Attitudes and the Implicit-Explicit Dualism

Bertram Gawronski & Skylar M. Brannon
University of Texas at Austin

Since the mid-1990s, research on attitudes has been shaped by a dualism that has gained enormous popularity across all areas of psychology: the implicit-explicit dualism (see Gawronski & Payne, 2010). This dualism has its roots in the development of a new class of indirect measurement instruments, which are distinguished from direct measurement instruments based on self-report. A central feature of these instruments is that they rely on experimental procedures adapted from cognitive psychology, such as sequential priming and response interference tasks (for an overview, see Gawronski & De Houwer, 2014). Researchers often label these instruments *implicit measures* and self-report measures *explicit measures*.

A central characteristic of implicit measures as applied to the study of attitudes is that evaluative responses are inferred from objective performance indicators, such as participants' speed and accuracy in responding to attitudinal stimuli. Based on this characteristic, it has been argued that implicit measures are capable of assessing attitudes that people are either unwilling or unable to report. Resonating with these concerns, self-report measures have been criticized for their susceptibility to socially desirable responding (Crowne & Marlowe, 1960; Paulhus, 1984) and for being unable to capture mental contents that are inaccessible to introspection (Greenwald & Banaji, 1995; Nisbett & Wilson, 1977). Implicit measures have been claimed to overcome these limitations, because (a) responses on implicit measures are much more difficult to control compared to explicit measures, and (b) implicit measures do not require introspection for the assessment of mental contents.

Although these assumptions are very common in the attitude literature, the available evidence suggests that differences between implicit and explicit measures cannot be boiled down to self-presentation and unawareness (Gawronski, LeBel, & Peters, 2007). Instead, the exponentially growing body of research in this area suggests that interpretations of different outcomes on the two kinds of measures require more nuanced theoretical frameworks. The current chapter provides an overview of extant theories and research on the implicit-explicit dualism in the attitude literature. Toward this end, we first provide a brief overview of the most popular instruments and then review different interpretations of the implicit-explicit dualism. Expanding on this discussion, we review theories and empirical findings regarding (a) the relation between implicit and explicit measures, (b) their predictive relation to behavior, (c) their usefulness in understanding the processes underlying attitude formation and change, (d) context effects on implicit measures, and (e) controversies and current themes in research using implicit measures.

Measurement Instruments

Although there are more than a dozen performance-based measures of attitudes that have been described as implicit, some of these measures tend to be more popular than others. In the current section, we briefly describe the procedural details of the most popular instruments and provide a list of less frequently used instruments for the sake of comprehensiveness. For readers who are interested in learning more about their implementation and scoring, we recommend the method-focused overviews by Teige-Mocigemba, Klauer, and Sherman (2010) and Wentura and Degner (2010). Broader issues in the measurement of attitudes are discussed in the chapter by Krosnick, Judd, and Wittenbrink (this volume).

Evaluative Priming Task

Historically, the first performance-based instrument to measure attitudes is the *evaluative priming task* (EPT; Fazio, Sanbonmatsu, Powell, & Kardes, 1986). In a typical EPT, participants are briefly presented with an attitudinal prime stimulus, which is followed by a positive or negative target word. Participants' task is to indicate as quickly as possible whether the target word is positive or negative.[1] The basic idea underlying the EPT is that quick and accurate responses to the targets should be facilitated when they are evaluatively congruent with participants' attitude toward the prime stimulus. In contrast, quick and accurate responses to the targets should be impaired when they are evaluatively incongruent with participants' attitude toward the prime

[1] Alternative variants include sequential priming with lexical decision tasks in which participants are asked to classify the targets as meaningful words versus meaningless non-words (e.g., Wittenbrink, Judd, & Park, 1997) and sequential priming with semantic categorization tasks in which participants are asked to classify the targets in terms of a non-evaluative, semantic dimension (e.g., Banaji & Hardin, 1996). Because these variants do not capture evaluative responses, they are less common in research on attitudes and more frequently used in research on semantic aspects of mental contents (e.g., stereotypes).

Chapter to appear in: D. Albarracín & B. T. Johnson (Eds.). *The handbook of attitudes* (2nd edition). New York: Taylor & Francis.

stimulus (e.g., Fazio, Jackson, Dunton, & Williams, 1995).

For example, if a person has a positive attitude toward Coke, this person should be faster and more accurate in identifying the valence of positive words when the person has been primed with an image of Coke compared to when they have been primed with a neutral baseline stimulus. Conversely, classifications of negative words should be slower and less accurate when the person has been primed with an image of Coke compared to when they have been primed with a neutral baseline stimulus. Although the EPT is among the most widely used instruments, it has been criticized for its low reliability, which rarely exceeds Cronbach's Alpha values of .50 (Gawronski & De Houwer, 2014).

Implicit Association Test (and Variants)

The most prominent instrument in attitude research using implicit measures is the implicit association test (IAT; Greenwald, McGhee, & Schwartz, 1998). The critical blocks of the IAT ask participants to complete two binary categorization tasks that are combined in a manner that is either congruent or incongruent with the to-be-measured attitude. For example, in the widely used race IAT, participants categorize pictures of Black and White faces in terms of their race and positive and negative words in terms of their valence. In one critical block of the task, participants are asked to press one response key for Black faces and negative words and another response key for White faces and positive words. In the other critical block, participants are asked to complete the same categorization tasks with a reversed key assignment for the faces, such that they have to press one response key for White faces and negative words and the other response key for Black faces and positive words. The basic idea underlying the IAT is that quick and accurate responses should be facilitated when the response mapping is congruent with participants' attitude and impaired when the response mapping is incongruent with participants' attitude. For example, a person who has a more favorable attitude toward Whites than Blacks should show faster and more accurate responses when White faces share the same response key with positive words and Black faces and share the same response key with negative words, compared with the reversed mapping.

IAT scores are inherently relative in the sense that they conflate four conceptually independent constructs. For example, in the race IAT, a participant's performance is jointly determined by (a) positivity toward Whites, (b) positivity toward Blacks, (c) negativity toward Whites, and (d) negativity toward Blacks (see Blanton, Jaccard, Gonzales, & Christie, 2006). This limitation makes the IAT inferior to the EPT, which permits the calculation of separate priming scores for each of the four determinants if the task includes appropriate baseline primes (see Wentura & Degner, 2010). Yet, the IAT is superior in terms of its internal consistency, which is typically in the range of .70 to .90 (Gawronski & De Houwer, 2014). At the same time, the IAT has been criticized for its blocked presentation of attitude-congruent and attitude-incongruent trials, which has been linked to several sources of systematic measurement error (see Teige-Mocigemba et al., 2010). To address these and various other limitations, researchers have developed several variants of the standard IAT that avoid blocked presentations of congruent and incongruent trials, permit non-relative measurements for individual targets and attributes, and reduce the overall length of the task. These IAT variants include the Recoding-Free IAT (IAT-RF; Rothermund, Teige-Mocigemba, Gast, & Wentura, 2009), the Single-Block IAT (SB-IAT; Teige-Mocigemba, Klauer, & Rothermund, 2008), the Single-Category IAT (SC-IAT; Karpinski & Steinman, 2006), the Single-Attribute IAT (SA-IAT; Penke, Eichstaedt, & Asendorpf, 2006), and the Brief IAT (BIAT; Sriram & Greenwald, 2009).

Go/No-Go Association Task

Another task that has been developed with the goal of overcoming the relative nature of measurement scores in the standard IAT is the go/no-go association task (GNAT; Nosek & Banaji, 2001). On the GNAT, participants are asked to press a button (*go*) in response to some stimuli, and to withhold a response (*no go*) to other stimuli. Different types of stimuli are then paired with the "go" response on different blocks of the task. For example, in one block of a GNAT to measure racial attitudes, participants may be asked to press the "go" button when they see a picture of a Black face or a positive word, and not respond to any other stimuli (which may include pictures of White faces, negative words, and distractor stimuli). In another block, participants may be asked to press the "go" button for pictures of Black faces and negative words, and not respond to any other stimuli. The same task may be repeated in two additional blocks for White instead of Black faces. Because GNAT scores are calculated on the basis of participants' error rates (rather than response times) using signal detection theory (Green & Swets, 1966), the GNAT typically includes a response deadline (e.g., 600 ms) to increase the number of systematic errors. The GNAT has shown lower reliability estimates compared with the standard IAT (Gawronski & De Houwer, 2014). Yet, a clear advantage is the possibility to calculate GNAT scores for individual target objects (e.g., attitudes toward Blacks) instead of relative scores involving two target objects (e.g., relative preference for Whites over Blacks).

Extrinsic Affective Simon Task

Another measure that has been designed to address structural limitations of the IAT is the extrinsic affective

Simon task (EAST; De Houwer, 2003). On the EAST, participants are presented with attitudinal target words (e.g., Pepsi) that are shown in two different colors (e.g., yellow vs. blue) and positive and negative words that are shown in white. Participants then categorize the attitudinal target words in terms of their color and the white words in terms of their valence. In the critical block of the task, participants respond to positive white words and attitudinal target words of one color (e.g., yellow) with the same key and to negative white words and attitudinal target words of the other color (e.g., blue) with another key (or vice versa). Because the attitudinal target words appear in different colors over the course of the task, each target is sometimes paired with the response key for positive words and sometimes with the response key for negative words. The critical question is whether participants respond faster and more accurately to a given target depending on whether its color requires a response with the "positive" or the "negative" key (e.g., are responses faster and more accurate when participants have to respond to the word *Pepsi* with the "positive" or the "negative" key). A major advantage of the EAST is that it does not include blocked presentations of congruent and incongruent trials, which resolves the problems associated with the blocked structure of the IAT (see Teige-Mocigemba et al., 2010). Yet, the EAST has been shown to be inferior to the IAT in terms of its reliability and construct validity, which has been attributed to the fact that participants do not have to process the semantic meaning of the colored target words (De Houwer & De Bruycker, 2007a). To address this limitation, De Houwer and De Bruycker (2007b) have developed a modified variant of the EAST that ensures semantic processing of the target words, which they called the Identification-EAST (ID-EAST).

Affect Misattribution Procedure

The affect misattribution procedure (AMP; Payne, Cheng, Govorun, & Stewart, 2005) was designed to combine the structural advantages of the EPT with the superior psychometric properties of the IAT (for a review, see Payne & Lundberg, 2014). Similar to the EPT, the AMP utilizes presentations of a prime followed by a target stimulus, and participants are asked to respond to the targets and ignore the primes. However, two central differences between the EPT and the AMP are that (a) the target stimuli in the AMP are evaluatively ambiguous and (b) participants are asked to report their subjective evaluations of the targets. The basic idea is that participants may misattribute the affective feelings elicited by primes to the neutral targets, and therefore judge the targets more favorably when they were primed with a positive stimulus than when they were primed with a negative stimulus. For example, in an AMP to measure racial attitudes, participants may be asked to indicate whether they find Chinese ideographs—which tend to be evaluatively neutral to people who are unfamiliar with the meaning of Chinese ideographs—visually more pleasant or visually less pleasant than average after being primed with pictures of Black versus White faces. A preference for Whites over Blacks would be indicated by a tendency to evaluate the Chinese ideographs more favorably when the ideographs followed the presentation of a White face than when they followed the presentation of a Black face. Interestingly, priming effects in the AMP emerge even when participants are explicitly informed about the nature of the task and instructed not to let the prime stimuli influence their evaluations of the targets (Payne et al., 2005).

Although the AMP has shown satisfactory reliability estimates that are comparable to those of the IAT (Gawronski & De Houwer, 2014; Payne & Lundberg, 2014), the task has been criticized for being susceptible to intentional use of the primes in evaluations of the targets (Bar-Anan & Nosek, 2012). Nonetheless, several follow-up studies have refuted this criticism by showing that correlations between AMP effects and self-reported intentional use of the primes reflect retrospective confabulations of intentionality (i.e., participants infer that they must have had such an intention when asked afterwards) rather than actual effects of intentional processes (e.g., Gawronski & Ye, 2015; Payne, Brown-Iannuzzi, Burkley, Arbuckle, Cooley, Cameron, & Lundberg, 2013).

Approach-Avoidance Tasks

Approach-avoidance tasks (AAT) are based on the idea that positive stimuli should elicit spontaneous approach reactions, whereas negative stimuli should elicit spontaneous avoidance reactions. In line with this idea, Solarz (1960) found that participants were faster at pushing a lever towards them (approach) in response to positive as opposed to negative stimuli, and pushing it away from them (avoidance) for negative as opposed to positive stimuli. Chen and Bargh (1999) expanded on this finding by instructing participants to make either an approach or an avoidance movement as soon as a stimulus appeared on screen. They then calculated participants' response time to a given stimulus depending on whether they had to show an approach or an avoidance movement in response to that stimulus. Their results showed that participants were faster in making an approach movement in response to positive compared to negative stimuli. Conversely, participants were faster in making an avoidance movement in response to negative compared to positive stimuli.

Initial accounts of approach-avoidance tasks interpreted the obtained response patterns as reflecting direct links between particular motor actions and motivational orientations (e.g., contraction of arm extensor = avoidance; contraction of arm flexor muscle

= approach; see Neumann, Förster, & Strack, 2003). However, in contrast to these accounts, more recent findings suggest that congruency effects in AATs depend on the evaluative meaning that is ascribed to a particular motor action in the task. For example, Eder and Rothermund (2008) found that participants were faster in moving a lever backward in response to positive words than negative words when this movement was described as “pull” (positive) and the opposite movement as “push” (negative). In contrast, participants were faster in moving a lever backward in response to negative words than positive words when this movement was described as “downward” (negative) and the opposite movement as “upward” (positive). Corresponding patterns emerged for forward movements. These results suggest that the labels used to describe particular motor actions in AATs are essential for accurate interpretations of their measurement outcomes. Although some versions of AATs have shown satisfactory estimates of internal consistency, their reliability varies considerably depending on the variant that is used (Krieglmeyer & Deutsch, 2010).

Other Instruments

Although the instruments reviewed above are the most popular examples in the current list of available measures, there are several other instruments with unique features that make them better suited for particular research questions. Although a detailed description of these instruments goes beyond the scope of this chapter, we briefly list them for the sake of comprehensiveness. For example, the action interference paradigm (AIP; Banse, Gawronski, Rebetez, Gutt, & Morton, 2010) has been developed for research with very young children who may not be able to follow the complex instructions of other tasks. The implicit relational assessment procedure (IRAP; Barnes-Holmes, Barnes-Holmes, Stewart, & Boles, 2010) and the relational responding task (RRP; De Houwer, Heider, Spruyt, Roets, & Hughes, 2015) have been designed to measure attitudes involving more complex relations between stimuli (e.g., belief that a pharmaceutical product causes or prevents a negative health condition) rather than mere associations between an attitude object and evaluative concepts (e.g., association between the pharmaceutical product and a negative health condition). Other instruments have targeted various methodological limitations of existing measures (e.g., blocked structure, relative measurement, low reliability), including the evaluative movement assessment (EMA; Brendl, Markman, & Messner, 2005), the implicit association procedure (IAP; Schnabel, Banse, & Asendorpf, 2006), and the sorting paired features task (SPFT; Bar-Anan, Nosek, & Vianello, 2009).

Interpretations of the Implicit-Explicit Dualism

The implicit-explicit dualism is not only one of the most common distinctions in the attitude literature, but it is also one of the most confusing dualities because different researchers use it in different ways. Whereas some researchers use the implicit-explicit dualism to refer to two distinct kinds of attitudes (e.g., Greenwald & Banaji, 1995; Wilson, Lindsey, & Schooler, 2000), others use it to describe different types of measurement instruments (e.g., Fazio, 2007; Petty, Fazio, & Briñol, 2009). Yet, other researchers use the dualism to describe the processes by which attitudes influence responses on a given measure (e.g., De Houwer, Teige-Mocigemba, Spruyt, & Moors, 2009); still others use it to refer to two distinct kinds of evaluative responses (e.g., Gawronski & Bodenhausen, 2011). In the current chapter, we use the terms *implicit* and *explicit* in line with this last interpretation. For the sake of conceptual clarity, we first review other interpretations of the implicit-explicit dualism and then explain the theoretical basis for one adopted in the current chapter.

Implicit vs. Explicit Attitudes

One of the most common interpretations of the implicit-explicit dualism uses the distinction to refer to two distinct kinds of attitudinal representations. The central assumption underlying this interpretation is that people can have different attitudes toward the same object stored in memory, one implicit and the other explicit. The most prominent example is Greenwald and Banaji’s (1995) conceptualization of implicit attitudes as “introspectively unidentified (or inaccurately identified) traces of past experience that mediate favorable or unfavorable feeling, thought, or action” (p. 8). Based on this conceptualization, it is often assumed that self-report measures reflect conscious attitudes that are introspectively accessible, whereas performance-based measures reflect unconscious attitudes to which people have no introspective access (e.g., Bosson, Swann, & Pennebaker, 2000; Cunningham, Nezlek, & Banaji, 2004; Rudman, Greenwald, Mellott, & Schwartz, 1999).

