Response Interference as a Mechanism Underlying Implicit Measures

Some Traps and Gaps in the Assessment of Mental Associations with Experimental Paradigms

Bertram Gawronski[1], Roland Deutsch[2], Etienne P. LeBel[1], and Kurt R. Peters[1]

[1]University of Western Ontario, Canada, [2]University of Würzburg, Germany

Abstract. Over the last decade, implicit measures of mental associations (e.g., Implicit Association Test, sequential priming) have become increasingly popular in many areas of psychological research. Even though successful applications provide preliminary support for the validity of these measures, their underlying mechanisms are still controversial. The present article addresses the role of a particular mechanism that is hypothesized to mediate the influence of activated associations on task performance in many implicit measures: response interference (RI). Based on a review of relevant evidence, we argue that RI effects in implicit measures depend on participants' attention to association-relevant stimulus features, which in turn can influence the reliability and the construct validity of these measures. Drawing on a moderated-mediation model (MMM) of task performance in RI paradigms, we provide several suggestions on how to address these problems in research using implicit measures.

Keywords: associative processes, attention, implicit measures, reliability, response interference, validity

Introduction

Over the last decade, a new class of indirect measurement procedures has become increasingly popular in many areas of psychological research (for reviews, see Fazio & Olson, 2003; Petty, Fazio, & Briñol, in press; Wittenbrink & Schwarz, 2007). In contrast to standard self-report measures, these *implicit measures* are based on experimental paradigms derived from cognitive psychology, such as sequential priming (Neely, 1977) or other types of compatibility paradigms (Kornblum, Hasbroucq, & Osman, 1990). The most prominent examples are Greenwald, McGhee, and Schwartz's (1998) Implicit Association Test (IAT) and Fazio, Jackson, Dunton, and Williams' (1995) affective priming paradigm. Other examples include the Extrinsic Affective Simon Task (EAST; De Houwer, 2003a), the Go/No-go Association Task (GNAT; Nosek & Banaji, 2001), semantic priming (Wittenbrink, Judd, & Park, 1997), and the affect misattribution procedure (AMP; Payne, Cheng, Govorun, & Stewart, 2005).

Researchers commonly assume that implicit measures assess mental associations that are activated automatically in response to a given stimulus. Although successful applications provide preliminary support for the validity of these measures (for reviews, see Fazio & Olson, 2003; Petty et al., in press), their underlying mechanisms are still controversial (Wittenbrink & Schwarz, 2007). In the present article, we argue that implicit measures do not directly reflect the activation of mental associations. Instead, implicit measures provide only an indirect proxy for mental associations, such that the impact of activated associations on task performance is mediated by mechanisms that are specific to the task.

In the following sections, we will review empirical findings regarding a particular mechanism that has been the subject of our research over the last years: response interference (RI) (for alternative mechanisms, see Brendl, Markman, & Messner, 2001; Klauer & Musch, 2003; Rothermund & Wentura, 2004). In a nutshell, this research has shown that RI effects in implicit measures depend on participants' attention to association-relevant stimulus-features, which in turn has important implications for the reliability and the construct validity of these measures. For this purpose, we will first illustrate the notion of RI and how it is reflected in many (though not all) implicit measures (see Table 1). Based on this discussion, we will then review empirical evidence showing how attentional processes influence the reliability and the construct validity of RI tasks. This review will be supplemented by specific suggestions on how to address potential problems arising from attentional influences in research using implicit measures.

DOI 10.1027/1015-5759.24.4.218

Table 1. Presence versus absence of response interference mechanisms in different kinds of implicit measures

Implicit measure		Response interference?
Affect Misattribution Procedure	Payne et al. (2005)	No
Affective Priming with Evaluative Decision Task	Fazio et al. (1995)	Yes
Affective Priming with Pronunciation Task	Bargh et al. (1996)	No
Approach-Avoidance Paradigm with Evaluative Judgment Task	Chen & Bargh (1999; Exp. 1)	Yes
Approach-Avoidance Paradigm with Nonevaluative Judgment Task	Chen & Bargh (1999; Exp. 2)	Yes
Extrinsic Affective Simon Task	De Houwer (2003)	Yes
Go/No-go Association Task	Nosek & Banaji (2001)	Yes
Implicit Association Test	Greenwald et al. (1998)	Yes
Personalized Implicit Association Test	Olson & Fazio (2004)	Yes
Semantic Priming with Lexical Decision Task	Wittenbrink et al. (1997)	No
Semantic Priming with Semantic Categorization Task	Banaji & Hardin (1996; Exp. 1)	Yes
Single Category Implicit Association Test	Karpinski & Steinman (2006)	Yes
Weapon Identification Task	Payne (2001)	Yes

