The Effect of the Validity of Co-occurrence on Automatic and Deliberate Evaluation

Tal Moran and Yoav Bar-Anan
Ben-Gurion University of the Negev, Beer-Sheva, Israel
Brian A. Nosek
University of Virginia and Center for Open Science, USA

In press, *European Journal of Social Psychology*

Author's note: Correspondence should be addressed to: Tal Moran, Department of Psychology, Ben-Gurion University in the Negev Be'er Sheva, Israel. E-mail: tmo@post.bgu.ac.il. This project was supported by grants from the Israeli Science Foundation [1012/10] to Y. B.-A, and from Project Implicit Inc. and the United States – Israel Binational Science Foundation [2013214] to Y. B.-A and B.A.N.. B.A.N. is an officer of Project Implicit Inc., a nonprofit organization that provided financial and technical support to this project, and includes in its mission "To develop and deliver methods for investigating and applying phenomena of implicit social cognition, including especially phenomena of implicit bias based on age, race, gender or other factors." The authors declare that there are no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Abstract

Co-occurrence of an object and affective stimuli does not always mean that the object and the stimuli are the same valence (e.g., false accusations that Richard is a crook). Contemporary theory posits that information about the (in)validity of co-occurrence has stronger influence on deliberate than automatic evaluation. However, available evidence supports that hypothesis only when the (in)validity information is delayed. Further, the existing evidence is open to alternative methodological accounts. In six high-powered experiments (total N = 1,750), we modified previous procedures to minimize alternative explanations and examine whether delayed (in)validity information has discrepant effect on automatic versus deliberate evaluation. Casting doubt on the generality of the hypothesis, we found more sensitivity of deliberate than automatic evaluation to delayed validity information only when automatic evaluation was measured with the Implicit Association Test and not with the Evaluative Priming task or the Affective Misattribution Procedure.

Key words: Automatic evaluation, Evaluative learning, Attitude formation, Implicit measure, Validity

The Effect of the validity of Co-occurrence on Automatic and Deliberate Evaluation

People evaluate everything: a new job, a co-worker, a potential pet. The evaluation can be deliberate and thoughtful, after serious consideration, or automatic and swift, with little care or intention. It is of particular interest to understand the factors that produce discrepancies between automatic and deliberate evaluation (Gregg, Seibt & Banaji, 2006; Moran & Bar-Anan, 2013; Peters & Gawronski, 2011; Petty, Tormala, Briñol & Jarvis, 2006; Prestwich, Perugini, Hurling, & Richetin, 2010; Ranganath & Nosek, 2008; Ratliff & Nosek, 2010; Rydell, McConnell, Mackie & Strain, 2006; Whitfield & Jordan, 2009).

Contemporary theory (Gawronski & Bodenhausen, 2006, 2011; Petty & Briñol, 2006; Petty, Briñol & DeMarree, 2007; Rydell & McConnell, 2006; Rydell et al., 2006) suggests that an important factor is that automatic evaluation is less sensitive than deliberate evaluation to validity information. For example, according to the Associative-Propositional Evaluation (APE) model automatic evaluation always reflects the activation of evaluative associations, formed mainly as a result of spatiotemporal co-occurrence between stimuli (Gawronski & Bodenhausen, 2006, 2011). In contrast, deliberate evaluation is based on propositional processes that judge the validity of the evaluation implied by the activation of evaluative associations, and relies only on associations that provide valid information regarding the target's "true" valence.

Despite the widespread theoretical assumption that automatic evaluation is less sensitive than deliberate evaluation to validity information, there is very little direct empirical support for it. The purpose of the present research was to provide strong direct tests of this hypothesis by minimizing alternative explanations for previous evidence.

Existing Empirical Evidence

A few studies tested the effect of co-occurring affective stimuli, when there is other source of information suggesting that the target has valence opposite to the co-occurring valence. The co-occurrence was expected to form an association between the target object and the co-occurring affective stimuli. In these studies the co-occurrence information was more impactful on automatic evaluation of the target object than on deliberate evaluation (Petty et al., 2006, Study 1; Rydell & McConnell,

2006; Rydell et al., 2006). Because these studies did not explicitly present the co-occurrence information as invalid, they only suggest that when there are two conflicting sources of information, the co-occurrence information has a stronger effect on automatic than on deliberate evaluation.