Although it seems possible that people have unconscious attitudes that differ from their conscious attitudes, any claims about systematic relations between measurement instruments and conscious awareness are theoretical hypotheses that require empirical evidence. The most common piece of evidence cited in support of the unconsciousness claim is that performance-based measures tend to show rather low correlations with self-report measures (for meta-analyses, see Cameron, Brown-Iannuzzi, & Payne, 2012, Hofmann, Gawronski, Gschwendner, Le, & Schmitt, 2005). The central assumption underlying this argument is that lack of introspective access to implicit attitudes makes it impossible to report these attitudes on a self-report measure, which should lead to low correlations between self-report measures and performance-based measures.

To be sure, correlations between the two kinds of measures can be expected to be low if performance-based measures tap into unconscious attitudes. However, as we will explain in more detail below, correlations between the two kinds of measures can be low for various reasons that have nothing to do with lack of introspective access. More seriously, the available evidence suggests that people are fully aware of the attitude that is captured by performance-based measures, and often report a different attitude on self-report measures for variety of other reasons (for a review, see Gawronski, Hofmann, & Wilbur, 2006).

The strongest evidence that people are fully aware of their attitudes captured by performance-based measures comes from research by Hahn, Judd, Hirsh, and Blair (2014). In a series of studies, participants were asked to predict their scores on multiple IATs capturing attitudes toward different social groups and then completed the same IATs. Counter to the assumption that the IAT provides a window into unconscious attitudes, participants were able to predict the pattern of their IAT scores with a high degree of accuracy. Moreover, predicted and actual IAT scores were highly correlated although self-reported evaluations showed the same low correlations with IAT scores that are typically observed in this area (see Cameron et al., 2012; Hofmann et al., 2005). These findings pose a challenge to the claim that people have no introspective access to the attitudes captured by performance-based measures, and thus to the common interpretation of the two kinds of measures in terms of conscious versus unconscious attitudes.

Another interpretation that proposes two distinct kinds of attitudes is based on the idea that old attitudes may not be erased from memory when people change their attitudes in response to new information. In line with this idea, Wilson et al.'s (2000) dual-attitudes model assumes that performance-based measures capture highly-overlearned old attitudes that are activated automatically in response to an attitude object. In contrast, self-report measures are assumed to reflect more recently acquired attitudes that require cognitive effort to be retrieved from memory (assuming that participants engage in the effortful process of retrieving their new attitude from memory). Adopting the implicit-explicit dualism, Wilson et al. use the term *implicit attitude* to refer to highly-overlearned old attitudes captured by performance-based measures and the term *explicit attitude* to refer to more recently acquired attitudes captured by self-report measures.

Although there is evidence that attitude change can sometimes be limited to self-report measures without generalizing to performance-based measures (e.g., Gawronski & Strack, 2004; Gregg, Seibt, & Banaji, 2006), a large number of studies have shown the opposite pattern. In these studies, experimental manipulations aimed to induce attitude change effectively influenced participants' responses on performance-based measures without affecting their responses on self-report measures (e.g., Gawronski & LeBel, 2008; Gibson, 2008; Grumm, Nestler, & von Collani, 2009; Olson & Fazio, 2006). These findings pose a challenge to interpretations of the two kinds of measures in terms of old "implicit" versus new "explicit" attitudes.

Implicit vs. Explicit Measurement Instruments

Other researchers use the implicit-explicit dualism to describe different types of measurement approaches to assess the same underlying attitude rather than conceptualizing the two kinds of measures in terms of two distinct attitudes. According to Fazio (2007), the primary difference between self-report and performance-based measures is that self-reported evaluations can be influenced by various processes over and above the to-be-measured attitude, whereas performance-based measures reduce the impact of such non-attitudinal processes. For example, in the domain of racial prejudice, White participants may be motivated to report a more favorable attitude toward African Americans on a self-report measure. In this case, participants may show more favorable racial attitudes on self-report measures compared to performance-based measures, the latter of which are assumed to provide a more accurate reflection of participants' real attitudes (e.g., Fazio et al., 1995). In terms of this conceptualization, the implicit-explicit dualism refers to two types of measurement instruments rather than two distinct types of underlying attitudes. Whereas self-report measures are described as *explicit measures*, performance-based measures are described as *implicit measures*.

Expanding on this interpretation of the implicit-explicit dualism, implicit measures have been characterized by the feature that participants are unaware of what the measure is assessing, whereas explicit measures are characterized by the feature that participants are fully aware that the measure is assessing their attitudes (e.g., Petty et al., 2009). Although this conceptualization provides an accurate characterization of various unobtrusive measures in the history of attitude research (see Webb, Campbell, Schwartz, & Sechrist, 1966), it seems less suitable to characterize the new type of indirect measures that are based on objective performance indicators. For example, in the race IAT, most participants are fully aware that the measure aims to assess their racial attitudes. Similarly, in the EPT, participants may be aware that the measure aims to assess their attitudes toward the prime stimuli, unless the primes are presented subliminally. More seriously, some studies using the AMP explicitly informed participants about what the measure is supposed to assess, and how the measure captures the to-be-assessed construct. Yet, this information had no effect on the measurement outcome (Payne et al., 2005). Based on these findings, it seems problematic to interpret the implicit-explicit dualism in terms of participants' awareness of what is assessed by a given measure.

Implicit vs. Explicit Measurement Outcomes

To address the problems of interpretations in terms of measurement instruments, De Houwer et al. (2009) proposed an alternative conceptualization in terms of measurement outcomes. According to this conceptualization, the implicit-explicit dualism reflects properties of the processes by which the to-be-measured attitude influences the outcomes on a given measure. Moreover, instead of limiting the interpretation of the implicit-explicit dualism to the question of whether the to-be-measured attitude influences participants' responses outside of awareness, De Houwer et al. suggested a broader interpretation of the dualism that makes it synonymous with the distinction between automatic and controlled processes. According to their conceptualization, the observed outcomes of a given measure can be described as implicit to the extent that the to-be-measured attitude influences measurement outcomes (a) in the absence of an intention to evaluate the attitude object, (b) in the absence of awareness, (c) in the absence of cognitive resources, or (d) despite the intention to counteract this influence (see Bargh, 1994). Conversely, measurement outcomes can be described as explicit to the extent that the to-be-measured attitude influences measurement outcomes only when participants (a) have an intention to evaluate the attitude object, (b) are aware of how their attitude influences responses on the measure, (c) have sufficient cognitive resources, or (d) do not have an intention to counteract this influence. Because any given measure may be characterized by some features of automatic processing and some features of controlled processing (e.g., attitudes may influence measurement outcomes in the absence of an intention to evaluate the attitude object, but participants may be fully aware of the influence of their attitude on measurement outcomes), De Houwer et al. suggested that descriptions of measurement outcomes as implicit should specify in which sense the measurement outcome is assumed to be implicit (i.e., unintentional, unconscious, efficient, uncontrollable).

Although De Houwer et al.'s conceptualization resolves the problems of the aforementioned interpretations, it involves empirical assumptions in the sense that claims about automatic versus controlled effects of attitudes on measurement outcomes require supportive evidence. This issue poses a terminological dilemma, because (a) it prohibits the use of the implicit-explicit dualism in the absence of empirical evidence, and (b) previous labeling practices may be deemed erroneous in light of new evidence regarding the mechanisms by which attitudes influence outcomes on a given measure.

Implicit vs. Explicit Evaluations

One way to overcome the limitations of the reviewed interpretations of the implicit-explicit dualism is to (1) clearly distinguish between behavioral responses and mental constructs and (2) use the dualism in manner that refers to behavioral responses rather than mental constructs (see De Houwer, Gawronski, & Barnes-Holmes, 2013). In line with this idea, the implicit-explicit dualism has been used to describe two distinct kinds of evaluative responses, one being referred to as *implicit evaluations* and the other one as *explicit evaluations* (e.g., Gawronski & Bodenhausen, 2011). According to this conceptualization, a behavioral response can be described as explicit evaluation if the evaluative connotation of the response is explicit in the observed response (e.g., when evaluations are inferred from participants' self-reported liking of Black people). Conversely, a behavioral response can be described as implicit evaluation if the evaluative connotation of the response is implicit in the observed response (e.g., when evaluations of Black people are inferred from participants' reaction times in responding to positive and negative words after being presented with a Black face or from their self-reported liking of a neutral object that is quickly presented after a Black face).

One advantage of this conceptualization is that it allows for a priori classifications of behavioral responses as implicit or explicit evaluations on the basis of objective features of the measured response. In addition, it incorporates earlier calls to clearly distinguish between attitude as an inner psychological tendency and the behavioral expression of an attitude that is reflected overt evaluations (Eagly & Chaiken, 2007). As such, a conceptualization in terms of distinct types of evaluations remains agnostic about their underlying mental processes and representations, which are treated as theoretical questions that have to be answered on the basis of empirical data. For example, the claim that implicit evaluations reflect unconscious attitudes is treated as a theoretical hypothesis that requires empirical tests, and the available evidence suggests that this hypothesis is incorrect (e.g., Hahn et al., 2014). Similarly, the claim that implicit evaluations reflect old attitudes is treated as a theoretical hypothesis that requires empirical tests, and the available evidence suggests that this hypothesis is incorrect (e.g., Gawronski & LeBel, 2008; Gibson, 2008; Grumm et al., 2009; Olson & Fazio, 2006). Finally, claims that implicit evaluations are the result of unintentional, unconscious, efficient, and uncontrollable processes are treated theoretical hypotheses, and these hypotheses may be correct in some regards but not in others (for a review, see De Houwer et al., 2009). Thus, to avoid a conflation between descriptions of behavioral responses and theoretical hypotheses about their underlying mental constructs, we will use the terms *implicit evaluation* and *explicit evaluation* throughout the remainder of this chapter, referring to responses on performance-based measures and self-report measures, respectively.

Implicit-Explicit Relations

A common assumption in attitude research is that measures of implicit evaluation provide valuable information that cannot be gained from measures of explicit evaluations. This idea is prominently reflected in the argument that correlations between implicit and explicit evaluations tend to be rather low overall. Yet, as we noted above, correlations between the two kinds of evaluations may be low for a variety of reasons. Thus, before drawing any conclusions from observed differences between implicit and explicit evaluations, it is essential to understand the conditions under which they do or do not correspond to each other.

Methodological Factors

Several meta-analyses have found average correlations between implicit and explicit evaluations in the range of .20 to .25 (e.g., Cameron et al., 2012; Hofmann et al., 2005). However, there is evidence that these correlations may underestimate their actual correspondence, in that various methodological factors led to attenuated meta-analytic estimates. Many of these factors can be broadly interpreted in terms of the correspondence principle in research on attitude-behavior relations (Ajzen & Fishbein, 1977), in that correlations between implicit and explicit evaluations tend to be higher if they correspond in terms of their dimensionality and content. For example, Hofmann et al. (2005) found that measurement scores reflecting implicit preferences for one group over another tend to show higher correlations to measurement scores of the same explicit preferences compared to non-relative explicit evaluations of one of the two groups. Similarly, measures of implicit evaluation using Black and White faces as stimuli tend to show higher correlations to explicit evaluative judgments of the same faces compared to explicit evaluative judgments of anti-discrimination policies and perceptions of racial discrimination (e.g., Payne, Burkley, & Stokes, 2008).[2] Thus, when their correspondence in terms of the focal attitude objects is taken into account, correlations between implicit and explicit evaluations tend to be much higher compared to the average correlations found in meta-analyses (e.g., Cameron et al., 2012; Hofmann et al., 2005). This issue is important not only for accurate interpretations of correlations between implicit and explicit evaluations. It also has fundamental implications for research on attitude formation and change, in that differential effects on the two kinds of evaluations may be due to differences in their focal attitude object (e.g., change in evaluations of anti-discrimination policies, but no change in evaluations of Black and Whites faces) rather than differences in the type of evaluation (e.g., change in explicit, but not implicit, evaluations).

The MODE Model

Historically, the development of performance-based measures has its origin in the idea that attitudes can be activated automatically and, thereby, influence behavior in the absence of a goal to evaluate the relevant target object. This idea is at the core of Fazio's (1990) motivation and opportunity as determinants (MODE) model, which includes specific assumptions about the relation between implicit and explicit evaluations. A central assumption of the MODE model is that attitudes are represented as object-valence associations in memory (Fazio, 1995, 2007).[3] To the extent that the link between an object and its associated valence is sufficiently strong, encountering the object should automatically activate the valence associated with the object, which is assumed to be the driving force behind the effects of attitudes on performance-based measures (Fazio et al., 1986).

The downstream effects of automatically activated attitudes on self-reported evaluative judgments are further assumed to depend on people's motivation and opportunity to deliberate. To the extent that either the motivation or the opportunity to deliberate is low, evaluative judgments are assumed to reflect the automatically activated attitude. Yet, if both the motivation and the opportunity to deliberate are high, people are assumed to engage in an elaborate analysis of specific attributes of the target object, which in turn provide the basis for self-reported evaluative judgments. Together, these assumptions imply that correlations between implicit and explicit evaluations should be high when either the motivation or the opportunity to engage in deliberate processing is low. In contrast, correlations between implicit and explicit evaluations should be low when both the motivation and the opportunity to engage in deliberate processing are high.

These predictions have been supported by several studies showing that correlations between implicit and explicit evaluations tend to be higher when evaluative judgments are provided under time pressure, which

[2] Evaluative judgments of anti-discrimination policies and perceptions of racial discrimination are central themes in the Modern Racism Scale (McConahay, 1986), which is often used as a self-report measure in research on racial attitudes.

[3] Fazio (1995) uses the term *object-evaluation association*. To avoid potential confusion with the current use of the term *evaluation* to refer to the behavioral expression of an attitude, we use the term *object-valence associations*.

reduces the opportunity to engage in deliberate processing (e.g., Ranganath, Smith, & Nosek, 2008). Conversely, correlations between implicit and explicit evaluations tend to decrease as a function of the time participants take to make an evaluative judgment, which presumably reflects a higher motivation to engage in deliberate processing (e.g., LeBel, 2010). Further evidence for the predictions of the MODE model comes from research on racial attitudes, showing that participants with a high motivation to control prejudiced reactions show lower correlations between implicit and explicit evaluations of racial outgroups compared to participants with a low motivation to control prejudiced reactions (e.g., Dunton & Fazio, 1997; Fazio et al., 1995). In terms of the MODE model, such findings reflect a higher motivation to engage in deliberate processing among those who are motivated to control prejudiced reactions, which should reduce the impact of automatically activated attitudes on self-report evaluative judgments.

The APE Model

Another theory that explains the relation between implicit and explicit evaluations is the associative-propositional evaluation (APE) model (Gawronski & Bodenhausen, 2006a, 2011). According to the APE model, implicit evaluations are the behavioral outcomes of associative processes; explicit evaluations are the behavioral outcomes of propositional processes. *Associative processes* are defined as the activation of mental associations on the basis of feature similarity and spatio-temporal contiguity; *propositional processes* are defined as the validation of the information implied by activated associations. A central assumption of the APE model is that the propositional validation of activated associations involves an assessment of consistency, in that inconsistency requires a reassessment and potential revision of one's beliefs (Gawronski, 2012). From this perspective, the relation between implicit and explicit evaluations should depend on whether the evaluation implied by activated associations is consistent with other information that is considered for a self-reported evaluative judgment. This information may include non-evaluative beliefs about the world or evaluative beliefs about other attitude objects (Jones & Gerard, 1967). To the extent that the evaluation implied by activated associations is consistent with other salient information, it is usually regarded as valid and therefore used as a basis for self-reported evaluative judgments. However, if it is inconsistent with other salient information, people may reject the evaluation implied by activated associations in order to restore cognitive consistency (Gawronski & Strack, 2004).

Although the MODE and the APE model make similar predictions in most cases, the two theories differ in terms of two central assumptions. First, whereas the MODE model assumes that motivation and opportunity are the primary determinants of implicit-explicit relations, the APE model proposes cognitive consistency as the central proximal factor. To illustrate this difference, consider Fazio et al.'s (1995) finding that the relation between implicit and explicit evaluations of racial outgroups is higher for participants with a low motivation to control prejudice compared to participants with a high motivation to control prejudice. From the perspective of the APE model, implicit evaluations reflect spontaneous affective reactions that result from the associations that are activated in response to members of the target group (e.g., negative affective reaction to African Americans resulting from negative associations). These reactions may serve as the basis for self-reported evaluative judgments (e.g., *I dislike African Americans*), unless such a judgment would be inconsistent with other salient information. In the case of racial prejudice, other salient information may include non-evaluative beliefs about racial discrimination (e.g., *African Americans represent a disadvantaged group*) and evaluative beliefs about the expression of negative racial judgments (e.g., *Negative evaluations of disadvantaged groups are wrong*). According to APE model, consistency among these beliefs may be restored by rejecting one's spontaneous affective reaction as a basis for a self-reported evaluative judgment (e.g., *I like African Americans*). Yet, consistency may also be restored by changing one's evaluative beliefs about the expression of negative racial judgments (e.g., *Negative evaluations of disadvantaged groups are okay*) or one's non-evaluative beliefs about racial discrimination (e.g., *African Americans do not represent a disadvantaged group*). These considerations lead to the alternative prediction that high motivation to control prejudice—rooted in negative beliefs about the expression of negative racial judgments—should be insufficient to reduce the relation between implicit and explicit evaluations of racial outgroups when participants maintain cognitive consistency by denying racial discrimination. In this case, a person may report negative feelings towards African Americans and nevertheless maintain their belief that one should not express negativity towards disadvantaged groups, because the person denies that African Americans represent a disadvantaged group (akin to the concept of "modern racism", McConahay, 1982). This prediction has been confirmed in several studies, showing high correlations between implicit and explicit evaluations of stigmatized groups when *either* motivation to control prejudice *or* perceived discrimination is low (Brochu, Gawronski, & Esses, 2011; Gawronski, Peters, Brochu, & Strack, 2008). Correlations between implicit and explicit evaluations were reduced only when *both* motivation to control prejudice *and* perceived discrimination were high. These results support the APE model's hypothesis that cognitive consistency functions as the primary proximal determinant of implicit-explicit relations, whereas motivation and opportunity to deliberate are better understood as distal determinants.