Response Interference in Implicit Measures

A useful example to illustrate the notion of RI is the Stroop color-naming task. In this task, participants are asked to name the color of a word presented on a screen as quickly as possible. The critical items in this task are words that themselves represent a color label. On these items, people usually show better performance when the ink color of the word corresponds to the color label depicted by the word (e.g., the word *RED* written in red ink) than when ink color and color label do not correspond to one another (e.g., the word *RED* written in blue ink). These differences in performance can be explained by the influence of two independent response tendencies elicited by the ink color and the semantic meaning of the stimulus. For instance, the word *RED* written in blue ink may elicit two response tendencies that interfere with a quick and accurate response to that stimulus, namely a response tendency to say "red" elicited by the word meaning and a response tendency to say "blue" elicited by the ink color. Conversely, the word *RED* written in red ink may elicit two response tendencies that facilitate quick and accurate responses, namely a response tendency to say "red" elicited by the word meaning and a response tendency to say "red" elicited by the ink color. Put differently, the first case results in two response tendencies that have antagonistic effects on participants' responses, whereas the latter case results in two response tendencies that have synergistic effects.

Such Stroop-like RI effects play a significant role in many implicit measures (De Houwer, 2003b). One example is Greenwald et al.'s (1998) IAT. In this task, participants are asked to categorize individual stimuli (e.g., Black and White faces) as quickly as possible into a pair of target categories (e.g., Black vs. White). The strength of mental associations between the target categories is usually assessed by combining category-related responses in an association-congruent and an association-incongruent manner. For example, in an IAT designed to assess White participants' implicit preference for Whites over Blacks, participants are asked to respond to pictures of Black and White individuals and to pleasant and unpleasant words with a key assignment representing a prejudice-congruent combination (i.e., Black-negative; White-positive), and with a key assignment representing a prejudice-incongruent combination (i.e., White-negative; Black-positive). The difference between the mean response latencies for prejudice-congruent and prejudice-incongruent assignments is usually interpreted as an index of participants' implicit preference for Whites over Blacks.

Applied to the present question, responses on the IAT can be understood in terms of the same RI mechanism that underlies performance on the Stroop task (De Houwer, 2003b; Gawronski & Bodenhausen, 2005). For example, in the combined blocks of an IAT designed to assess implicit preferences for Whites over Blacks, the target items (e.g., Black and White faces) may elicit a prepotent response tendency that is driven by the evaluative associations that are activated in response to the stimulus (i.e., negative vs. positive) and another response tendency that is driven by the category membership of the stimulus (i.e., Black vs. White). In the congruent block, both response tendencies lead to correct responses in terms of the key assignment (synergistic effect). In the incongruent block, however, only category-based response tendencies lead to correct responses, whereas evaluation-based response tendencies lead to incorrect responses (antagonistic effect). Hence, the IAT involves an RI mechanism comparable to the one in the Stroop task.

A similar logic applies to affective priming effects in Fazio et al.'s (1995) sequential priming paradigm using an evaluative decision task. In this task, participants have to indicate the valence of positive and negative target words as quickly as possible. Shortly before the presentation of a target word, participants are briefly presented with either a positive or a negative prime stimulus. The standard affective priming effect is reflected in faster responses to positive words after priming with positive as compared to negative stimuli, and in faster responses to negative words after priming with negative as compared to positive stimuli (for a review, see Klauer & Musch, 2003).

As with the IAT, affective priming effects in Fazio et al.'s (1995) paradigm can be understood in terms of Stroop-like RI processes (e.g., De Houwer, Hermans, Rothermund, & Wentura, 2002; Gawronski, Deutsch, & Seidel, 2005; Klinger, Burton, & Pitts, 2000). Specifically, the valence of the prime stimulus may trigger a prepotent response tendency that can be compatible or incompatible with the response tendency triggered by the target word. If the prime stimulus and the target word share the same valence, the two response tendencies have synergistic effects. If, however, the prime stimulus and the target word have a different valence, the two response tendencies have antagonistic effects. In this sense, affective priming effects in Fazio et al.'s (1995) paradigm are driven by two competing response tendencies, thereby implying an RI mechanism similar to the one in the IAT and the Stroop task (De Houwer, 2003b; Klauer & Musch, 2003).

