletin, 116,* 117-142.

Woiczyk, T. K. A., & Le Mens, G. (2021). Evaluating categories from experience: The simple averaging heuristic. *Journal of Personality and Social Psychology, 121*, 747-773.

Wolsiefer, K., Westfall, J., & Judd, C. M. (2017). Modeling stimulus variation in three common implicit attitude tasks. *Behavior Research Methods*, *49*, 1193-1209.

Zajonc, R. B. (1968). Attitudinal effects of mere exposure. *Journal of Personality and Social Psychology, 9*, 1-27.

Table 1. Methodological recommendations for studies investigating the (un)awareness of attitudes, their environmental causes, and their behavior effects.

1. Do not conflate unawareness with other features of automaticity (e.g., unintentionality, efficiency, uncontrollability).
2. Be specific about the awareness question you are interested in (e.g., awareness of what? awareness at which processing stage?).
3. Do not assume that the mere use of a particular procedure (e.g., short presentation times; indirect measures) guarantees unawareness. Instead, check for unawareness in these procedures using independent criteria.
4. Ensure that measures of evaluative responses and measures of awareness are held constant on procedural factors (e.g., stimuli, timing of measured responses).
5. Ensure that measures of evaluative responses and measures of awareness have comparably high reliability and sensitivity in capturing the to-be-measured constructs.
6. Rule out alternative explanations before inferring unawareness from unreported attitudes or influences (e.g., unwillingness to report).
7. Consider that between-subjects approaches to studying (un)awareness confound self-insight with knowledge about other participants in the sample.
8. Rule out spurious effects of attitudes driven by confounds with non-evaluative representations (e.g., semantic beliefs, stereotypes).
9. Whenever possible, investigate unawareness questions using experimental approaches (e.g., manipulating attention level) rather than correlational designs (e.g., linking performance to retrospective memory reports or responses in funnel debriefings).

Figure 1. Three aspects of attitudes for which people may lack awareness. First, people may be unaware of the attitude itself, defined as psychological tendency that is expressed by evaluating a particular entity with some degree of favor or disfavor. Second, people may be unaware of the environmental cause of the attitude, including stimulus events that are responsible for an attitude and the causal influence of stimulus events on an attitude. Third, people may be unaware of the behavioral effect of the attitude, including behavioral responses that are influenced by the attitude and the causal influence of the attitude on behavioral responses.