Second, whereas the MODE model assumes that deliberate processing generally reduces the relation between implicit and explicit evaluations, the APE model assumes that such reductions should occur only when the additionally considered information is inconsistent with the evaluation implied by activated associations. To the extent that deliberate processing involves a selective search for information that supports the validity of this evaluation, deliberate processing may in fact increase rather than decrease the relation between implicit and explicit evaluations. This hypothesis is consistent with research showing that selective search for information that is consistent with the evaluation implied by activated associations increases the correlation between implicit and explicit evaluations (e.g., Galdi, Gawronski, Arcuri, & Friese, 2012; see also Peters & Gawronski, 2011a).

Attitude-Behavior Relations

A common question about performance-based measures is whether they predict behavior (for meta-analyses, see Cameron et al., 2012; Greenwald, Poehlman, Uhlmann, & Banaji, 2009; Oswald, Mitchell, Blanton, Jaccard, & Tetlock, 2013). Although this question is perfectly justified, it does not reflect the more nuanced theoretical views that have guided research on the prediction of behavior with performance-based and self-report measures (for reviews, see Friese, Hofmann, & Schmitt, 2008; Perugini, Richetin, & Zogmaister, 2010). Instead of focusing on zero-order relations between implicit evaluations and behavioral criteria, research guided by extant theories of attitude-behavior relations aims to provide answers to the following three questions: (a) What kinds of behaviors do implicit and explicit evaluations predict? (b) Under which conditions do implicit and explicit evaluations predict behavior? (c) For whom do implicit and explicit evaluations predict behavior? (see also Ajzen, Fishbein, Lohmann, & Albarracín, this volume).

The MODE Model

In addition to providing a theoretical framework for understanding the relation between implicit and explicit evaluations, the MODE model includes precise assumptions about their predictive relations to behavior (Fazio, 1990, 2007). According to the MODE model, responses on performance-based and self-report measures do not reflect two distinct types of attitudes (e.g., implicit versus explicit attitudes). Instead, responses on either type of measure are behavioral expressions of the same underlying attitude, conceptualized as object-valence association of varying strength. The central difference between the two kinds of measures is that performance-based measures limit participants' opportunity to engage in deliberate processing. As a result, responses on performance-based measures are relatively independent of participants' motivation to engage in deliberate processing, such that participants are assumed show the same evaluative response regardless of whether their motivation to engage in deliberate processing is high or low. This situation differs for self-report measures, which typically do not constrain participants' opportunity to engage in deliberate processing (unless judgments have to be provided under time pressure; see Ranganath et al., 2008). Hence, responses on self-report measures depend on participants' motivation to engage in deliberate processing, such that participants may report different evaluative judgments depending whether their motivation to engage in deliberate processing is high or low.

These assumptions have important implications for the prediction of behavior with performance-based and self-report measures. According to the MODE model, the predictive relations between either type of measure and behavior depend on the processing conditions imposed by the measurement instrument and the processing conditions of the to-be-predicted behavior. That is, predictive relations should be high to the extent that the processing conditions imposed by the measurement instrument are equivalent to the processing conditions of the to-be-predicted behavior. In contrast, predictive relations should be low to the extent that the processing conditions imposed by the measurement instrument are different from the processing conditions of the to-be-predicted behavior.

These hypotheses provide clear and empirically supported answers to the three central questions in theory-based research on attitude-behavior relations. Regarding what kinds of behaviors are predicted by implicit and explicit evaluations, several studies have found a double-dissociation pattern, such that implicit evaluations outperformed explicit evaluations in the prediction of spontaneous behavior, whereas explicit evaluations outperformed implicit evaluations in the prediction of deliberate behavior (for reviews, see Friese et al., 2008; Perugini et al., 2010). For example, a common finding in research on racial attitudes is that nonverbal behavior in interracial interactions shows stronger relations to implicit compared to explicit evaluations, whereas the content of verbal responses in interracial interactions shows stronger relations to explicit compared to implicit evaluations (e.g., Dovidio, Kawakami, & Gaertner, 2002; Fazio et al., 1995; see also Penner, Albrecht, Dovidio, Manning, & van Ryn, this volume).

Regarding the conditions under which implicit and explicit evaluations predict behavior, several studies have found a moderation pattern, such that implicit evaluations outperformed explicit evaluations in the prediction of a given behavior when the opportunity to engage in deliberate processing is low. In contrast, explicit evaluations outperformed implicit evaluations in the prediction of the same behavior when the opportunity to engage in deliberate processing is high (for reviews, see Friese et al., 2008; Perugini et al., 2010). For example, in a series of studies, Hofmann and colleagues found that the amount of high-calorie foods participants consumed showed a stronger relation with implicit than with explicit evaluations when participants' cognitive resources were taxed while eating the foods. Yet, consumption of high-calorie foods showed a stronger relation with explicit than with implicit evaluations when participants' cognitive resources were not taxed (e.g., Hofmann, Rauch, & Gawronski, 2007; Friese, Hofmann, & Wänke, 2008). Similar findings have been obtained for the prediction of interpersonal behavior in interracial interactions (Hofmann, Gschwendner, Castelli, & Schmitt, 2008).

Finally, regarding whether there are individual differences in the prediction of behavior, several studies have found a moderation pattern. Consistent with the predictions of the MODE model, implicit evaluations have been shown to outperform explicit evaluations in the prediction of behavior for people who do not have the motivation or the cognitive capacity to engage in deliberate processing. In contrast, explicit evaluations have been shown to outperform implicit evaluations in the prediction of behavior for people with high motivation and high cognitive capacity to engage in deliberate processing (for reviews, see Friese et al., 2008; Perugini et al., 2010). For example, Hofmann, Gschwendner, Friese, Wiers, and Schmitt (2008) found that individual differences in working memory capacity (WMC) moderated the predictive relation of implicit and explicit evaluations to a behavioral criterion (e.g., amount of time participants' spent looking at pornographic images). In their study, implicit evaluations outperformed explicit evaluations in the prediction of behavior for people with low WMC, whereas explicit evaluations outperformed implicit evaluations for people with high WMC. Similar patterns have been found for individual differences in thinking styles, such that explicit evaluations are better predictors of behavior for people with a preference for a deliberative thinking style, whereas implicit evaluations are better predictors of behavior for people with a preference for an intuitive thinking style (e.g., Richetin, Perugini, Adjali, & Hurling, 2007; see also Briñol & Petty, this volume).

Although the MODE model puts a strong emphasis on the correspondence between the processing conditions of the measurement instrument and the to-be-predicted behavior, the theory includes a number of additional assumptions that permit a predictive relation of implicit evaluations to behavior even when their processing conditions do not align. A central assumption of the MODE model is that automatically activated attitudes have the potential to influence the immediate perception of a stimulus (Fazio, 1990). In such cases, automatically activated attitudes may influence deliberate behavioral decisions to the extent that these decisions are based on people's immediate perceptions. Several studies suggest that effects of automatically activated attitudes on immediate perceptions tend to be more pronounced for ambiguous stimuli. For example, in a series of studies by Hugenberg and Bodenhausen (2003), participants were presented with short video clips showing changes in the facial expressions of Black and White men. In one study, the facial expressions changed from hostile to friendly; in a second study, the facial expressions changed from friendly to hostile. Participants in the first study were asked to press a key when they saw no hostility in the facial expression anymore; participants in the second study were asked to press a key as soon as they noticed hostility in the facial expression. Results revealed a greater readiness to perceive hostility in ambiguous expressions of Black faces compared to White faces, and the relative size of this bias was positively correlated to participants' implicit preference for Whites over Blacks. There was no relation between biased perceptions of facial expressions and participants' explicit preference for Whites over Blacks. To the extent that such influences on immediate perceptions occur outside of conscious awareness, they likely remain uncorrected even when people have the motivation and the opportunity to engage deliberate processing (see Gawronski, Geschke, & Banse, 2003).

Effects of automatically activated attitudes on the perception of ambiguous information help to explain why implicit evaluations sometimes predict deliberate behavior over and above explicit evaluations (see Perugini et al., 2010). For example, several studies have shown that implicit evaluations can help to predict future voting decisions over and above explicit evaluations (e.g., Galdi, Arcuri, & Gawronski, 2008; Lundberg & Payne, 2014). The central idea underlying this research is that automatically activated attitudes may bias the perception of ambiguous information about the available options (e.g., perception of a candidate's ambiguous performance in a televised debate). Thus, to the extent that a person's voting decision is based on their biased perception of ambiguous information, implicit evaluations may contribute to the prediction of future voting decisions via their effect on the perception of decision-relevant information.

The PAST Model

Although the MODE model includes precise assumptions about the relative superiority of implicit and explicit evaluations in predicting (a) different kinds of behavior, (b) behavior under different processing conditions, and (c) behavior of different individuals, it does not capture multiplicative patterns in which implicit and explicit evaluations interact in the prediction of behavior (see Perugini et al., 2010). Such multiplicative patterns play a central role in the past-attitudes-are-still-there (PAST) model (Petty, Tormala, Briñol, & Jarvis, 2006). Similar to Wilson et al.'s (2000) dual-attitudes model, the PAST model assumes that implicit evaluations reflect old attitudes that have not been erased from memory when explicit evaluations changed in response to counterattitudinal information. Such asymmetric effects of counterattitudinal information are assumed to cause an aversive state of implicit ambivalence, which describes the discrepancy between implicit and explicit evaluations resulting from changes in explicit, but not implicit, evaluations.

According to the PAST model, people are motivated to reduce aversive states of implicit ambivalence by engaging in extensive processing of attitude-relevant information. Consistent with this assumption, several studies have shown interactive effects of implicit and explicit evaluations in predicting enhanced processing of attitude-relevant information. The general pattern obtained in these studies is that participants with large discrepancies between their implicit and explicit evaluations of an attitude object show more elaborate processing of information about the object compared to participants with small discrepancies (e.g., Johnson, Petty, Briñol, & See, in press; Petty et al., 2006; Rydell, McConnell, & Mackie, 2008).

Although the predictions of the PAST model have been confirmed in several studies, the model does not include any assumptions about directional effects of implicit ambivalence. According to the theory, it does not matter whether explicit evaluations are positive and implicit evaluations are negative, or vice versa (see Johnson et al., in press; Petty et al., 2006). Hence, the theory is unable to explain asymmetric interactions in which a given behavior is related to one type of discrepancy but not the other. For example, in research on self-esteem, combinations of high explicit positivity and low implicit positivity toward the self have been shown to predict narcissistic tendencies, favoritism of one's ingroup over outgroups, and dissonance-related attitude change (e.g., Jordan, Spencer, Zanna, Hoshino-Browne, & Correll, 2003). These predictive patterns did not generalize to combinations of low explicit positivity and high implicit positivity toward the self. Without additional assumptions, the PAST model is unable to explain why these behaviors are predicted by one type of discrepancy, but not by the other type. Another challenge for the PAST model is the large body of research showing changes in implicit, but not explicit, evaluations (e.g., Gawronski & LeBel, 2008; Gibson, 2008; Grumm et al., 2009; Olson & Fazio, 2006). As we noted earlier in this chapter, these findings contradict the hypothesis that explicit evaluations reflect recently acquired attitudes, whereas implicit evaluations reflect old attitudes that have been erased from memory.

The Reflective-Impulsive Model

Although Strack and Deutsch's (2004) reflective-impulsive model (RIM) is not an attitude theory in the strict sense of the term, its broader assumptions about the processes underlying human behavior have inspired a considerable body of research on attitude-behavior relations. The RIM argues that human behavior is guided by two systems of information processing: the *reflective system* (RS) and the *impulsive system* (IS). Even though the two systems are assumed to operate in parallel, the IS enjoys priority over the RS, in that the operation of the RS depends on the availability of cognitive resources whereas the operation the IS is claimed to be resource-independent. The RIM further assumes that the IS operates on the basis of simple associative links between elements that are formed and activated according to the principles of similarity and contiguity. Information processing in the RS is assumed to involve propositionally represented relations between elements that are tagged with truth values (i.e., true vs. false). These operating characteristics make the RS capable of performing various operations that cannot be performed by the IS, the most important being the processing of negations and representations of the future. Thus, although activated associations in the IS provide the basis for propositional representations in the RS, the two systems can lead to different behavioral outcomes if processing in the RS involves the negation of activated associations in the IS (see Deutsch, Gawronski, & Strack, 2006) or the discounting of currently available options in the light of future options (see Metcalfe & Mischel, 1999).

In terms of the RIM, implicit evaluations can be understood as the antecedent of impulsive tendencies to approach or avoid an object generated by the IS. Such impulsive tendencies may sometimes conflict with a person's explicit evaluations, which can be understood as reflective judgments of that object generated by the RS. According to the RIM, the critical difference between the two kinds of responses is that reflective judgments can go beyond impulsive tendencies by (a) negating the associations that gave rise to an impulsive tendency and (b) consider future outcomes that leave impulsive tendencies unaffected. As such, the RIM has been particularly influential in research on self-regulatory conflicts between impulsive tendencies and reflective judgments, including research on food consumption and sexual behavior (for a review, see Hofmann, Friese, & Strack, 2009). For example, Friese et al. (2008) found that consumption of high-calorie foods under conditions of limited capacity was positively related to implicit, but not explicit, evaluations of the relevant foods. In contrast, consumption of high-calorie foods under control conditions were positively related to explicit, but not implicit, evaluations. Similarly, Hofmann et al. (2008) showed that the amount of time participants' spent looking at pornographic images was predicted implicit, but not explicit, evaluations of pornography for participants with low WMC. In contrast, for participants with high WMC, looking times were predicted by explicit, but not implicit, evaluations of pornography. Although these findings are broadly consistent with the predictions of the MODE model, the RIM suggests a slightly different interpretation by treating implicit evaluations as a precursor of impulsive tendencies to approach or avoid an object and explicit evaluations as a precursor of reflective action plans that go beyond immediate hedonic tendencies (see Hofmann et al., 2009).

Attitude Formation and Change

Expanding on the unique roles of implicit and explicit evaluations in the prediction of behavior, a substantial body of research has aimed to identify the antecedents of the two kinds of evaluations. This agenda is particularly prominent in research on attitude formation and change, which has shown various dissociations in the antecedents of implicit and explicit evaluations (for a review, see Gawronski & Bodenhausen, 2006a). Whereas some studies demonstrated change on explicit but not implicit evaluations (e.g., Gawronski & Strack, 2004; Gregg, Seibt, & Banaji, 2006), others demonstrated change on implicit but not explicit evaluations (e.g., Gibson, 2008; Olson & Fazio, 2006). Yet, other studies found corresponding effects on explicit and implicit evaluations (e.g., Olson & Fazio, 2001; Whitfield & Jordan, 2009) and some studies found changes in opposite directions on the two kinds of evaluations (e.g., Moran & Bar-Anan, 2013; Rydell, McConnell, Mackie, & Strain, 2006). These inconsistent patterns posed a challenge to earlier theories of attitude formation and change (e.g., Chaiken, Liberman, & Eagly, 1989; Kruglanski & Thompson, 1999; Petty & Cacioppo, 1986), which inspired the development of new theories to explain (and ideally predict) dissociations in the antecedents of implicit and explicit evaluations (e.g., Gawronski & Bodenhausen, 2006a; Petty, Briñol, & DeMarree, 2007).

The APE Model

One example of a theory aimed to explain these dissociations is the associative-propositional evaluation (APE) model (Gawronski & Bodenhausen, 2006a, 2011), which distinguishes between the activation of associations in memory (*associative process*) and the validation of momentarily activated information (*propositional process*). As we explained earlier in this chapter, the APE model assumes that processes of association activation are driven by principles of similarity and contiguity; processes of propositional validation are assumed to be guided by principles of cognitive consistency. The distinction between associative and propositional processes is further linked to implicit and explicit evaluations, such that implicit evaluations are assumed to reflect the outcomes of associative processes, whereas explicit evaluations are assumed to reflect the outcomes of propositional processes. Drawing on several assumptions about mutual interactions between associative and propositional processes, the APE model implies precise predictions regarding the conditions under which a given factor should lead to (a) changes in implicit but not explicit evaluations; (b) changes in explicit but not implicit evaluations; and (c) corresponding changes in explicit and implicit evaluations.