To further illustrate the mediating role of RI processes in implicit measures, it seems useful to juxtapose the aforementioned paradigms with other measures that do not involve an RI component (see Table 1). One example of the latter category is Wittenbrink et al.'s (1997) sequential priming paradigm using a lexical decision task. Even though this measure seems superficially comparable to Fazio et al.'s (1995) paradigm, its underlying mechanisms are quite different. In Wittenbrink et al.'s (1997) task, participants have to indicate as quickly as possible whether a letter string presented on the screen constitutes a meaningful word or a meaningless nonword. Before the presentation of a target letter string, participants are briefly presented with a meaningful prime stimulus. Priming effects are reflected in faster responses to meaningful target words when these words follow a semantically related prime word (e.g., *bread* followed by *butter*) than when these words follow a semantically unrelated prime word (e.g., *bread* followed by *car*). Although such semantic priming effects may appear similar to affective priming effects in Fazio et al.'s (1995) task, priming effects on lexical decisions are generally independent of RI processes (De Houwer, 2003b; Klauer & Musch, 2003). Prime words in Wittenbrink et al.'s (1997) task are always meaningful words and, thus, generally imply a response tendency to press the key for meaningful words. This response tendency, however, is irrelevant for the facilitated identification of related target words. Critically, both semantically related and semantically unrelated target-words should benefit from the prime-related response tendency to press the "meaningful" key. There is no response-related incompatibility in responses to semantically related vs. semantically unrelated target words. Instead, semantic priming effects on lexical decisions exclusively depend on the semantic relatedness of primes and targets at the stimulus level (De Houwer, 2003b).

Another measure that does not include an RI component is Payne et al.'s (2005) affect misattribution procedure (AMP). In this paradigm, participants are briefly presented with a positive or a negative prime stimulus, which is followed by a neutral Chinese character. Participants' task is to indicate whether they consider the Chinese character as visually more pleasant or less pleasant than the average Chinese character. Affective priming effects in this paradigm are reflected in assimilation effects, such that the neutral Chinese ideographs are evaluated more positively when they were preceded by a positive prime stimulus than when they were preceded by a negative prime stimulus. Again, even though Payne et al.'s (2005) AMP may appear very similar to Fazio et al.'s (1995) paradigm, the mechanisms underlying the two measures are quite different. Unlike Fazio et al.'s (1995) paradigm, the AMP is assumed to be driven by a misattribution mechanism, whereby the affect elicited by the prime is (mistakenly) used to evaluate the Chinese character. In Payne et al.'s (2005) studies, the Chinese characters appeared on the screen for only 100 ms, after which they were replaced by a masking stimulus. Such conditions make it difficult to deliberately process specific features of the target stimulus. At the same time, the affective state elicited by the prime is likely to persist during the presentation of the Chinese character, thereby biasing participants' evaluations of the target. Thus, as Payne et al. (2005) argued, participants seem to mistakenly assume that their affective reaction stems from the target character, which may result from their inability to disentangle the relative contributions of prime-related vs. target-related responses to their momentary affective state. Moreover, the target stimuli in the AMP are typically selected to be evaluatively neutral in order to maximize the likelihood that both positive and negative valence can be attributed to them. Thus, they do not trigger the same kind of response tendencies as the target words in Fazio et al.'s (1995) task, which could be congruent or incongruent with the response tendencies elicited by the primes. Hence, the mechanisms underlying the two measures are quite different, such that Fazio et al.'s (1995) paradigm is based on the interference of two independent response tendencies resulting from the prime and target stimuli, whereas the AMP is based on the misattribution of prime characteristics to the neutral target stimuli.

The basic logic of RI can be applied to many other currently employed measures, namely the extrinsic affective Simon task (EAST, De Houwer, 2003a), the go/no-go association task (GNAT; Nosek & Banaji, 2001), various approach-avoidance paradigms (Chen & Bargh, 1999), and different modifications of the IAT (e.g., Karpinski & Steinman, 2006; Olson & Fazio, 2004). In fact, the proportion of measures that do not involve an RI component seems relatively small, including semantic priming using a lexical decision task (Wittenbrink et al., 1997), the AMP (Payne et al., 2005), and affective priming using a pronunciation task (Bargh, Chaiken, Raymond, & Hymes, 1996). For the sake of brevity, we will refrain from a detailed discussion

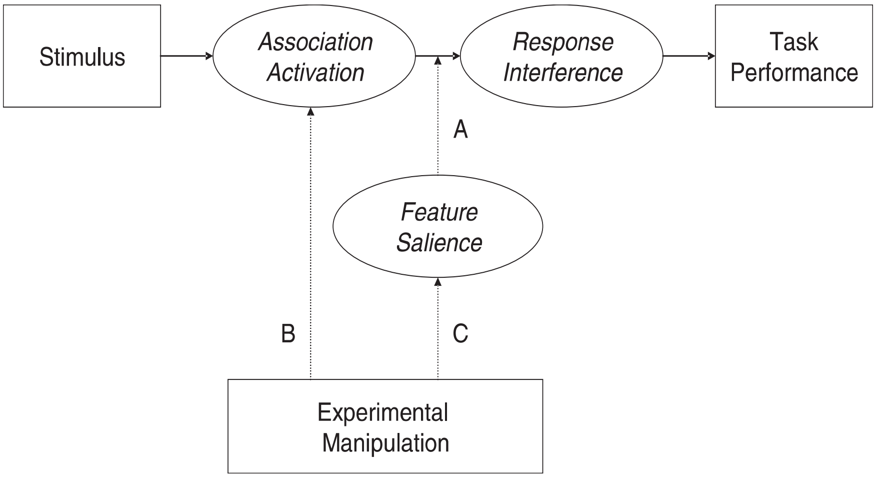

Figure 1. Hypothesized sequence of processes mediating between stimulus presentation and task performance on an implicit measure based on response interference, with feature salience functioning as a moderator of response interference effects elicited by activated associations (A). Experimental manipulations may influence task performance via two routes: primary effects on association activation (B), or secondary effects on feature salience, moderating response interference effects resulting from activated associations (C).

of these measures. Instead, we will provide a more detailed discussion of how attentional processes can influence the reliability and the validity of implicit measures that involve an RI component.

Reliability of Response Interference Tasks

The fact that many implicit measures are based on Stroop-like RI processes (see Table 1) provides important insights into factors that may influence the reliability of these measures. One example is the role of attentional processes. Several studies have shown that participants' performance on the Stroop task is significantly enhanced when they are able to ignore the semantics of the color word (e.g., Besner, Stolz, & Boutilier, 1997). Using the present terminology, these results suggest that the response tendency elicited by the semantics of the color word depends on the relative salience of this particular stimulus feature. Thus, to the degree that participants are able to ignore this feature (e.g., by directing their attention to alternative aspects of the stimulus), the response tendency elicited by the word meaning should be reduced, thereby reducing RI effects in the task.

These insights have important implications for implicit measures that are based on RI processes. One implication is that the prepotent response tendency elicited by activated associations depends on the salience of the association-relevant stimulus feature (see Figure 1). For instance, in an implicit measure designed to assess mental associations between *Black* and *negative*, the prepotent response tendency to press the *negative* key in response to a Black face may depend on the salience of the stimulus feature *Black*. Thus, to the extent that participants are able to ignore this particular stimulus feature, RI effects resulting from evaluative associations regarding Blacks may be reduced, thereby undermining a reliable assessment of these associations. In other words, existing associations regarding a particular stimulus feature may be picked up by the measure only when participants pay attention to the association-relevant feature, but not when they are able to ignore that feature.

This issue is obviously less relevant for measures that require participants to pay attention to the association-relevant stimulus feature. For example, in the IAT participants are required to categorize the target stimuli in terms of the association-relevant feature (e.g., Black faces need to be categorized as Black), which makes it quite difficult (if not impossible) to ignore this feature. However, other measures do not have such a built-in mechanism to control participants' attention, which makes these measures more susceptible to attentional influences. One example is Fazio et al.'s (1995) affective priming paradigm using an evaluative decision task. In this paradigm, participants are typically not required to attend to the association-relevant feature of the primes. Thus, the prepotent response tendencies resulting from these associations may be reduced when participants attend to alternative stimulus features (e.g., Olson & Fazio, 2003). For instance, if participants pay attention to the age rather than the race of a Black face prime, prepotent response tendencies resulting from race-related associations may be reduced, thereby undermining a reliable assessment of these associations by means of RI mechanisms.