According to the APE model, changes in implicit but not explicit evaluations should occur when (a) a given factor alters the structure of associations in memory, and at the same time (b) these newly created associations are rejected as a basis for self-reported judgments because of their inconsistency with other salient information. Resonating with the idea of associative learning in evaluative conditioning (EC; see De Houwer, Thomas, & Baeyens, 2001), this pattern has been observed when (a) a well-known conditioned stimulus (CS) was repeatedly paired with a positive or negative unconditioned stimulus (US) and (b) participants considered other information about the CS that led them to reject the newly the formed association as a basis for their self-reported judgments about the CS (e.g., Gawronski & LeBel, 2008; Gibson, 2008; Grumm et al., 2009; Karpinski & Hilton, 2001; Olson & Fazio, 2006). However, when participants were encouraged to rely on their spontaneous "gut" feelings toward the CS, implicit and explicit evaluations typically showed corresponding effects, in that both reflected the valence implied by the CS-US pairings (e.g., Gawronski & LeBel, 2008; Grumm et al., 2009). The latter finding is consistent with the APE model's prediction that both implicit and explicit evaluations should show change when (a) a given factor alters the structure of associations in memory, and (b) these newly created associations are accepted as a valid basis for self-reported judgments.

Another prediction of the APE model is that changes in explicit but not implicit evaluations should occur when (a) a given factor influences the perceived validity of an existing association, and at the same time (b) mere negation of validity does not result in the formation of new associations. According to the APE model, this case is most likely when newly acquired information leads to inconsistency within a set of salient beliefs, and the resulting inconsistency is resolved by rejecting activated associations as a basis for self-reported judgments. In line with these assumptions, Gawronski and Strack (2004) found that cognitive dissonance arising from induced compliance (Festinger & Carlsmith, 1959; see also Harmon-Jones, Armstrong, & Olson, this volume) led to changes in explicit but not implicit evaluations (see also Wilson et al., 2000). The same pattern has been observed in paradigms where previously acquired information is discredited as invalid, and participants are asked to mentally reverse the previously presented information. For example, Gregg et al. (2006) presented participants with positive information about a Group A and negative information about another Group B. After the impression formation task, participants were told to mentally reverse this information, such that the positive information was supposed to refer to Group B and the negative information was supposed to refer to Group A. Whereas explicit evaluations showed a full reversal, implicit evaluations continued to reflect the content of the initial information.

A critical aspect in these studies is that the discrediting information involves a simple “negation” of activated associations, which may lead to a rejection of these associations for self-reported judgments. Yet, mere rejection of a given association does not necessarily lead to a deactivation of this association (see Deutsch et al., 2006). In fact, repeated negations may often have ironic effects, in that they strengthen the associative link that is supposed to be undone. For example, rejecting the proposition “old people are bad drivers” as false may have counterintentional effects at the associative level, in that it may strengthen the associative link between *old people* and *bad drivers*. Consistent with this hypothesis, Gawronski, Deutsch, Mbirkou, Seibt, and Strack (2008) found that repeated negation of racially biased associations (e.g., Black-hostile) enhanced rather than reduced racial bias on implicit evaluations (but see Johnson, Kopp, & Petty, in press). A successful reduction occurred only when participants repeatedly affirmed the opposite (e.g., Black-friendly). The latter finding resonates with the APE model’s prediction that both implicit and explicit evaluations should show change when (a) a given factor influences the perceived validity of propositional information, and at the same time (b) affirmation of this information leads to the formation of new associations. Consistent with this prediction, several studies have shown that that newly acquired verbal information about positive or negative characteristics of an object influences both implicit and explicit evaluations of that object (e.g., Cone & Ferguson, 2015; Gawronski & Walther, 2008; Whitfield & Jordan, 2009).

To summarize the different patterns of change that can emerge as a result of interactions between associative and propositional processes, Gawronski and Bodenhausen (2006a) provided a schematic overview that includes the following four cases.

Case 1: A direct effect on associative representations with the newly formed associations being accepted by a propositional validity assessment. This pattern is assumed to lead to corresponding changes in implicit and explicit evaluations, with changes in explicit evaluations being mediated by changes in implicit evaluations (see Figure 1, upper left panel).

Case 2: A direct effect on associative representations with the newly formed associations being rejected by a propositional validity assessment. This pattern is assumed to lead to changes in implicit but not explicit evaluations (see Figure 1, upper right panel).

Case 3: A direct effect on the process of propositional validity assessment that leads to a rejection of existing associations. This pattern is assumed to lead to changes in explicit but not implicit evaluations (see Figure 1, lower left panel).

Case 4: A direct effect on the process of propositional validity assessment that leads to the formation of new associations. This pattern is assumed to lead to corresponding changes on implicit and explicit evaluations, with changes in implicit evaluations being mediated by changes in explicit evaluations (see Figure 1, lower right panel).

Expanding on these cases, Gawronski and Bodenhausen (2006a) also discussed various combinations of the four basic patterns involving multiple factors with different effects. For example, opposite effects on implicit and explicit evaluations have been observed when repeated CS-US pairings suggest an evaluation that is opposite to the one implied by newly acquired propositional information (e.g., Moran & Bar-Anan, 2013; Rydell et al., 2006). In such cases, implicit evaluations reflected the valence implied by the CS-US pairings, whereas explicit evaluations reflected the valence of the newly acquired propositional information.

The Meta-Cognitive Model

The meta-cognitive model (MCM; Petty et al., 2007) is a conceptual extension of the PAST model (Petty et al., 2006), designed to reconcile some inconsistencies between the theory and the available evidence. As we noted earlier in this chapter, the PAST model assumes that implicit evaluations reflect old attitudes that have not been erased from memory after explicit evaluations have changed in response to counterattitudinal information. Yet, in contrast to this assumption, a considerable body of research has shown changes in implicit but not explicit evaluations (e.g., Gawronski & LeBel, 2008; Gibson, 2008; Grumm et al., 2009; Olson & Fazio, 2006). To reconcile this inconsistency, the MCM draws on Fazio’s (1995) assumption that attitudes are represented as object-valence associations in memory. A central assumption of the MCM is that exposure to counterattitudinal information creates a new counterattitudinal association in addition to the pre-existing attitudinal association. Which of these conflicting associations determines implicit and explicit evaluations is claimed to depend on (a) the relative strength of each association and (b) stored validity tags that qualify one of these associations as true and the other one as false (e.g., one validity tag qualifying the new counterattitudinal association as true and another validity tag qualifying the initial attitudinal association as false; see Figure 2).

According to the MCM, implicit evaluations exclusively depend on the relative strength of the conflicting object-valence associations. Because associations involving validity tags are assumed to be weaker compared to object-valence associations, the impact of validity tags should depend on whether people are motivated and able to engage in the effortful process of retrieving validity tags from memory. Thus, stored validity tags should influence only explicit, but not implicit, evaluations. From this perspective, the effectiveness of a given factor in changing implicit evaluations should depend on whether this factor is capable of creating a new counterattitudinal association that is stronger than the pre-existing attitudinal association. To the extent that the new counterattitudinal association is stronger than the pre-existing attitudinal association, implicit evaluations should reflect the newly acquired counterattitudinal information. Yet, if the new counterattitudinal association is weaker than the pre-existing attitudinal association, implicit evaluations should reflect the valence of the initial attitude. Moreover, if the two kinds of associations are equal in strength, implicit evaluations should show a pattern of ambivalence, such that exposure to the attitude object should activate both associations to the same extent (e.g., De Liver, Van der Pligt, & Wigboldus, 2007; Petty et al., 2006).

As for changes in explicit evaluations, the MCM assumes that the effectiveness of a given factor depends on which of the two conflicting associations is tagged as true and which association is tagged as false. If the pre-existing attitudinal association is tagged as true and the newly formed counterattitudinal association is tagged as false, explicit evaluations should reflect the valence of the initial attitude. In contrast, if the pre-existing attitudinal association is tagged as false and the newly formed counterattitudinal association is tagged as true, explicit evaluations should reflect the valence of the newly acquired counterattitudinal information.

By virtue of these assumptions, the MCM is able to explain almost every possible pattern of change in implicit and explicit evaluations (see Figure 2). For example, changes in implicit, but not explicit, evaluations are explained by the formation of a new counterattitudinal association that is (a) stronger than the initial attitudinal association and (b) tagged as false (see Figure 2, upper left panel). Conversely, changes in explicit, but not implicit, evaluations are explained by the formation of a new counterattitudinal association that is (a) weaker than the initial attitudinal association and (b) tagged as true, with the initial attitudinal association being tagged as false (see Figure 2, upper right panel). Moreover, corresponding changes in implicit and explicit evaluations are explained by the formation of a new counterattitudinal association that is (a) stronger than the initial attitudinal association and (b) tagged as true, with the initial attitudinal association being tagged as false (see Figure 2, lower left panel). Finally, there should be no change in either implicit or explicit evaluations when a given factor produces a new counterattitudinal association that is (a) weaker than the initial attitudinal association and (b) tagged as false (see Figure 2, lower right panel).

Importantly, these predictions are based on the assumptions that (a) a person has conflicting evaluative associations with an attitude object and (b) the person is motivated and able to engage in the effortful process of retrieving stored validity tags from memory. If there are no conflicting evaluative associations, retrieval of validity tags is not necessary and implicit and explicit evaluations should directly reflect the existing object-valence association (e.g., Olson & Fazio, 2001). Moreover, if the person is not motivated or unable to engage in the effortful process of retrieving stored validity tags, explicit evaluations should show the pattern predicted for implicit evaluations, such that explicit evaluations should reflect the relative strength of the two conflicting associations irrespective of stored validity tags (e.g., Wilson et al., 2000).

Although these assumptions allow the MCM to explain a broad range of empirical findings, the model does not include specific assumptions about the conditions under which each of the observed patterns should occur (see Gawronski & Bodenhausen, 2006b). As such, the model has strong explanatory, but weak predictive, power. Another limitation of the MCM is that it is unable to explain different mediation patterns when implicit and explicit evaluations show corresponding effects. For example, repeated CS-US pairings have been shown to influence both implicit and explicit evaluations, with changes in explicit evaluations being fully mediated by changes in implicit evaluations (e.g., Whitfield, & Jordan, 2009). Conversely, acquisition of new propositional information has been shown to influence both implicit and explicit evaluations, with changes in implicit evaluations being fully mediated by changes in explicit evaluations (e.g., Whitfield, & Jordan, 2009). These findings are consistent with the predictions of the APE model, but they are not captured by the MCM.

Context Effects

Counter to initial claims that implicit evaluations may be resistant to context effects, a substantial body of research has shown that implicit evaluations of the same object can differ depending on the context in which it is encountered (for a review, see Gawronski & Sritharan, 2010). For example, in one of the first demonstrations of such context effects, Wittenbrink, Judd, and Park (2001) found that a picture of an African American man elicited less favorable implicit evaluations when this person was presented against a graffiti wall than when the same person was presented in the context of a family barbeque. Similarly, Roefs et al. (2006) found that implicit evaluations of high-fat foods were more favorable when these foods were presented in a context emphasizing palatability (i.e., restaurant) than when they were presented in a context emphasizing health (i.e., hospital). To date, research has identified a wide range of contextual factors that influence implicit evaluations, including exposure to liked or disliked exemplars (Dasgupta & Greenwald, 2001), salient categories (e.g., Mitchell, Nosek, & Banaji, 2003), social roles (e.g., Richeson & Ambady, 2003), affiliation motivation (e.g., Sinclair, Lowery, Hardin, & Colangelo, 2005), goal pursuit (e.g., Ferguson & Bargh, 2004), and emotional states (e.g., DeSteno, Dasgupta, Bartlett, & Cajdric, 2004). In fact, the pervasive evidence for context effects on implicit evaluations has led some researchers to conclude that it seems more difficult to find evidence for context-independence than context-dependence (Schwarz, 2007). To account for these findings, attitude researchers have proposed distinct mechanisms that explain context effects on implicit evaluations.

Context-Dependent Categorization

The first account, most prominently represented by Fazio (2007), argues that people have relatively stable object-valence associations stored in memory. To the extent that the associative link between the two is sufficiently strong, the valence that is associated with an object becomes activated automatically upon encountering the object, which in turn influences implicit evaluations of that object. Context effects on implicit evaluations are attributed to the fact that virtually all objects can be categorized in multiple ways. For example, a young African American man may be categorized in terms of his age, race, or gender. Yet, categorization usually occurs in terms of a single dimension instead of all possible categories (Kawakami, Amodio, & Hugenberg, 2017). Hence, contextual cues can influence implicit evaluations when they influence how a given object is categorized (e.g., Mitchell et al., 2003; Olson & Fazio, 2003; but see Gawronski, Cunningham, LeBel, & Deutsch, 2010). In the above example, the young African American man may elicit a more favorable response when he is categorized in terms of his age (activating positive stereotypical associations related to the category of young people) than when he is categorized in terms of his race (activating negative stereotypical associations related to the category African Americans). Thus, to the extent that contextual cues influence which feature is used to categorize an attitude object, it may moderate the associations that are activated upon encountering the object and, thus, implicit evaluations of that object. Such context effects are not limited to orthogonal categories but may also involve the use of hierarchically structured subcategories. For example, the same African American man may elicit a more favorable implicit evaluation when contextual cues promote a categorization of this person as a Black lawyer than when they promote a categorization in terms of the superordinate category African American (e.g., Barden, Maddox, Petty & Brewer, 2004). Thus, from the perspective of this account, context effects on implicit evaluations are explained by the hypotheses that (a) people have relatively stable category representations, (b) any object can be categorized in multiple ways, and (c) contextual cues influence which of the applicable category representations is used to categorize the target object.

Context-Dependent Construction

The second account, most prominently represented by Schwarz (2007), rejects the notion of enduring dispositional tendencies as it is reflected in many definitions of the attitude construct (e.g., Eagly & Chaiken, 1993). Instead, this account argues that both implicit and explicit evaluations are constructed on the spot on the basis of momentarily accessible concepts (see also Schwarz & Lee, this volume). Accessibility of mental concepts is further assumed to depend on specific features of the context. For example, environmental cues may influence the momentary accessibility of positive or negative exemplars of a given category (e.g., the context of a basketball court may activate different exemplars of the category African American than the context of a graffiti wall), which may moderate the quality of evaluative responses to other members of the same category (see Lord & Lepper, 1999). From a constructivist perspective, context effects on implicit evaluations do not represent exceptions to the presumed rule of context-independence. Instead, context-dependence is regarded as the default, with context-independence being the incidental outcome of highly similar contexts that activate the same concepts. Varying levels of context-similarity can also explain different levels of stability over time, in that implicit evaluations may show higher levels of temporal stability when they occur in the same context than when they occur different contexts (e.g., Gschwendner, Hofmann, & Schmitt, 2008).

Context-Dependent Activation

A third explanation adopts a view that is somewhere in-between representational and constructivist accounts. According to this account, people can have a wide range of conflicting evaluative associations with regard to a particular object. Yet, encountering the object usually activates only a subset of these dormant associations, such that implicit evaluations depend on the net valence of the associated concepts that are activated in response to a given object (e.g., Gawronski & Bodenhausen, 2006a, 2011). Drawing on the notion of pattern matching in connectionist models (Smith, 1996), this account further assumes that the particular concepts that are activated in response to an object are constrained by (a) the overall set of input stimuli and (b) the pre-existing structure of associations in memory. Importantly, the overall set of input stimuli is assumed to include not only the relevant target object but also momentarily available context cues. However, whereas constructivist accounts imply a direct activation of mental concepts by contextual cues, the notion of pattern matching implies that context effects on the activation of associated concepts are constrained by the pre-existing structure of dormant associations in memory. For example, the representation of a given person may involve mental associations with both positive and negative experiences and contextual cues may influence which of these associations are activated in response to that person. Thus, contextual cues should influence implicit evaluations of a given object when they promote the activation of associated concepts of different valence (e.g., Ma, Correll, & Wittenbrink, 2016).

Context-Dependent Renewal

Although the three accounts can explain the available evidence for context effects on implicit evaluations, their explanations may be criticized as vague, in that they can explain almost every possible finding in a post-hoc fashion without allowing a derivation of novel predictions. A particularly important limitation in this regard is that the three accounts do not include testable predictions about the conditions under which implicit evaluations should be context-dependent or context-independent. One account that has been proposed to overcome this limitation is based on the notion of context-dependent renewal in animal learning (Gawronski & Cesario, 2013). A common finding in animal learning is that effects of extinction or counterconditioning of a previously conditioned response are often limited to the context in which extinction and counterconditioning occurred, in that the initial conditioned response recurs in the initial conditioning context or a novel context (for a review, see Bouton, 2004). Animal researchers have used the term *ABA renewal* to describe the finding that an initial response that was acquired in Context A reemerges in Context A after a different response was learned in a different Context B (e.g., Bouton & Bolles, 1979; Bouton & Peck, 1989). The term *ABC renewal* is used to describe the finding that an initial response that was acquired in Context A reemerges in a novel Context C after a different response was learned in Context B (e.g., Bouton & Bolles, 1979; Bouton & Brooks, 1993). Finally, the term *AAB renewal* is used to describe the finding that an initial response that was acquired in Context A reemerges in a novel Context B after a different response was learned in Context A (e.g., Bouton & Ricker, 1994; Tamai & Nakajima, 2000). Expanding on these findings, several studies with human participants have found similar patterns of context effects after experimentally induced changes in implicit evaluations (e.g., Gawronski, Rydell, Vervliet, & De Houwer, 2010; Rydell & Gawronski, 2009; Ye, Tong, Chiu & Gawronski, 2017).