A recent study by Gawronski, Cunningham, LeBel, and Deutsch (2007) confirmed these assumptions. In this study, participants completed an affective priming task that was based on the procedure by Fazio et al. (1995). As prime stimuli, Gawronski et al. used pictures of Black or White individuals of either young or old age. Adopting a manipulation from Olson and Fazio (2003), half of the participants were instructed to keep a mental tally of how many Black vs. White individuals were presented over the course of the task; the remaining half were asked to keep a mental tally of how many young vs. old individuals were presented during the task. Based on the considerations outlined above, Gawronski et al. expected that instructing participants to pay attention to a particular feature of the prime stimuli would enhance the reliability of the measure in assessing evaluative associations pertaining to that feature, whereas a reliable assessment of evaluative associations pertaining to the respective alternative feature should be undermined. Thus, given that most White North American undergraduates tend to show implicit preferences for Whites over Blacks (implicit racism; see Nosek et al., 2007) and for young over old people (implicit ageism; see Nosek et al., 2007), Gawronski et al. expected reliable scores of implicit racism when their participants paid attention to race, but not when they paid attention to age. Conversely, Gawronski et al. expected reliable scores of implicit ageism when participants paid attention to age, but not when they paid attention to race.

These predictions were generally confirmed. Specifically, Gawronski et al. found that internal consistency estimates of implicit racism scores were significantly higher when participants paid attention to race than when they paid attention to age. A similar pattern emerged for implicit ageism scores, which showed higher reliability estimates when participants paid attention to age than when they paid attention to race. Further reflecting the reduced reliability of the measure in the noncorresponding conditions, mean values of implicit racism scores were significantly lower when participants paid attention to age than when they paid attention to race, whereas mean values of implicit ageism were significantly lower when participants paid attention to race than when they paid attention to age. These results are consistent with earlier findings by Olson and Fazio (2003) who found higher implicit racism scores and higher reliability estimates in Fazio et al.'s (1995) affective priming task when participants paid attention to the race of the face primes than when they did not.

Even though these findings are consistent with our claims regarding the role of RI processes in implicit measures, one could object that participants' attention to a given stimulus feature may have influenced which particular associations were activated in response to that stimulus (Fazio, 2007; Olson & Fazio, 2003). For instance, attending to race-related features may facilitate the activation of race-related associations, whereas attending to age-related features may facilitate the activation of age-related associations. In this case, the obtained differences may reflect meaningful changes in the activation of evaluative associations, rather than secondary changes resulting from the moderating effect of feature salience on the impact of activated associations on RI processes (see Figure 1).

To resolve this ambiguity, Gawronski et al. (2007) conducted a second study in which they used an affective priming variant without an RI component, namely, Payne et al.'s (2005) AMP. As outlined above, this measure is similar to Fazio et al.'s (1995) task, in that the presentation of a prime stimulus influences evaluative responses to the target stimuli. However, Payne et al.'s (2005) paradigm is different from Fazio et al.'s (1995) in that it does not involve an RI component. To minimize procedural and material-related confoundings, Gawronski et al. (2007) designed an AMP that used the same procedural parameters (e.g., prime duration) and the same stimulus materials employed in their first study. Thus, if the results obtained by Gawronski et al. (2007) are a result of the activation of different associations as a function of participants' attention to particular features of the primes, Payne et al.'s (2005) measure should produce the same effects that were obtained for Fazio et al.'s (1995) paradigm. If, however, Gawronski et al.'s (2007) results are a result of the moderating effect of attentional processes on the elicitation of prepotent response tendencies resulting from activated associations, the reliability of the scores revealed by Payne et al.'s (2005) measure should be independent of participants' attention.

Gawronski et al.'s (2007) findings provided clear support for the latter prediction. Specifically, reliability estimates of implicit racism scores were reasonably high irrespective of whether participants paid attention to race or age. The same was true for implicit ageism scores, which also showed high estimates irrespective of whether participants paid attention to age or race. Moreover, mean values of preferences for Whites over Blacks and for young over old people were significantly higher than zero, irrespective of whether participants were instructed to pay attention to race or age.

Even though the reviewed findings regarding Fazio et al.'s (1995) measure may seem somewhat discouraging, they clearly have some useful practical implications. Specifically, Gawronski et al.'s (2007) results suggest that the reliability of RI tasks can be enhanced by directing participants' attention toward the association-relevant stimulus feature – for instance, by instructing participants to keep a mental tally of the number of prime stimuli of a particular stimulus category (see Olson & Fazio, 2003). This insight seems important, given the low reliability estimates often obtained for Fazio et al.'s (1995) measure (e.g., Banse, 1999; Cunningham, Preacher, & Banaji, 2001; Olson & Fazio, 2003). Note, however, that even in Gawronski et al.'s (2007) studies, reliability estimates for Fazio et al.'s (1995) paradigm were substantially lower than the ones obtained for Payne et al.'s (2005) task. Thus, even though explicit attention instructions may help to increase the reliability of Fazio et al.'s (1995) measure, more research is needed to improve its reliability to a level that would be acceptable from a measurement perspective.

Construct Validity of Response Interference Tasks

In addition to the obtained influences on internal consistencies, Gawronski et al.'s (2007) findings demonstrate that changes in the attention to association-relevant stimulus features can also influence the mean values obtained with RI tasks. To be sure, the latter finding is almost trivial, given that these changes can be easily explained by the reduced reliability of the measure. Nevertheless, changes in mean values as a function of attentional shifts have important implications for the interpretation of experimentally induced effects on implicit measures. In recent years, researchers in many areas of psychology have become increasingly interested in identifying factors that are capable of changing automatic associations (for a review, Gawronski & Bodenhausen, 2006). The findings by Gawronski et al. (2007) indicate that such effects should be interpreted with caution. Based on the MMM depicted in Figure 1, we argue that any factor that influences participants' attention to association-relevant stimulus features can influence the scores revealed by RI tasks, even when these factors leave the activation of associations unaffected.

The most dramatic example of such effects is a series of studies conducted by Deutsch and Gawronski (2007; Gawronski et al., 2005). Their findings indicate that experimental manipulations can influence the scores revealed by RI tasks in a manner that is in direct opposition to the effects on activated associations. In one set of studies, Gawronski et al. (2005) used a variant of Fazio et al.'s (1995) affective priming task, including two sequential primes rather than a single prime. According to spreading activation models of associative processing (Collins & Loftus, 1975), two sequential primes of the same valence should produce stronger affective priming effects, whereas two sequential primes of the opposite valence should attenuate their individual impact, and thereby affective priming effects.

These predictions stand in contrast to alternative hypotheses by Gawronski et al. (2005) that have been derived from accounts in terms of RI. Specifically, Gawronski et al. argued that the salience of the valence of a given prime stimulus should be enhanced if this stimulus is preceded by another prime of the opposite valence. Conversely, the salience of a prime's valence should be reduced if it is preceded by another prime of the same valence (see Cacioppo, Crites, Berntson, & Coles, 1993). Thus, to the degree that the prepotent response tendency elicited by a given prime depends on the salience of the association-relevant stimulus feature, affective priming effects in Fazio et al.'s (1995) paradigm should be stronger when the prime is preceded by another prime of the opposite valence than when it is preceded by another prime of the same valence. Supporting these assumptions, Gawronski et al. (2005) found clear evidence for contextual contrast effects resulting from two sequential primes, rather than for the additive effects predicted by spreading activation models.

Notwithstanding the consistency of these findings with our MMM (see Figure 1), one could object that evaluative processing may operate according to principles of hedonic contrast rather than spreading activation. Just as lukewarm water is experienced as hot after having placed one's hand in ice water, a negative response to a given stimulus may be more pronounced when this stimulus is encountered in a positive context. According to this perspective, evaluative responses should not be determined by the absolute hedonic level of a given stimulus or event, but by the direction and size of change in the hedonic level (Brickman, Coates, & Janoff-Bulman, 1978). To test these assumptions, Deutsch and Gawronski (2007) conducted a series of studies that compared the effects of two sequential primes in Fazio et al.'s (1995) affective priming paradigm and Payne et al.'s (2005) AMP. To further investigate potential differences between semantic and evaluative processing, Deutsch and Gawronski (2007) also included modified variants of the two paradigms in which the evaluative response dimension (i.e., positive vs. negative) was replaced with a semantic response dimension (i.e., animate vs. inanimate), using prime and target stimuli that varied in terms of this semantic dimension rather than valence. Consistent with the claim that the contextual contrast effects obtained by Gawronski et al. (2005) are driven by differential feature salience rather than hedonic contrast effects, Deutsch and Gawronski (2007) obtained contextual contrast effects in both the evaluative and the semantic variant of Fazio et al.'s (1995) paradigm. In contrast, both the evaluative and the semantic variant of Payne et al.'s (2005) paradigm revealed additive effects, as predicted by spreading activation models (Collins & Loftus, 1975). These results indicate that one and the same experimental manipulation – in this case, the presentation of a context prime – can produce opposite effects on implicit measures as a function of whether or not they involve an RI component. More specifically, Deutsch and Gawronski's (2007) findings suggest that attentional processes can influence RI effects in a manner that is in direct opposition to the effects of activated associations (i.e., contrastive vs. additive effects).