To account for these patterns, Gawronski, Rydell, et al. (2010) argued that exposure to expectancy-violating counterattitudinal information enhances attention to the context, which leads to an integration of the context into the representation of the counterattitudinal information (e.g., Gawronski, Ye, Rydell, & De Houwer, 2014). As a result, counterattitudinal information influences implicit evaluations only in the context in which the counterattitudinal information was learned, whereas initial attitudinal information continues to determine implicit evaluations in any other context, including the context of the initial attitudinal information or a novel context in which the attitude object had not been encountered before.

An interesting aspect of context-dependent renewal is that it implies specific predictions regarding the conditions under which implicit evaluations should show evidence for context-dependence and the conditions under which they should show evidence for context-independence. First, if initial attitudinal information about a novel object is acquired in a particular Context A and then challenged by counterattitudinal information in another Context B, comparing implicit evaluations across Contexts A and B should reveal inconsistent responses across the two contexts. Whereas implicit evaluations in Context A should reflect the initial attitudinal information, implicit evaluations in Context B should reflect the counterattitudinal information. Second, if initial attitudinal information about a novel object is acquired in a particular Context A and then challenged by counterattitudinal information in another Context B, comparing implicit evaluations in Context B to implicit evaluations in a novel Context C should also reveal inconsistent responses. Whereas implicit evaluations in Context B should reflect the counterattitudinal information, implicit evaluations in Context C should reflect the initial attitudinal information. Third, if initial attitudinal information about a novel object is acquired in a particular Context A and then challenged by counterattitudinal information in another Context B, implicit evaluations in a novel Context C should be consistent with implicit evaluations in Context A. In this case, implicit evaluations should reflect the initial attitudinal information in both Context A and Context C. Finally, if initial attitudinal information about a novel object is acquired in a particular Context A and then challenged by counterattitudinal information in the same Context A, comparing implicit evaluations in Context A to implicit evaluations in a novel Context B should reveal inconsistent responses. Whereas implicit evaluations in Context A should reflect the counterattitudinal information, implicit evaluations in Context B should reflect the initial attitudinal information. These patterns have been empirically confirmed in several studies on the formation and change of interpersonal attitudes (for reviews, see Gawronski & Cesario, 2013; Gawronski, Rydell, De Houwer, Brannon, Ye, Vervliet, & Hu, in press). Although the notion of context-dependent renewal does not explain the full range of context effects that have been demonstrated in the literature (for a review, see Gawronski & Sritharan, 2010), it generates testable predictions about the conditions under which implicit evaluations should be context-dependent or context-independent.

Controversies, Caveats, and Current Themes

Although performance-based measures have inspired an exponentially growing body of research (Gawronski & Payne, 2010), they have also ignited some heated debates about their conceptual meaning. Current research in this area is further characterized by controversies about the stability of implicit evaluations, their actual usefulness in prediction of behavior, and the nature of their underlying processes and representations. In the final section, we address these debates, discussing (a) the contribution of personal and cultural factors to implicit evaluations, (b) the prediction of behavior by implicit evaluations, (c) their presumed stability and resistance to change, (d) alternative accounts of implicit evaluation that reject the idea of mental associations, and (e) measurement issues arising from the processes underlying performance-based instruments.

Person vs. Culture

A common question about performance-based measures is whether they capture a person's real attitudes or just cultural associations. The former idea resonates with the argument that performance-based measures are less susceptible to strategic control than self-report measures. The latter idea is based on the argument that implicit evaluations may be influenced by incidental aspects of one's environment that do not reflect a person's true beliefs. This concern has been raised about performance-based measures in general (e.g., Arkes & Tetlock, 2004) and particularly about the IAT (e.g., Karpinski & Hilton, 2001; Olson & Fazio, 2004). To evaluate the validity of this concern, we deem it important to distinguish between a philosophical and an empirical aspect of the debate.

The philosophical aspect concerns the question of which type of behavior should be regarded as a reflection of a person's true self. On the one hand, there is the view that a person's true self is revealed when intentional control fails. On the other hand, there is the equally plausible view that a person's true self is reflected in what the person consciously intends to do or say. Whereas the first interpretation equates the true self with uncontrolled behavior, the second interpretation equates the true self with intentionally controlled behavior. To the extent that implicit evaluations reflect responses under conditions of limited control and explicit evaluations reflect intentionally controlled responses, the two philosophical views have conflicting implications about whether implicit or explicit evaluations reflect a person's true self (Gawronski, Peters, & LeBel, 2008). However, the preference for either of the two interpretations is a matter of philosophical views rather than empirical observation. Thus, even though implicit evaluations clearly fall into the category of uncontrolled behavior, any depiction of implicit evaluations as revealing the true self depends on the subjectively preferred conceptualization of the true self, which is a philosophical question, not an empirical one.

The empirical aspect of the debate concerns the questions of whether implicit evaluations are influenced by aspects of one's cultural environment that are not reflected in one's personal beliefs, and if so, whether behavior is more strongly influenced by a person's endorsed beliefs or by associations arising from a person's cultural environment. Both questions can be answered on the basis of research reviewed in this chapter. As for the first question, research on EC suggests that implicit evaluations are highly sensitive to incidental co-occurrences between stimuli in the environment even when explicit evaluations do not show any effect of the observed co-occurrences (e.g., Gawronski & LeBel, 2008; Gibson, 2008; Grumm et al., 2009; Olson & Fazio, 2006). Importantly, whether observed co-occurrences influence explicit evaluations has been shown to depend on (a) the consideration of other information about the target object and (b) the consistency of this information with the evaluation implied by the observed co-occurrences (e.g., Gawronski

& LeBel, 2008; Grumm et al., 2009). From this perspective, the apparent conflict between the two views does not map onto two distinct types of mental associations (e.g., personal associations vs. cultural associations; see Olson & Fazio, 2004). Instead, the debate becomes obsolete because the overt endorsement of mental associations depends on the processes that determine their use for evaluative judgments. Moreover, the reviewed research on the prediction of behavior suggests that mental associations can influence behavior even when they are rejected as a basis for evaluative judgments. Yet, as we noted in the preceding sections, their behavioral impact is moderated by various factors related to the type of behavior (e.g., Fazio et al., 1995), the conditions under which the behavior is performed (e.g., Hofmann et al., 2007), and individual characteristics of the person who is performing the behavior (e.g., Richetin et al., 2007). From this perspective, the presumed boundary between personal and cultural associations becomes rather blurry and difficult to defend at a conceptual level, which undermines the basis for the debate about whether implicit evaluations reflect a person's real attitudes or just cultural associations (for a more detailed discussion, see Gawronski, Peters, & LeBel, 2008).

Prediction of Behavior

Although numerous individual studies support the predictive validity of performance-based measures (for reviews, see Friese et al., 2008; Perugini et al., 2010), some researchers have raised doubts about whether their measurement outcomes are indeed related to meaningful behavior. The most serious challenge in this regard is a meta-analysis on the predictive validity of intergroup IATs by Oswald et al. (2013). The central finding of their meta-analysis is that IATs designed to measure intergroup bias were relatively poor predictors of intergroup judgments and behavior and performed no better than direct self-report measures. Overall, IATs of racial bias showed a meta-analytic correlation of r = .15 with intergroup behavior and IATs of interethnic bias showed a meta-analytic correlation of r = .12, which were similar in size to the correlations obtained for self-report measures.

Although these findings are frequently cited as evidence for methodological flaws of the IAT, it is important to keep in mind that dual-process theories of attitude-behavior relations do not predict high zero-order correlations for aggregated behavioral criteria (e.g., Fazio, 2007; Strack & Deutsch, 2004). Instead, these theories put a strong emphasis on the notion that predictive relations of implicit and explicit evaluations to overt behavior depend on the type of behavior that is predicted (e.g., spontaneous vs. deliberate), the conditions under which the to-be-predicted behavior is performed (e.g., high versus low cognitive capacity), and characteristics of the person who is performing the to-be-predicted behavior (e.g., preference for intuitive versus deliberate thinking styles). Depending on these theoretically derived moderators, behavior should be predicted by either implicit or explicit evaluations. Thus, to the extent that these moderators are not taken into account, aggregate analyses should reveal small positive relationships, as in Oswald et al.'s (2013) meta-analysis. Of course, there is no guarantee that future meta-analyses will support the moderators proposed by dual-processes theories of attitude-behavior relations. Yet, the results of earlier meta-analyses are largely consistent with these theories, showing that implicit evaluations outperform explicit evaluations in the prediction of spontaneous behavior, whereas explicit evaluations outperform implicit evaluations in the prediction of deliberate behavior (Cameron et al., 2012; Greenwald et al., 2009; see also Dovidio, Brigham, Johnson, & Gaertner, 1996).[4]

Another problem is that the notion of measurement correspondence (see Ajzen & Fishbein, 1977) is rarely taken into account in studies on the prediction of behavior with performance-based measures. A notable exception is a study by Amodio and Devine (2006) that used two IATs of implicit race bias—one measuring implicit evaluations and one measuring implicit stereotyping—to predict different aspects of interracial interactions. In line with their hypotheses, Amodio and Devine found that the evaluative IAT, but not the stereotyping IAT, predicted the physical distance participants kept from an African American interaction partner. Conversely, the stereotyping IAT, but not the evaluative IAT, predicted participants' expectations about the performance of the African American interaction partner in an upcoming academic test. To the extent that the correspondence between the constructs measured by the two IATs and the to-be-predicted

[4] Greenwald et al. (2009) questioned the consistency of their findings with the hypotheses of dual-process theories, emphasizing that the predictive validity of the IAT was unaffected by the spontaneous versus deliberate nature of the to-be-predicted behavior, whereas the predictive validity of self-report measures was significantly lower for spontaneous compared to deliberate behavior. However, to the extent that deliberate behavior shows a higher relation to attitudinal predictors compared to spontaneous behavior (e.g., as a result of more reliable measurement of deliberate behavior), the invariance observed for the IAT does not conflict with the hypotheses of dual-process theories. In fact, the pattern of predictive relations obtained by Greenwald et al. (2009) is perfectly consistent with these theories, in that self-report measures outperformed the IAT in the prediction of deliberate behavior, whereas the IAT outperformed self-report measures in the prediction of spontaneous behavior (for similar meta-analytic findings on the predictive validity of sequential priming tasks, see Cameron et al., 2012).

behaviors is ignored and predictive relations are aggregated across measures in meta-analytic reviews (e.g., Oswald et al., 2013), the outcome is a small positive average correlation that fails to capture the nuances of the original, theoretically predicted finding. Similar concerns can be raised about individual studies, many of which have paid little attention to issues of measurement correspondence in the relation between implicit evaluations and to-be-predicted behavior. Thus, if measurement correspondence is taken into account, the observed relations may turn out to be much higher compared to what is suggested by previous meta-analytic reviews (e.g., Cameron et al., 2012; Greenwald et al., 2009; Oswald et al., 2012).

Nevertheless, it is important to note a more fundamental problem that can undermine the utility of performance-based measures in predicting behavior. Counter to the widespread assumption that implicit evaluations are highly stable, longitudinal studies suggest that implicit evaluations are actually *less* stable over time than explicit evaluations. For example, across two longitudinal studies that compared the temporal stability of implicit and explicit evaluations (measured by the IAT and the AMP) in multiple content domains (e.g., racial attitudes, political attitudes) over a period of one to two months, Gawronski, Morrison, Phills, and Galdi (2017) found a weighted average stability of r = .54 for implicit evaluations and a weighted average stability of r = .75 for explicit evaluations (see also Cunningham, Preacher & Banaji, 2001). These results suggest that a person's implicit evaluation today may not necessarily reflect this person's implicit evaluation at a later time. Needless to say, such temporal fluctuations can be detrimental if the goal is to predict a person's future behavior on the basis of this person's implicit evaluation measured at an earlier time. Explicit evaluations fare better in this regard, in that they show significantly higher stability over time compared to implicit evaluations (Gawronski et al., 2017). From this perspective, explicit evaluations may be superior predictors of future behavior regardless of the moderators hypothesized by dual-process theories, simply because implicit evaluations tend to fluctuate to a greater extent than explicit evaluations.

Although the low temporal stability of implicit evaluations can undermine their usefulness in predicting future behavior, it is important to note that this limitation does not necessarily question the validity of performance-based measures in capturing implicit evaluations. In fact, temporal fluctuations in implicit evaluations would not be particularly surprising to the extent that implicit evaluations reflect momentary states of an individual. In line with this idea, Gawronski and Bodenhausen (2006a) argued that implicit evaluations reflect the momentary activation of associations in memory, which depends contextual cues and other situational factors over and above the chronic structure of associations in memory.[5] In contrast, explicit evaluations are assumed reflect the outcome of propositional validation processes, in that they reflect what a person believes to be true or false. Although activated associations are an important determinant of such beliefs, the informational input for propositional inferences is often much more complex. For example, after reading an article about potential positive effects of capital punishment, an opponent of the death penalty may show enhanced activation of favorable associations regarding capital punishment. However, such changes in the activation of associations may not necessarily lead to corresponding changes in overtly expressed beliefs, which may be supported by a much more complex set of propositional information. From this perspective, temporal fluctuations in the momentary activation of associations can still be detrimental for the prediction of future behavior via implicit evaluations, but this limitation does not necessarily question the validity of performance-based measures if implicit evaluations are interpreted as momentary evaluative states.

Prediction of Aggregate Outcomes

Although most research on implicit evaluations has focused on predictive relations at the individual level, an accumulating body of research has used aggregate scores of implicit evaluations at the level of counties, states, or countries to predict outcomes at the macro level (e.g., Hehman, Flake, & Calanchini, in press; Leitner, Hehman, Ayduk, & Mendoza-Denton, 2016; Nosek et al., 2009; Orchard & Price, 2017). For example, Hehman et al. (in press) found that average scores of implicit race bias at the regional level (obtained via a large database of IAT data from the Project Implicit website) predicted disproportionate use of lethal force by police officers against African Americans at the same level of analysis. Similarly, Leitner et al. (2016) found that implicit race bias at the county level predicted county-level death rates from circulatory-related diseases among African

[5] The interactive nature of person-related and situation-related factors in the activation of associations can be illustrated with a finding by Gschwendner et al. (2008). Consistent with many other studies (e.g., Cunningham et al., 2001; Gawronski et al., 2017), the authors found rather low levels of stability in implicit evaluations over a period of two weeks when they used a standard variant of the IAT (r = .29). Yet, temporal stability of implicit evaluations over the same period was significantly higher when the measure included background images to provide additional information about the context of the target stimuli (r = .72). These results are consistent with the assumption that the activation of associations is interactively determined by the chronic structure of associations and the overall set of input stimuli, including the target stimulus and the context in which it is encountered (see Gawronski & Bodenhausen, 2006a).

Americans. Interestingly, many studies that investigated correlates of aggregate levels of implicit bias have found relatively strong relations with aggregate levels of disparities and discrimination, which stands in contrast to the relatively small meta-analytic relations between implicit bias and behavior at the individual level (e.g., Cameron et al., 2012; Greenwald et al., 2009; Oswald et al., 2013).

To account for this paradox, Payne, Vuletich, and Lundberg (in press) proposed that implicit biases reflect the situational accessibility of bias-related concepts rather than early-learned and highly stable attitudes that drive discrimination among individuals who are high in bias. According to Payne et al., implicit biases are highly context-dependent, in that social interactions, media exposure, and various other contextual factors influence the accessibility of concepts from one situation to another. At the same time, people's broader demographic environments tend to be relatively stable, leading to systematic relations between implicit biases and social disparities at the macro level. Together, these assumptions explain why implicit biases at the individual level tend to be relatively unstable over time (because accessibility of bias-related concepts is highly context-sensitive), why zero-order relations between implicit biases and behavior at the individual level tend to be relatively low (because accessibility of bias-related concepts fluctuates at the individual level), and why relations between implicit biases and aggregate outcomes at the macro level can nonetheless be quite substantial (because stable disparities in people's broader demographic environment produce robust differences in the accessibility of bias-related concepts at the macro level). Although Payne et al.'s (in press) theory has been criticized for overemphasizing situational factors and for ignoring the interactive contribution of person-related and situation-related factors in determining the accessibility of mental concepts (e.g., Gawronski & Bodenhausen, in press), its ability to account for a wide range of seemingly conflicting findings makes it an interesting alternative to previous conceptualizations of implicit bias.

Resistance to Change

Another ongoing debate concerns the ease versus difficulty with which implicit evaluations can be changed. Challenging the widespread assumption that changes in implicit evaluations require excessive amounts of counterattitudinal information (e.g., Rydell, McConnell, Strain, Claypool, & Hugenberg, 2007), research by Ferguson and colleagues suggests that implicit evaluations can change rapidly in response to a single piece of novel information, including diagnostic counterattitudinal information (e.g., Cone & Ferguson, 2015) and information that suggests a reinterpretation of earlier information (e.g., Mann & Ferguson, 2015, 2017). These findings stand in contrast to the results of a large-scale study by Lai et al. (2014) that tested the relative effectiveness of 17 interventions to change implicit racial preferences. Overall, Lai et al.'s findings raise doubts about the effectiveness of several commonly accepted interventions, including ones that involve perspective-taking, increased egalitarian values, or induced positive emotions. Consistent with the predictions of dual-process theories (e.g., Gawronski & Bodenhausen, 2006a), the only type of interventions that effectively reduced implicit race bias involved various ways of linking the relevant target groups with positivity or negativity, such as evaluative conditioning or mental simulation of counterattitudinal exemplars. Yet, even these interventions failed to produce changes that remained stable over time, in that implicit evaluations returned to baseline after a delay that ranged from several hours to several days (Lai et al., 2016). The latter findings suggest that the interventions that turned out to be effective in the short-term led to a contextually induced shift in implicit evaluations that dissipated when the intervention was not salient anymore and other salient factors had a more dominant influence. This conclusion is consistent with the findings of longitudinal studies, indicating that implicit evaluations show considerable fluctuations over time (Gawronski et al., 2017).

One potential way to reconcile these findings can be derived from the notion of context-dependent renewal reviewed earlier in this chapter (see Gawronski & Cesario, 2013). Because participants in Lai et al.'s (2016) research completed the studies online in an environment of their choice, it is possible that the observed results reflect the ineffectiveness of the tested interventions in producing changes that generalize across contexts (rather their ineffectiveness in producing changes that remain stable over time). To the extent that participants completed the intervention and the initial measurements in one context and the follow-up measures in a different context, participants' old attitude may continue to influence implicit evaluations when they are measured in a context that is different from the context in which the intervention took place (e.g., Gawronski, Rydell, et al., 2010). In this case, it would be essential to identify interventions that produce changes in implicit evaluations that generalize across contexts even if the observed changes remain stable over time in the context of the intervention. Preliminary evidence in this regard comes from a study by Brannon and Gawronski (2017) who found that diagnostic counterattitudinal information (e.g., Cone & Ferguson, 2015) and information that suggests a reinterpretation of earlier information (e.g., Mann & Ferguson, 2015) led to changes in implicit evaluations that generalized across contexts. Together, these findings suggest that the interventions to change implicit evaluations should be evaluated not only on the basis of their effectiveness in producing immediate change within the same context, but also on the basis of whether the observed changes remain stable over time and generalize across contexts.

Dual-Process vs. Single-Process Accounts

A substantial body of research using performance-based measures has been guided by dual-process theories, assuming that implicit evaluations reflect behavioral outcomes of associative processes whereas explicit evaluations reflect behavioral outcomes of propositional processes (e.g., Gawronski & Bodenhausen, 2006a; Strack & Deutsch, 2004). These theories have been criticized by proponents of single-process theories who argue that both implicit and explicit evaluations are outcomes of a single propositional process (e.g., De Houwer, 2014; Kruglanski & Gigerenzer, 2011). For the topic of the current chapter, De Houwer (2014) has presented the most relevant single-process account, which states that implicit evaluations reflect the automatic formation and activation of mental propositions about the relation between co-occurring stimuli. To support this argument, De Houwer reviewed several studies showing that implicit evaluations (a) can be influenced by verbal instructions and inferences (e.g., De Houwer, 2006; Gast & De Houwer, 2012) and (b) are sensitive to information about how stimuli are related (e.g., Zanon, De Houwer, & Gast, 2012). According to De Houwer (2014), dissociations between implicit and explicit evaluations occur because performance-based measures involve constrained processing conditions during the retrieval of information, not because they tap into two distinct processes or representations. Whereas some information may be activated quickly without requiring much cognitive effort, other information may require time and cognitive resources to be retrieved from memory. Thus, whereas the former type of information should have a strong effect on implicit evaluations, the latter type of information may influence only explicit but not implicit evaluations (for similar arguments, see Cunningham, Zelazo, Packer, & Van Bavel, 2007; Wojnowicz, Ferguson, Dale, & Spivey, 2009).

In evaluating the empirical support for two competing accounts, we deem it important to clarify the specific assumptions about which they disagree (see Gawronski, Brannon, & Bodenhausen, 2017). A central issue in this context is that effects of propositional processes on implicit evaluations are explicitly addressed by dual-process theories that allow for mutual interactions between associative and propositional processes (e.g., Gawronski & Bodenhausen, 2006a; Strack & Deutsch, 2004). A central assumption of these theories is that propositional inferences can function as a distal determinant of implicit evaluations to the extent that they alter the structure or momentary activation of associations in memory (see Figure 1, Case 4). From this perspective, effects of verbal instructions and inferences on implicit evaluations (e.g., De Houwer, 2006; Gast & De Houwer, 2012) are perfectly consistent with dual-process accounts. The two theories lead to different predictions only when verbal instructions conflict with the effects of previously observed co-occurrences, such as repeated CS-US pairings (see Figure 1, Case 3). In this case, dual-process theories predict a dissociation, in that explicit evaluations should reflect the valence implied by the verbal instructions, whereas implicit evaluations should reflect the valence implied by the previously observed co-occurrences. In contrast, single-process propositional theories imply that both implicit and explicit evaluations should reflect the valence implied by the verbal instructions. The available evidence on these conflicting hypotheses supports the predictions of dual-process theories, and conflicts with the predictions of single-process theories (Hu, Gawronski, & Balas, in press).

There are two additional cases for which the two kinds of theories lead to different predictions. First, dual-process theories predict that propositional information about the validity of observed stimulus contingencies should influence only explicit evaluations, whereas implicit evaluations should reflect stimulus contingencies regardless of their perceived validity. This prediction stands in contrast to the one implied by single-process propositional theories, which imply that both explicit and implicit evaluations should reflect the perceived validity of stimulus contingencies. The available evidence on these hypotheses supports the predictions of single-process propositional accounts, and poses a challenge to dual-process accounts. For example, Peters and Gawronski (2011b) found that information about the truth or falsity of evaluative statements about several impression targets influenced both implicit and explicit evaluations when the validity information was available immediately after the encoding of the evaluative statements (see also Moran, & Bar-Anan, & Nosek, 2015). Validity information showed a reduced effect on implicit evaluations only when it was presented after a delay. This finding stands in contrast to the prediction of dual-process accounts that implicit evaluations should reflect the valence of the evaluative statements irrespective of whether they are described as true or false. Yet, it is consistent with the predictions of single-process propositional accounts, suggesting that both implicit and explicit evaluations should reflect the perceived validity of the evaluative statements.

Second, dual-process theories predict that information about contrastive relations between two co-occurring stimuli (e.g., A prevents B; A dislikes B) should influence only explicit evaluations, whereas implicit evaluations should reflect the mere co-occurrence of stimuli irrespective of their relation. In contrast, single-process propositional theories predict that information about contrastive relations between two co-occurring stimuli should have equivalent effects on both implicit and explicit evaluations. The available evidence on these competing hypotheses is rather mixed. Whereas some studies found evidence for the dissociation predicted by dual-process accounts (e.g., Moran & Bar-Anan, 2013) other studies found equivalent effects as predicted by single-process propositional theories (e.g., Gawronski, Walther, & Blank, 2005). Yet, other studies found that information about contrastive relations only reduced, but not reversed, the effects of co-occurrences (e.g., Zanon et al., 2012), suggesting that co-occurrence and relational information may jointly influence implicit evaluations. Based on the conflicting patterns of results, a major challenge for both accounts is to specify the conditions under which information about contrastive relations should reverse or merely attenuate effects of observed co-occurrences and when information about contrastive relations should be ineffective in qualifying effects of observed co-occurrences (e.g., Hu, Gawronski, & Balas, 2017).

Measurement Issues

Another important issue is that responses on performance-based measures do not provide direct reflections of underlying attitudinal processes or representations (De Houwer et al., 2013). That is, multiple qualitatively distinct attitudinal and non-attitudinal processes may contribute to responses on performance-based measures. To disentangle the effects of these distinct processes, theorists have developed formal models that provide quantitative estimates of these processes, including applications of process dissociation (Payne & Bishara, 2009), multinomial modeling (Conrey, Sherman, Gawronski, Hugenberg, & Groom, 2005; Meissner & Rothermund, 2013; Stahl & Degner, 2007), and diffusion modeling (Klauer, Voss, Schmitz, & Teige-Mocigemba, 2007).

An illustrative example is Conrey et al.’s (2005) quad-model, which distinguishes between four qualitatively distinct processes underlying responses on measures based on response interference: activation of an association (*AC*), detection of the correct response required by the task (*D*), success at overcoming associative bias (*OB*), and guessing (*G*). Research using the quad-model has provided more fine-grained insights into the mechanisms underlying previous findings obtained with performance-based measures. Whereas some effects have been shown to reflect genuine effects on attitudinal processes (e.g., reduced racial bias scores resulting from extended training to associate racial groups with positive or negative attributes; see Calanchini, Gonsalkorale, Sherman, & Klauer, 2013), others have been shown to reflect effects on non-attitudinal processes, such as successful versus unsuccessful inhibition of activated associations (e.g., increased racial bias scores after alcohol consumption; see Sherman et al., 2008).

Another important caveat is that different measurement instruments are based on different underlying processes. Although the majority of performance-based measures rely on the notion of response interference, there are a few notable exceptions that rely on other processes (for an overview, see Gawronski, Deutsch, LeBel, & Peters, 2008). Because effects on either type of measure may reflect influences on either attitudinal or measurement-related processes, they may not always show equivalent effects of the same experimental manipulation (e.g., Deutsch & Gawronski, 2009; Gawronski & Bodenhausen, 2005). For example, Gawronski, Cunningham et al. (2010) found that racial bias scores in the EPT were significantly reduced when participants were instructed to attend to an alternative category dimension (e.g., age). Yet, racial bias scores in the AMP were unaffected by attention to alternative categories, showing significant priming effects and meaningful correlations to criterion measures regardless of attention. These results suggest that reliable measurement of implicit evaluations depends on attention to the relevant feature of the primes in the EPT, but not the AMP. This limitation of the EPT can lead to incorrect conclusions, for example, when the effect of attention to alternative categories is interpreted as an effective strategy to control racial bias in implicit evaluations. Thus, to avoid premature inferences about effects on underlying attitudinal processes, it seems prudent to replicate a given finding with an alternative measure that relies on a different underlying mechanism.

Conclusion

Performance-based measures are often claimed to overcome the limitations of self-report measures in assessing attitudes that people might be unwilling or unable to report. Counter to these assumptions, the available evidence suggests that (a) implicit evaluations do not reflect unconscious attitudes, and (b) the relation between implicit and explicit evaluations cannot be boiled down to self-presentational distortions in self-report measures. Nevertheless, the exponentially growing body of research on implicit and explicit evaluation attests to the value of performance-based measures in providing deeper insights into the processes underlying evaluative judgments, the processes by which attitudes influence behavior, and the processes underlying attitude formation and change. Although there are still some open questions and unresolved controversies in research using performance-based measures, the insights they provided are so fundamental that the implicit-explicit dualism has arguably become one of the most central distinctions in attitude research.

References

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

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.

Arkes, H. R., & Tetlock, P. E. (2004). Attributions of implicit prejudice, or “Would Jesse Jackson ‘fail’ the Implicit Association Test? *Psychological Inquiry, 15*, 257-278.

Banaji, M. R., & Hardin, C. D. (1996). Automatic stereotyping. *Psychological Science, 7,* 136-141.

Banse, R., Gawronski, B., Rebetez, C., Gutt, H., & Morton, J. B. (2010). The development of spontaneous gender stereotyping in childhood: Relations to stereotype knowledge and stereotype flexibility. *Developmental Science*, *13*, 298-306.

Bar-Anan, Y., & Nosek, B. A. (2012). Reporting intentional rating of the primes predicts priming effects in the Affective Misattribution Procedure. *Personality and Social Psychology Bulletin, 38,* 1194-1208.

Bar-Anan, Y., Nosek, B. A., & Vianello, M. (2009). The sorting paired features task: A measure of association strengths. *Experimental Psychology*, *56*, 329-343.

Barden, J., Maddux, W. W., Petty, R. E., & Brewer, M. B. (2004). Contextual moderation of racial bias: The impact of social roles on controlled and automatically activated attitudes. *Journal of Personality and Social Psychology, 87*, 5-22.

Bargh, J. A. (1994). The four horsemen of automaticity: Awareness, intention, efficiency, and control in social cognition. In R. S. Wyer & T. K. Srull (Eds.), *Handbook of social cognition* (pp. 1-40). Hillsdale, NJ: Erlbaum.

Barnes-Holmes, D., Barnes-Holmes, Y., Stewart, I., & Boles, S. (2010). A sketch of the Implicit Relational Assessment Procedure (IRAP) and the Relational Elaboration and Coherence (REC) model. *The Psychological Record*, *60*, 527-542.

Blanton, H., Jaccard, J., Gonzales, P. M., & Christie, C. (2006). Decoding the Implicit Association Test: Implications for criterion prediction. *Journal of Experimental Social Psychology, 42,* 192-212.

Bosson, J. K., Swann, W. B., & Pennebaker, J. W. (2000). Stalking the perfect measure of implicit self-esteem: The blind men and the elephant revisited? *Journal of Personality and Social Psychology, 79,* 631-643.

Bouton, M. E. (2004). Context and behavioral processes in extinction. *Learning and Memory, 11,* 485-494.

Bouton, M. E., & Bolles, R. C. (1979). Contextual control of the extinction of conditioned fear. *Learning and Motivation, 10,* 445-466.

Bouton, M. E., & Brooks, D. C. (1993). Time and context effects on performance in a Pavlovian discrimination reversal. *Journal of Experimental Psychology: Animal Behavior Processes, 19,* 165-179.

Bouton, M. E., & Peck, C. A. (1989). Context effects on conditioning, extinction, and reinstatement in an appetitive conditioning preparation. *Animal Learning and Behavior, 17,* 188-198.

Bouton, M. E., & Ricker, S. T. (1994). Renewal of extinguished responding in a second context. *Animal Learning and Behavior*, *22*, 317-324.

Brannon, S. M., & Gawronski, B. (2017). A second chance for first impressions? Exploring the context (in)dependent updating of implicit evaluations. *Social Psychological and Personality Science, 8,* 275-283.

Brendl, C. M., Markman, A. B., & Messner, C. (2005). Indirectly measuring evaluations of several attitude objects in relation to a neutral reference point. *Journal of Experimental Social Psychology*, *41*, 346-368.

Brochu, P. M., Gawronski, B., & Esses, V. M. (2011). The integrative prejudice framework and different forms of weight prejudice: An analysis and expansion. *Group Processes and Intergroup Relations, 14,* 429-444.

Calanchini, J., Gonsalkorale, K., Sherman, J. W., & Klauer, K. C. (2013). Counter-prejudicial training reduces activation of biased associations and enhances response monitoring. *European Journal of Social Psycholgy*, *43*, 321-325.

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

Chaiken, S., Liberman, A., & Eagly, A. H. (1989). Heuristic and systematic processing within and beyond the persuasion context. In J. S. Uleman & J. A. Bargh (Eds.), *Unintended thought* (pp. 212-252). New York: Guilford Press.

Chen, M., & Bargh, J. A. (1999). Consequences of automatic evaluation: Immediate behavioral predispositions to approach or avoid the stimulus. *Personality and Social Psychology Bulletin, 25,* 215-224.

Cone, J., & Ferguson, M. J. (2015). He did *what*? The role of diagnosticity in revising implicit evaluations. *Journal of Personality and Social Psychology, 108.* 37-57.

Conrey, F. R., Sherman, J. W., Gawronski, B., Hugenberg, K., & Groom, C. (2005). Separating multiple processes in implicit social cognition: The Quad-Model of implicit task performance. *Journal of Personality and Social Psychology, 89,* 469-487.

Crowne, D. P., & Marlowe, D. (1960). A new scale of social desirability independent of psychopathology. *Journal of Consulting Psychology, 24,* 349-354.

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., Nezlek, J. B., & Banaji, M. R. (2004). Implicit and explicit ethnocentrism: Revisiting the ideologies of prejudice. *Personality and Social Psychology Bulletin, 30,* 1332-1346.

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.

Dasgupta, N., & Greenwald, A. G. (2001). On the malleability of automatic attitudes: Combating automatic prejudice with images of admired and disliked individuals. *Journal of Personality and Social Psychology, 81,* 800-814.

De Houwer, J. (2003). The extrinsic affective Simon task. *Experimental Psychology, 50,* 77-85.

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. (2014). A propositional model of implicit evaluation. *Social and Personality Psychology Compass, 8,* 342-353.

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 Houwer, J., & De Bruycker, E. (2007a). The Implicit Association Test outperforms the Extrinsic Affective Simon Task as an implicit measure of interindividual differences in attitudes. *British Journal of Social Psychology, 46,* 401-421.

De Houwer, J., & De Bruycker, E. (2007b). The identification-EAST as a valid measure of implicit attitudes toward alcohol-related stimuli. *Journal of Behavior Therapy and Experimental Psychiatry, 38,* 133-143.

De Houwer, J., Heider, N., Spruyt, A., Roets, A., & Hughes, S. (2015). The relational responding task: Toward a new implicit measure of beliefs. *Frontiers in Psychology, 6:319.*

De Houwer, J., Teige-Mocigemba, S., Spruyt, A., & Moors, A. (2009). Implicit measures: A normative analysis and review. *Psychological Bulletin, 135,* 347-368.

De Houwer, J., Thomas, S., & Baeyens, F. (2001). Associative learning of likes and dislikes: A review of 25 years of research on human evaluative conditioning. *Psychological Bulletin, 127,* 853-869.

De Liver, Y., Van der Pligt, J., & Wigboldus, D. (2007). Positive and negative associations underlying ambivalent attitudes. *Journal of Experimental Social Psychology, 43,* 319-326.