Needless to say, such attentional influences challenge the construct validity of experimentally induced effects on RI tasks as long as these effects are not replicated with a measure that does not involve an RI component. A useful supplement in this regard may be Payne et al.'s (2005) AMP or Wittenbrink et al.'s (1997) semantic priming paradigm using a lexical decision task. For instance, if a given manipulation produces the same effects on both Fazio et al.'s (1995) and Payne et al.'s (2005) paradigms, task-related influences that are unrelated to activated associations can be ruled out with much greater confidence than when a given effect occurs only for one of the two measures.

Summary and Conclusions

In the present article, we argued that implicit measures do not directly reflect the activation of mental associations. Instead, implicit measures provide only an indirect proxy for mental associations, such that the impact of activated associations on task performance is mediated by mechanisms that are specific to the task. Focusing on a particular mechanism that has been the subject of our own research, response interference (RI), the reviewed evidence suggests that prepotent response tendencies elicited by activated associations depend on participants' attention to association-relevant stimulus features, which in turn can influence the reliability and the construct validity of these measures. With regard to the reliability of RI tasks, the reviewed findings suggest that the internal consistency of RI tasks is significantly reduced when participants do not pay attention to the association-relevant stimulus feature. In many cases, this issue can be overcome by directing participants' attention toward the association-relevant stimulus feature (e.g., by instructing participants to count the number of prime stimuli of a particular category). With regard to the construct validity of RI tasks, our analysis implies that experimental effects on RI tasks may sometimes reflect secondary changes in the attention to stimulus features, rather than genuine changes in activated associations. Given these findings, it would seem wise to confirm the validity of any such effects with multiple measures, ideally measures that do vs. do not involve an RI component. To be sure, the evidence that is available so far is limited to a few implicit measures and future research is required to confirm the generalizability of our claims to other measures (see Table 1). However, given other findings showing similar differences between measures that do vs. do not involve an RI component (e.g., Gawronski & Bodenhausen, 2005), we are optimistic that our generalized claims will turn out to be accurate.

Another important issue related to RI in implicit measures is the notion of cognitive control. The studies reviewed in the present article were primarily concerned with the factors that determine the activation of a prepotent response tendency that may be compatible or incompatible with the correct response required in the task. This focus on the activation of prepotent response tendencies is not comprehensive, as the subsequent influence of these tendencies on task performance further depends on various other factors. For instance, in their quad-model of implicit task performance, Conrey, Sherman, Gawronski, Hugenberg, and Groom (2005) argued that the impact of a prepotent response tendency in RI tasks (labeled *AC* in the quad-model) depends on whether participants are able to identify the accurate response required in the task (labeled *D* in the quad-model) and participants' success at inhibiting the prepotent response tendency in favor of the required, accurate response, if the two are incongruent (labeled *OB* in the quad-model). Thus, to the degree that the identification of the accurate response as well as the inhibition of a prepotent response-tendency represent controlled processes that require cognitive resources (Conrey et al., 2005), measurement scores revealed by RI tasks may be susceptible to various other factors pertaining to the operation of controlled processes. Even though these issues were not the primary concern of the present article, they are nevertheless important for the construct validity of implicit measures, as they could lead to confoundings between individual differences in prepotent response tendencies resulting from activated associations and other individual differences pertaining to the execution of cognitive control (e.g., Mierke & Klauer, 2003).

Notwithstanding these "traps and gaps" in the measurement of mental associations, the reviewed findings do not generally question the usefulness of implicit measures as a tool for psychological research. After all, implicit measures have provided important insights into the workings of the human mind and the processes that guide behavior (Fazio & Olson, 2003; Petty et al., in press). Still, implicit measures – like any other measures – should be used with caution in order to avoid inaccurate interpretations of the data obtained with these measures. We hope that the present article provides some useful insights in this regard to further our understanding of our inner mental lives.

Acknowledgments

Preparation of this article has been supported by grants from the Canada Research Chairs program (CRC), the Social Sciences and Humanities Research Council of Canada (SSHRC), and the Academic Development Fund of the University of Western Ontario (ADF) to the first author, and by a grant by the German Science Foundation to the second author.

References

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

Banse, R. (1999). Automatic evaluation of self and significant others: Affective priming in close relationships. *Journal of Social and Personal Relationships, 16,* 803–821.

Bargh, J.A., Chaiken, S., Raymond, P., & Hymes, C. (1996). The automatic evaluation effect: Unconditional automatic activation with a pronunciation task. *Journal of Personality and Social Psychology, 32,* 104–128.