DeSteno, D. A., Dasgupta, N., Bartlett, M. Y., & Cajdric, A. (2004). Prejudice from thin air: The effect of emotion on automatic intergroup attitudes. *Psychological Science, 15,* 319-324.

Deutsch, R., & Gawronski, B. (2009). When the method makes a difference: Antagonistic effects on "automatic evaluations" as a function of task characteristics of the measure. *Journal of Experimental Social Psychology, 45,* 101-114.

Deutsch, R., Gawronski, B., & Strack, F. (2006). At the boundaries of automaticity: Negation as reflective operation. *Journal of Personality and Social Psychology, 91*, 385-405.

Dovidio, J. F., Brigham, J., Johnson, B. T., & Gaertner, S. L. (1996). Stereotyping, prejudice, and discrimination: Another look. In C. N. Macrae, C. Stangor, & M. Hewstone (Eds.), *Stereotypes and stereotyping* (pp. 276–319). New York: Guilford.

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.

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. (1993). *The psychology of attitudes*. Fort Worth, TX: Harcourt Brace Jovanovich.

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

Eder, A. B., & Rothermund, K. (2008). When do motor behaviors (mis)match affective stimuli? An evaluative coding view of approach and avoidance reactions. *Journal of Experimental Psychology: General*, *137*, 262.

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. (1995). Attitudes as object-evaluation associations: Determinants, consequences, and correlates of attitude accessibility. In R. E. Petty, & J. A. Krosnick (Eds.), *Attitude strength: Antecedents* *and consequences* (pp. 247-282). Hillsdale, NJ: Erlbaum.

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., 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.

Ferguson, M. J., & Bargh, J. A. (2004). Liking is for doing: The effects of goal pursuit on automatic evaluation. *Journal of Personality and Social Psychology, 87,* 557-572.

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

Friese, M., Hofmann, W., & Schmitt, M. (2008). When and why do implicit measures predict behavior? Empirical evidence for the moderating role of opportunity, motivation, and process reliance. *European Review of Social Psychology, 19,* 285-338.

Friese, M., Hofmann, W., & Wänke, M. (2008). When impulses take over: Moderated predictive validity of explicit and implicit attitude measures in predicting food choice and consumption behavior. *British Journal of Social Psychology, 47,* 397-419.

Galdi, S., Arcuri, L., & Gawronski, B. (2008). Automatic mental associations predict future choices of undecided decision makers. *Science, 321,* 1100-1102.

Galdi, S., Gawronski, B., Arcuri, L., & Friese, M. (2012). Selective exposure in decided and undecided individuals: Differential relations to automatic associations and conscious beliefs. *Personality and Social Psychology Bulletin, 38,* 559-569.

Gast, A., & De Houwer, J. (2012). Evaluative conditioning without directly experienced pairings of the conditioned and the unconditioned stimuli. *The Quarterly Journal of Experimental Psychology, 65,* 1657-1674.

Gawronski, B. (2012). Back to the future of dissonance theory: Cognitive consistency as a core motive. *Social Cognition, 30,* 652-668.

Gawronski, B., & Bodenhausen, G. V. (2005). Accessibility effects on implicit social cognition: The role of knowledge activation and retrieval experiences. *Journal of Personality and Social Psychology, 89,* 672-685.

Gawronski, B., & Bodenhausen, G. V. (2006a). 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. (2006b). Associative and propositional processes in evaluation: Conceptual, empirical, and meta-theoretical issues. Reply to Albarracín, Hart, and McCulloch (2006), Kruglanski and Dechesne (2006), and Petty and Briñol (2006). *Psychological Bulletin, 132,* 745-750.

Gawronski, B., & Bodenhausen, G. V. (2011). The associative-propositional evaluation model: Theory, evidence, and open questions. *Advances in Experimental Social Psychology, 44,* 59-127.

Gawronski, B., & Bodenhausen, G. V. (in press). Beyond persons and situations: An interactionist approach to understanding implicit bias. *Psychological Inquiry*.

Gawronski, B., Brannon, S. M., & Bodenhausen, G. V. (2017). The associative-propositional duality in the representation, formation, and expression of attitudes. In R. Deutsch, B. Gawronski, & W. Hofmann (Eds.), *Reflective and impulsive determinants of human behavior* (pp. 103-118). New York: Psychology Press.

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

Gawronski, B., Cunningham, W. A., LeBel, E. P., & Deutsch, R. (2010). Attentional influences on affective priming: Does categorization influence spontaneous evaluations of multiply categorizable objects? *Cognition and Emotion, 24,* 1008-1025.

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., Deutsch, R., LeBel, E. P., & Peters, K. R. (2008). Response interference as a mechanism underlying implicit measures: Some traps and gaps in the assessment of mental associations with experimental paradigms. *European Journal of Psychological Assessment, 24,* 218-225.

Gawronski, B., Deutsch, R., Mbirkou, S., Seibt, B., & Strack, F. (2008). When "just say no" is not enough: Affirmation versus negation training and the reduction of automatic stereotype activation. *Journal of Experimental Social Psychology, 44,* 370-377.

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., Hofmann, W., & Wilbur, C. J. (2006). Are “implicit“ attitudes unconscious? *Consciousness and Cognition, 15,* 485-499.

Gawronski, B., & LeBel, E. P. (2008). Understanding patterns of attitude change: When implicit measures show change, but explicit measures do not. *Journal of Experimental Social Psychology, 44,* 1355-1361.

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., Morrison, M., Phills, C. E., & Galdi, S. (2017). Temporal stability of implicit and explicit measures: A longitudinal analysis. *Personality and Social Psychology Bulletin, 43,* 300-312.

Gawronski, B., & Payne, B. K. (Eds.). (2010). *Handbook of implicit social cognition: Measurement, theory, and applications*. New York: Guilford Press.

Gawronski, B., Peters, K. R., Brochu, P. M., & Strack, F. (2008). Understanding the relations between different forms of racial prejudice: A cognitive consistency perspective. *Personality and Social Psychology Bulletin, 34,* 648-665.

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

Gawronski, B., Rydell, R. J., Vervliet, B., & De Houwer, J. (2010). Generalization versus contextualization in automatic evaluation. *Journal of Experimental Psychology: General, 139,* 683-701.

Gawronski, B., & Sritharan, R. (2010). Formation, change, and contextualization of mental associations: Determinants and principles of variations in implicit measures. In B. Gawronski, & B. K. Payne (Eds.), *Handbook of implicit social cognition: Measurement, theory, and applications* (pp. 216-240). New York: Guilford Press.

Gawronski, B., & Strack, F. (2004). On the propositional nature of cognitive consistency: Dissonance changes explicit, but not implicit attitudes. *Journal of Experimental Social Psychology, 40,* 535-542.

Gawronski, B., & Walther, E. (2008). The TAR effect: When the ones who dislike become the ones who are disliked. *Personality and Social Psychology Bulletin, 34,* 1276-1289.

Gawronski, B., Walther, E., & Blank, H. (2005). Cognitive consistency and the formation of interpersonal attitudes: Cognitive balance affects the encoding of social information. *Journal of Experimental Social Psychology, 41,* 618-626.

Gawronski, B., & Ye, Y. (2015). Prevention of intention invention in the affect misattribution procedure. *Social Psychological and Personality Science, 6,* 101-108.

Gawronski, B., Ye, Y., Rydell, R. J., & De Houwer, J. (2014). Formation, representation, and activation of contextualized attitudes. *Journal of Experimental Social Psychology, 54,* 188-203.

Gibson, B. (2008). Can evaluative conditioning change attitudes toward mature brands? New evidence from the Implicit Association Test. *Journal of Consumer Research, 35,* 178-188.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics.* New York: Wiley.

Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review, 102,* 4-27.

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.

Greenwald, A. G., Poehlman, T. A., Uhlmann, E., & Banaji, M. R. (2009). Understanding and using the Implicit Association Test: III. Meta-analysis of predictive validity. *Journal of Personality and Social Psychology, 97,* 17-41.

Gregg, A. P., Seibt, B., & Banaji, M. R. (2006). Easier done than undone: Asymmetry in the malleability of implicit preferences. *Journal of Personality and Social Psychology, 90,* 1-20.

Grumm, M., Nestler, S., & von Collani, G. (2009). Changing explicit and implicit attitudes: The case of self-esteem. *Journal of Experimental Social Psychology, 45,* 327-335.

Gschwendner, T., Hofmann, W., & Schmitt, M. (2008). Differential stability: The effects of acute and chronic construct accessibility on the temporal stability of the Implicit Association Test. *Journal of Individual Differences, 29,* 70-79.

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

Hehman, E., Flake, J. K., & Calanchini, J. (in press). Disproportionate use of lethal force in policing is associated with regional racial biases of residents. *Social Psychological and Personality Science*.

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.

Hofmann, W., Gschwendner, T., Castelli, L., & Schmitt, M. (2008). Implicit and explicit attitudes and interracial interaction: The moderating role of situationally available control resources. *Group Processes and Intergroup Relations, 11,* 69-87.

Hofmann, W., Gschwendner, T., Friese, M., Wiers, R., & Schmitt, M. (2008). Working memory capacity and self-regulatory behavior: Towards and individual differences perspective on behavior determination by automatic versus controlled processes. *Journal of Personality and Social Psychology, 95,* 962-977.

Hofmann, W., Friese, M., & Strack, F. (2009). Impulse and self-control from a dual-systems perspective. *Perspectives on Psychological Science, 4,* 162-176.

Hofmann, W., Rauch, W., & Gawronski, B. (2007). And deplete us not into temptation: Automatic attitudes, dietary restraint, and self-regulatory resources as determinants of eating behavior. *Journal of Experimental Social Psychology, 43,* 497-504.

Hu, X., Gawronski, B., & Balas, R. (2017). Propositional versus dual-process accounts of evaluative conditioning: I. The effects of co-occurrence and relational information on implicit and explicit evaluations. *Personality and Social Psychology Bulletin, 43,* 17-32.

Hu, X., Gawronski, B., & Balas, R. (in press). Propositional versus dual-process accounts of evaluative conditioning: II. The effectiveness of counter-conditioning and counter-instructions in changing implicit and explicit evaluations. *Social Psychological and Personality Science*.

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

Johnson, I. R., Kopp, B. M., & Petty, R. E. (in press). Just say no! (and mean it): Meaningful negation as a tool to modify automatic racial attitudes. *Group Processes and Intergroup Relations*.

Johnson, I. R., Petty, R. E., Briñol, P., & See. Y. M. (in press). Persuasive message scrutiny as a function of implicit-explicit discrepancies in racial attitudes. *Journal of Experimental Social Psychology*.

Jones, E. E., & Gerard, H. B. (1967). *Foundations of social psychology*. New York: Wiley.

Jordan, C. H., Spencer, S. J., Zanna, M. P., Hoshino-Browne, E., & Correll, J. (2003). Secure and defensive high self-esteem. *Journal of Personality and Social Psychology 85,* 969-978.

Karpinski, A., & Hilton, J. L. (2001). Attitudes and the Implicit Association Test. *Journal of Personality and Social Psychology, 81,* 774-788.

Karpinski, A., & Steinman, R. B. (2006). The single category implicit association test as a measure of implicit social cognition. *Journal of Personality and Social Psychology, 91,* 16-32.

Kawakami, K., Amodio, D. M., & Hugenberg, K. (2017). Intergroup perception and cognition: An integrative framework for understanding the causes and consequences of social categorization. *Advances in Experimental Social Psychology, 55,* 1-80.

Klauer, K. C., Voss, A., Schmitz, F., & Teige-Mocigemba, S. (2007). Process-components of the Implicit Association Test: A diffusion model analysis. *Journal of Personality and Social Psychology, 93,* 353-368.

Krieglmeyer, R., & Deutsch, R. (2010). Comparing measures of approach-avoidance behavior: the manikin task vs. two versions of the joystick task. *Cognition and Emotion, 24,* 810-828.

Kruglanski, A. W., & Gigerenzer, G. (2011). Intuitive and deliberative judgments are based on common principles. *Psychological Review, 118,* 97-109.

Kruglanski, A. W., & Thompson, E. P. (1999). Persuasion by a single route: A view from the unimodel. *Psychological Inquiry, 10,* 83-109.

Lai, C. K., Marini, M., Lehr, S. A., Cerruti, C., Shin, J. L., Joy-Gaba, J. A., Ho, A. K., Teachman, B. A., Wojcik, S. P., Koleva, S. P., Frazier, R. S., Heiphetz, L., Chen, E., Turner, R. N., Haidt, J., Kesebir, S., Hawkins, C. B., Schaefer, H. S., Rubichi, S., Sartori, G., Dial, C., Sriram, N., Banaji, M. R., & Nosek, B. A. (2014). A comparative investigation of 18 interventions to reduce implicit racial preferences. *Journal of Experimental Psychology: General, 143,* 1765-1785.

Lai, C. K., Skinner, A. L., Cooley, E., Murrar, S., Brauer, M., Devos, T., Calanchini, J., Xiao, Y. J., Pedram, C., Marshburn, C. K., Simon, S., Blanchar, J. C., Joy-Gaba, J. A., Conway, J., Redford, L., Klein, R. A., Roussos, G., Schellhaas, F. M. H., Burns, M., Hu, X., McLean, M. C., Axt, J. R., Asgari, S., Schmidt, K., Rubinstein, R., Marini, M., Rubichi, S., Shin, J. L., & Nosek, B. A. (2016). Reducing implicit racial preferences: II. Intervention effectiveness across time. *Journal of Experimental Psychology: General, 145*, 1001-1016.

LeBel, E. P. (2010). Attitude accessibility as a moderator of implicit and explicit self-esteem correspondence. *Self and Identity, 9,* 195-208.

Leitner, J. B., Hehman, E., Ayduk, O., & Mendoza-Denton, R. (2016). Racial bias is associated with ingroup death rate for Blacks and Whites: Insights from Project Implicit. *Social Science & Medicine, 170,* 220-227.

Lord, C. G., & Lepper, M. R. (1999). Attitude representation theory. *Advances in Experimental Social Psychology, 31,* 265-343.

Lundberg, K. B., & Payne, B. K. (2014). Decisions among the undecided: Implicit attitudes predict future voting behavior of undecided voters. *PLoS ONE 9(1):e85680*.

Ma, D. S., Correll, J., & Wittenbrink, B. (2016). Context dependency at recall: Decoupling context and targets at encoding. *Social Cognition, 34,* 119-132.

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

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

McConahay, J. B. (1986). Modern racism, ambivalence, and the Modern Racism Scale. In J. F. Dovidio, & S. L. Gaertner (Eds.), *Prejudice, discrimination, and racism* (pp. 91-126). New York: Academic Press.

Meissner, F., & Rothermund, K. (2013). Estimating the contributions of associations and recoding in the Implicit Association Test: The ReAL model for the IAT. *Journal of Personality and Social Psychology, 104,* 45-69.

Metcalfe, J., & Mischel, W. (1999). A hot/cool-system analysis of delay of gratification. Dynamics of willpower. *Psychological Review, 106,* 3-19.

Mitchell, J. P., Nosek, B. A., & Banaji, M. R. (2003). Contextual variations in implicit evaluation. *Journal of Experimental Psychology: General, 132,* 455-469.

Moran, T., & Bar-Anan, Y. (2013). The effect of object-valence relations on automatic evaluation. *Cognition and Emotion, 27,* 743-752.

Moran, T., Bar-Anan, Y., & Nosek, B. A. (2015). Processing goals moderate the effect of co-occurrence on automatic evaluation. *Journal of Experimental Social Psychology, 60,* 157-162.

Neumann, R., Förster, J., & Strack, F. (2003). Motor compatibility: The bidirectional link between behavior and emotion. In J. Musch & K. C. Klauer (Eds.), *The psychology of evaluation: Affective processes in cognition and emotion* (pp. 371–391). Mahwah, NJ: Erlbaum.

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

Nosek, B. A., & Banaji, M. R. (2001). The go/no-go association task. *Social Cognition, 19,* 625-666.

Nosek, B. A., Smyth, F. L., Sriram, N., Linder, N. M., Devos, T., Ayala, A.,… Greenwald, A. G. (2009). National differences in gender-science stereotypes predict national sex differences in science and math achievement. *Proceedings of the National Academy of Sciences, 106,* 10593-10597.

Olson, M. A., & Fazio, R. H. (2001). Implicit attitude formation through classical conditioning. *Psychological Science, 12,* 413-147.

Olson, M. A., & Fazio, R. H. (2003). Relations between implicit measures of prejudice: What are we measuring? *Psychological Science, 14,* 636-639.

Olson, M. A., & Fazio, R. H. (2004). Reducing the influence of extra-personal associations on the Implicit Association Test: Personalizing the IAT. *Journal of Personality and Social Psychology, 86,* 653-667.

Olson, M. A., & Fazio, R. H. (2006). Reducing automatically activated racial prejudice through implicit evaluative conditioning. *Personality and Social Psychology Bulletin, 32,* 421-433.

Orchard, J., & Price, J. (2017). County-level racial prejudice and the black-white gap in infant health outcomes. *Social Science & Medicine, 181,* 191-198.

Oswald, F. L., Mitchell, G., Blanton, H., Jaccard, J., & Tetlock, P. E. (2013). Predicting ethnic and racial discrimination: A meta-analysis of IAT criterion studies. *Journal of Personality and Social Psychology, 105,* 171-192.

Paulhus, D. L., (1984). Two component models of social desirable responding. *Journal of Personality and Social Psychology, 46,* 598-609.

Payne, B. K., & Bishara, A. J. (2009). An integrative review of process dissociation and related models in social cognition. *European Review of Social Psychology, 20, 272-314.*

Payne, B. K., Brown-Iannuzzi, J., Burkley, M., Arbuckle, N. L., Cooley, E., Cameron, C. D., & Lundberg, K .B. (2013). Intention invention and the Affect Misattribution Procedure: Reply to Bar-Anan and Nosek (2012). *Personality and Social Psychology Bulletin, 39,* 375-386.

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

Payne, B. K. & Lundberg, K. B. (2014). The affect misattribution procedure: Ten years of evidence on reliability, validity, and mechanisms. *Social and Personality Psychology Compass, 8,* 672-686.

Payne, B. K., Vuletich, H. A., & Lundberg, K. B. (in press). The bias of crowds: How implicit bias bridges personal and systemic prejudice. *Psychological Inquiry*.

Penke, L., Eichstaedt, J., & Asendorpf, J. B. (2006). Single Attribute Implicit Association Tests (SA-IAT) for the assessment of unipolar constructs: The case of sociosexuality. *Experimental Psychology, 53,* 283-291.

Perugini, M., Richetin, J., & Zogmeister, C. (2010). Prediction of behavior. In B. Gawronski, & B. K. Payne (Eds.), *Handbook of implicit social cognition: Measurement, theory, and applications* (pp. 255-277). New York: Guilford Press.

Peters, K. R., & Gawronski, B. (2011a). Mutual influences between the implicit and explicit self-concepts: The role of memory activation and motivated reasoning. *Journal of Experimental Social Psychology, 47,* 436-442.

Peters, K. R., & Gawronski, B. (2011b). Are we puppets on a string? Comparing the effects of contingency and validity on implicit and explicit evaluations. *Personality and Social Psychology Bulletin, 37,* 557-569.

Petty, R. E., Briñol, P., & DeMarree, K. G. (2007). The meta-cognitive model (MCM) of attitudes: Implications for attitude measurement, change, and strength. *Social Cognition*, *25*, 657-686.

Petty, R. E., & Cacioppo, J. T. (1986). The Elaboration Likelihood Model of persuasion. *Advances in Experimental Social Psychology*, *19*, 123-205.

Petty, R. E., Fazio, R. H., & Briñol, P. (2009). The new implicit measures: An overview. In R. E. Petty, R. H. Fazio, & P. Briñol (Eds.). *Attitudes: Insights from the new implicit measures* (pp. 3-18). New York: Psychology Press.

Petty, R. E., Tormala, Z. L., Brinol, P., & Jarvis, W. B. G. (2006). Implicit ambivalence from attitude change: An exploration of the PAST Model. *Journal of Personality and Social Psychology, 90,* 21-41.

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.

Richeson, J. A., & Ambady, N. (2003). Effects of situational power on automatic racial prejudice. *Journal of Experimental Social Psychology, 39,* 177-183.

Richetin, J., Perugini, M., Adjali, I., & Hurling, R. (2007). The moderator role of intuitive versus deliberative decision making for the predictive validity of implicit measures. *European Journal of Personality, 21,* 529-546.

Roefs, A., Quaedacjers, L., Werrij, M. Q., Wolters, G., Havermans, R., Nederkoorn, C., van Breukelen, G., & Jansen, A. (2006). The environment influences whether high-fat foods are associated with palatable or with unhealthy. *Behaviour Research and Therapy, 44,* 715-736.

Rothermund, K., Teige-Mocigemba, S., Gast, A., & Wentura, D. (2009). Minimizing the influence of recoding in the implicit association test: The recoding-free implicit association test (IAT-RF). *The Quarterly Journal of Experimental Psychology, 62*, 84-98.

Rudman, L. A., Greenwald, A. G., Mellott, D. S., & Schwartz, J. L. (1999). Measuring the automatic components of prejudice: Flexibility and generality of the Implicit Association Test. *Social Cognition*, *17*, 437-465.

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

Rydell, R. J., McConnell, A. R., & Mackie, D. M. (2008). Consequences of discrepant explicit and implicit attitudes: Cognitive dissonance and increased information processing. *Journal of Experimental Social Psychology, 44,* 1526-1532.

Rydell, R. J., McConnell, A. R., Mackie, D. M., & Strain, L. M. (2006). Of two minds: Forming and changing valence-inconsistent implicit and explicit attitudes. *Psychological Science, 17,* 954-958.

Rydell, R. J., McConnell, A. R., Strain, L. M., Claypool, H. M., & Hugenberg, K. (2007). Implicit and explicit attitudes respond differently to increasing amounts of counterattitudinal information. *European Journal of Social Psychology, 37,* 867-878.

Schnabel, K., Banse, R., & Asendorpf, J. (2006). Employing automatic approach and avoidance tendencies for the assessment of implicit personality self-concept: The Implicit Association Procedure (IAP). *Experimental Psychology, 53,* 69-76.

Schwarz, N. (2007). Attitude construction: Evaluation in context. *Social Cognition, 25*, 638-656.

Sherman, J. W., Gawronski, B., Gonsalkorale, K., Hugenberg, K., Allen, T. A., & Groom, C. (2008). The self-regulation of automatic associations and behavioral impulses. *Psychological Review, 115,* 314-335.

Sinclair, S., Lowery, B. S., Hardin, C. D., & Colangelo, A. (2005). Social tuning of automatic racial attitudes: The role affiliative motivation. *Journal of Personality and Social Psychology, 89,* 583-592.

Smith, E. R. (1996). What do connectionism and social psychology offer each other? *Journal of Personality and Social Psychology, 70,* 893-912.

Solarz, A. K. (1960). Latency of instrumental responses as a function of compatibility with the meaning of eliciting verbal signs. *Journal of Experimental Psychology, 59,* 239-245.

Sriram, N., & Greenwald, A. G. (2009). The brief implicit association test. *Experimental Psychology, 56,* 283-294.

Stahl, C., & Degner, J. (2007). Assessing automatic activation of valence: A multinomial model of EAST performance. *Experimental Psychology, 54,* 99-112.

Strack, F., & Deutsch, R. (2004). Reflective and impulsive determinants of social behavior. *Personality and Social Psychology Review*, *8*, 220-247.

Tamai, N., & Nakajima, S. (2000). Renewal of formerly conditioned fear in rats after extensive extinction training. *International Journal of Comparative Psychology*, *13*, 137-147.

Teige-Mocigemba, S., Klauer, K. C., & Rothermund, K. (2008). Minimizing method-specific variance in the IAT: A single block IAT. *European Journal of Psychological Assessment*, *24*, 237-245.

Teige-Mocigemba, S., Klauer, K. C., & Sherman, J. W. (2010). A practical guide to Implicit Association Tests and related tasks. In B. Gawronski & B. K. Payne (Eds.), *Handbook of implicit social cognition: Measurement, theory, and applications* (pp. 117-139). New York: Guilford Press.

Webb, E. J., Campbell, D. T., Schwartz, R. D., & Sechrist, L. (1966). *Unobtrusive measures: Nonreactive research in the social sciences*. Chicago, IL: Rand McNally.

Wentura, D., & Degner, J. (2010). A practical guide to sequential priming and related tasks. In B. Gawronski, & B. K. Payne (Eds.), *Handbook of implicit social cognition: Measurement, theory, and applications* (pp. 95-116). New York: Guilford Press.

Whitfield, M., & Jordan, C. H. (2009). Mutual influences of explicit and implicit attitudes. *Journal of Experimental Social Psychology, 45*, 748-759.

Wilson, T. D., Lindsey, S. & Schooler, T. Y. (2000). A model of dual attitudes. *Psychological Review, 107,* 101-126.

Wittenbrink, B., Judd, C. M., & Park, B. (1997). Evidence for racial prejudice at the implicit level and its relationships with questionnaire measures. *Journal of Personality and Social Psychology, 72,* 262-274.

Wittenbrink, B. Judd, C. M., & Park, B. (2001). Spontaneous prejudice in context: Variability in automatically activated attitudes. *Journal of Personality and Social Psychology, 81*, 815-827.

Wojnowicz, M., Ferguson, M., Dale, R., & Spivey, M. J. (2009). The self-organization of explicit attitudes. *Psychological Science, 20,* 1428-1435.

Ye, Y., Tong, Y.-Y., Chiu, C.-Y., & Gawronski, B. (2017). Attention to context during evaluative learning and context-dependent automatic evaluation: A cross-cultural analysis. *Journal of Experimental Social Psychology, 70,* 1-7.

Zanon, R., De Houwer, J., & Gast, A. (2012). Context effects in evaluative conditioning of implicit evaluations. *Learning & Motivation, 43*, 155-165.

Figure 1. Potential direct and indirect influences of an external factor on associative and propositional processes underlying changes in implicit and explicit evaluations according to the associative-propositional evaluation (APE) model. Thin arrows depict direct effects of an external factor on either associative or propositional processes and influences of the two processes on implicit and explicit evaluations; fat arrows depict mutual influences between associative and propositional processes, with solid arrows depicting the presence of an effect and open arrows the absence of an effect. Figure adapted from Gawronski and Bodenhausen (2011); reprinted with permission.

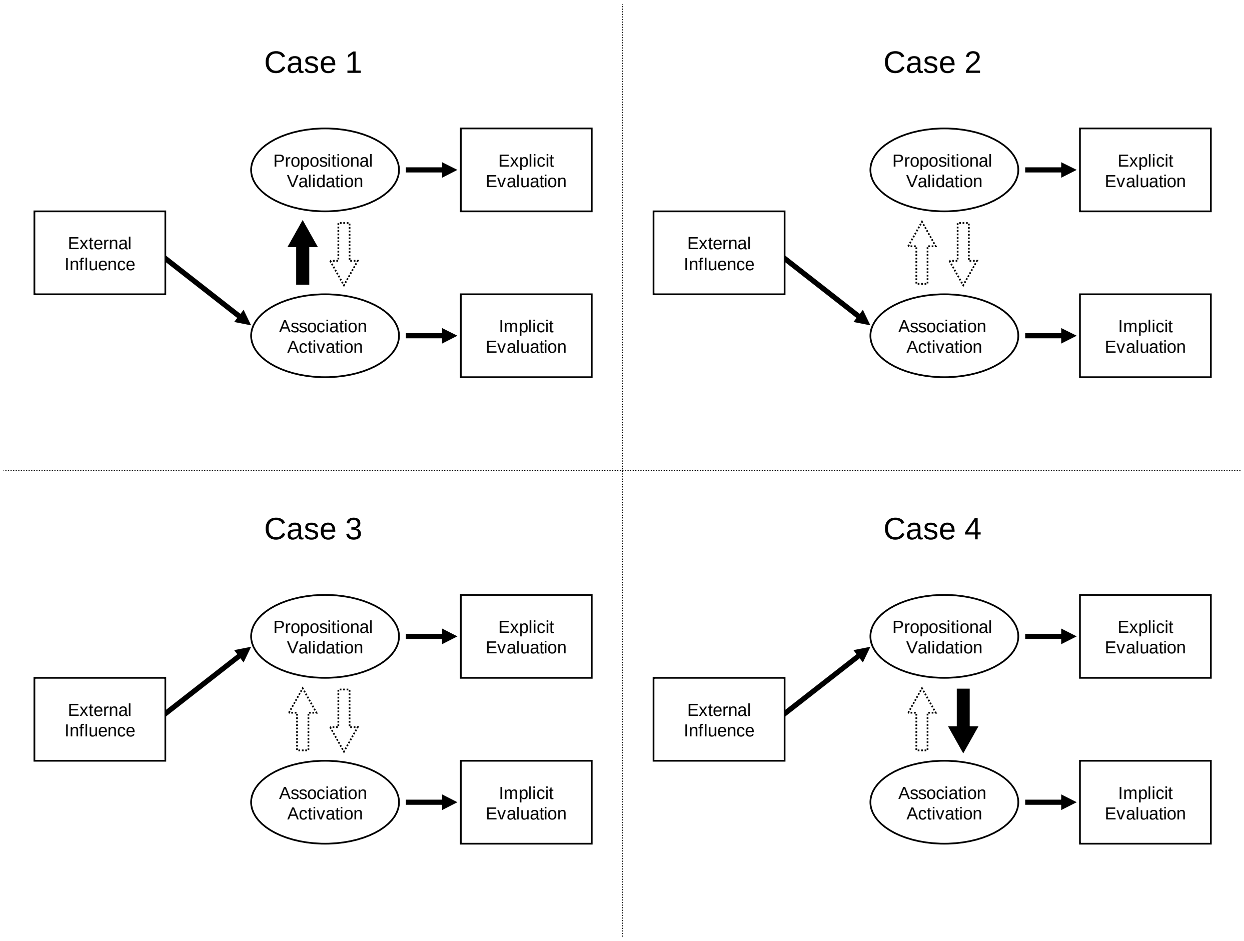
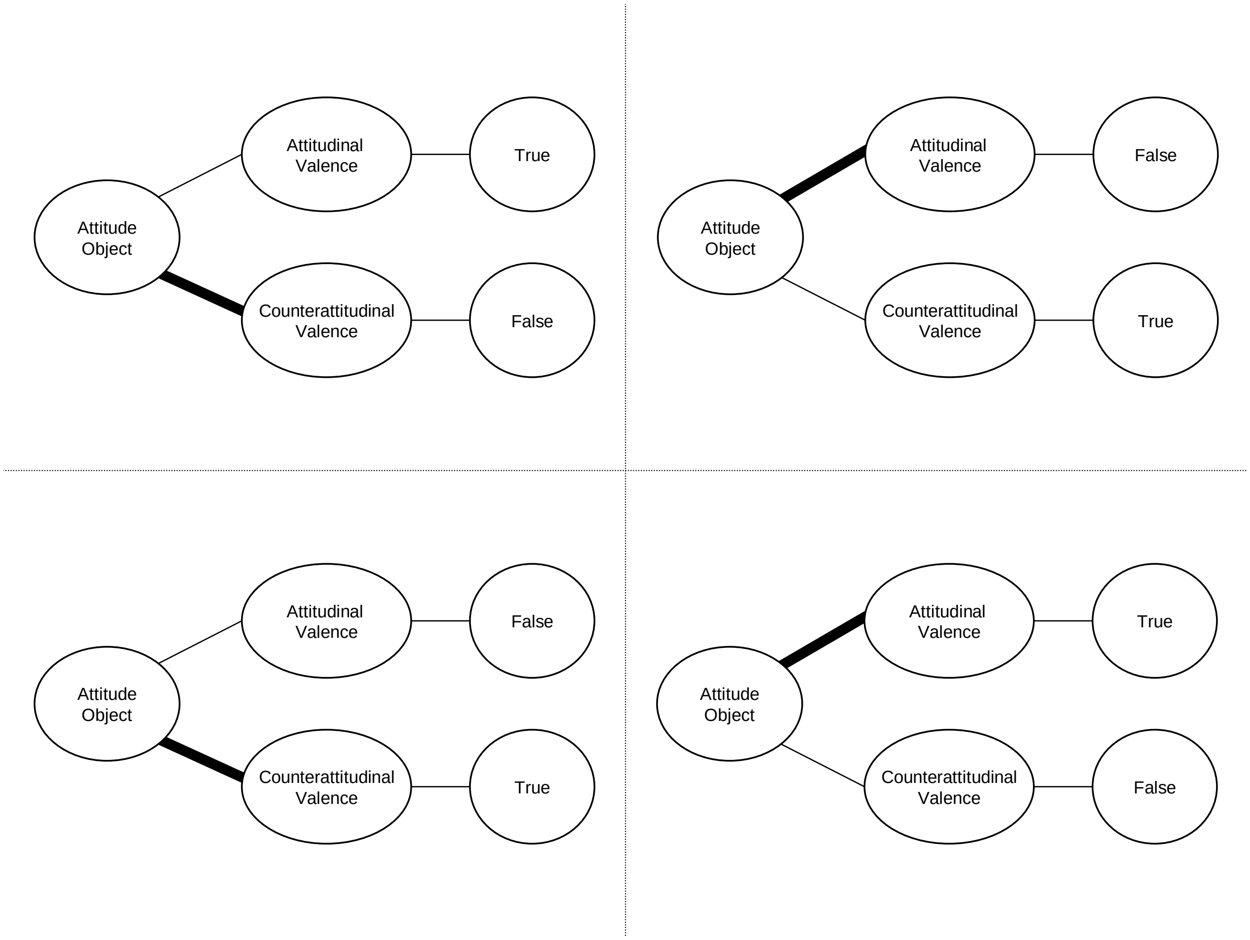


Figure 2. Associative representations of conflicting attitudinal and counterattitudinal information tagged as either true or false according to the meta-cognitive model (MCM). Thin lines depict weak associations; fat lines depict strong associations.


Attitude Object
Attitudinal Valence
True
Counterattitudinal Valence
False
Attitude Object
Attitudinal Valence
False
Counterattitudinal Valence
True
Attitude Object
Attitudinal Valence
False
Counterattitudinal Valence
True
Attitude Object
Attitudinal Valence
True
Counterattitudinal Valence
False