Besner, D., Stolz, J.A., & Boutilier, C. (1997). The Stroop effect and the myth of automaticity. *Psychonomic Bulletin and Review, 4,* 221–225.

Brendl, C.M., Markman, A.B., & Messner, C. (2001). How do indirect measures of evaluation work? Evaluating the inference of prejudice in the Implicit Association Test. *Journal of Personality and Social Psychology, 81,* 760–773.

Brickman, P., Coates, D., & Janoff-Bulman, R. (1978). Lottery winners and accident victims: Is happiness relative? *Journal of Personality and Social Psychology, 36,* 917–927.

Cacioppo, J.T., Crites, S.L., Berntson, G.G., & Coles, M.G. (1993). If attitudes affect how stimuli are processed, should they not affect the event-related brain potential? *Psychological Science, 4,* 108–112.

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.

Collins, A.M., & Loftus, E.F. (1975). A spreading-activation theory of semantic processing. *Psychological Review, 82,* 407–428.

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.

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

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

De Houwer, J. (2003b). A structural analysis of indirect measures of attitudes. In J. Musch & K.C. Klauer (Eds.), *The psychology of evaluation: Affective processes in cognition and emotion* (pp. 219–244). Mahwah, NJ: Erlbaum.

De Houwer, J., Hermans, D., Rothermund, K., & Wentura, D. (2002). Affective priming of semantic categorization responses. *Cognition and Emotion, 16,* 643–666.

Deutsch, R., & Gawronski, B. (2007). *When the method makes a difference: Antagonistic effects on "automatic evaluations" as a function of task characteristics of the measure*. Manuscript submitted for publication.

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., & Olson, M.A. (2003). Implicit measures in social cognition research: Their meaning and use. *Annual Review of Psychology, 54,* 297–327.

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

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

Gawronski, B., Cunningham, W.A., LeBel, E.P., & Deutsch, R. (2007). *Affective priming as a measure of implicit preferences: Reliability depends on attention to relevant features in response interference paradigms*. Manuscript submitted for publication.

Gawronski, B., Deutsch, R., & Seidel, O. (2005). Contextual influences on implicit evaluation: Additive versus contrastive effects of evaluative context stimuli in affective priming. *Personality and Social Psychology Bulletin, 31,* 1226–1236.

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.

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.

Klauer, K.C., & Musch, J. (2003). Affective priming: Findings and theories. In J. Musch & K.C. Klauer (Eds.), *The psychology of evaluation: Affective processes in cognition and emotion* (pp. 7–49). Mahwah, NJ: Erlbaum.

Klinger, M.R., Burton, P.C., & Pitts, G.S. (2000). Mechanisms of unconscious priming: I. Response competition, not spreading activation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26,* 441–455.

Kornblum, S., Hasbroucq, T., & Osman, A. (1990). Dimensional overlap: Cognitive basis for stimulus-response compatibility – a model and taxonomy. *Psychological Review, 97,* 253–270.

Mierke, J., & Klauer, K.C. (2003). Method-specific variance in the Implicit Association Test. *Journal of Personality and Social Psychology, 85,* 1180–1192.

Neely, J.H. (1977). Semantic priming and retrieval from lexical memory: Roles of inhibitionless spreading activation and limited-capacity attention. *Journal of Experimental Psychology: General, 106*, 226–254.

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

Nosek, B.A., Smyth, F.L., Hansen, J.J., Devos, T., Lindner, N.M., Ranganath, K.A. et al. (2007). Pervasiveness and correlates of implicit attitudes and stereotypes. *European Review of Social Psychology, 18*, 36–88.

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.

Payne, B.K. (2001). Prejudice and perception: The role of automatic and controlled processes in misperceiving a weapon. *Journal of Personality and Social Psychology, 81,* 181–192.

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

Petty, R.E., Fazio, R.H., & Briñol, P. (Eds.). (in press). *Attitudes: Insights from the new implicit measures*. Mahwah, NJ: Erlbaum.

Rothermund, K., & Wentura, D. (2004). Underlying processes in the Implicit Association Test: Dissociating salience from associations. *Journal of Experimental Psychology: General, 133,* 139–165.

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., & Schwarz, N. (Eds.). (2007). *Implicit measures of attitudes.* New York: Guilford.

Bertram Gawronski

Department of Psychology
University of Western Ontario
Social Science Centre
London, Ontario N6A 5C2
Canada
E-mail bgawrons@uwo.ca