There are a few studies that did clarify the validity of the co-occurring affective stimuli by explicitly stating that the co-occurring affective stimuli do not characterize the targets. These studies found, contrary to the discrepancy hypothesis, that when the validity information was provided before or immediately after presenting the co-occurrence, validity information had *the same* effect on automatic and deliberate evaluation (Boucher & Rydell, 2012; Peters & Gawronski, 2011; Siegel, Sigall, & Huber, 2012). The lack of discrepancy found between automatic and deliberate evaluation in studies that provided co-occurrence and validity information at the same time might be interpreted as evidence that automatic evaluation is as sensitive to validity information as deliberate evaluation. However, it is possible that in studies that provided the validity information together with the co-occurrence, the validity information prevented the *formation* of an association between the target and the co-occurring stimuli. The discrepancy assumption pertains to the *expression* of evaluative response, not to the *formation* of mental representations. Therefore, in order to test the discrepancy assumption, it is essential that the participants form associations between the target object and valence that does not characterize the target. Then, the experiment should test whether that association has a stronger influence on automatic than on deliberate evaluation.

One method to induce association formation is to delay the validity information until after the participants have been exposed to the co-occurrence information. Indeed, studies that employed that method provide the only existing evidence for different sensitivity of automatic versus deliberate evaluation to *explicit* validity information. In one demonstration, Peters and Gawronski (2011) presented two target men with negative behavioral descriptions and two with positive behaviors. After participants learned about *all* the behaviors of four men, they were provided with information whether the behavioral descriptions characterized or mischaracterized the target person. One man occurred with positive behaviors that characterized him, and one occurred with characteristic negative behaviors. The other two men co-occurred with behaviors presented as uncharacteristic of them: one with positive behaviors and one with negative behaviors. In their self-reported evaluation,

participants preferred the man who co-occurred with negative behaviors that were uncharacteristic of him over the man who co-occurred with positive behaviors that were uncharacteristic of him. In contrast, the automatic evaluation measure—the Evaluative Priming Task (EPT; Fazio, Jackson, Dunton, & Williams, 1995)—found no evidence for preference between these two targets.

In the other relevant study, participants first learned about two groups, one described favorably and the other unfavorably (Gregg, Seibt, & Banaji, 2006). Then, participants completed measures of deliberate and automatic evaluations of the two groups. After the first measurement, new information revealed to the participants that the previous information was mixed-up, and each group was in fact characterized by the information provided about the other group. Finally, the deliberate and automatic evaluations of the two social groups were measured again. The self-reported preference between the two groups was reversed after participants received the validity information. By contrast, the automatic evaluation measure—the Implicit Association Test (IAT; Greenwald, McGhee, & Schwartz, 1998; Nosek, Greenwald, & Banaji, 2007)—showed the initial preference.

In summary, two studies have found evidence suggesting that automatic evaluation is less sensitive than deliberate evaluation to *changes* in the perceived validity of the co-occurrence information. This is the only *direct* evidence that in comparison to deliberate evaluation, automatic evaluation is less sensitive to explicit validity information.

Alternative Explanations for the Existing Evidence

The results observed by Peters and Gawronski (2011) and Gregg et al. (2006), are open to alternative explanations not relevant to the discrepancy hypothesis. The discrepancy found by Peters and Gawronski could be explained by reduced *overall* sensitivity of the automatic measure in comparison to the deliberate evaluation measure. In their research, Peters and Gawronski found that the self-reported preference between the positive and negative men was weaker when the co-occurring information was uncharacteristic of the men compared to when it was characteristic of them. Moreover, this preference was the weakest in the delayed validity information condition – the only condition that showed a discrepancy between automatic and deliberate evaluation. Therefore, perhaps the induced preference in the invalid co-occurrence condition was too weak to be detected with the automatic evaluation measure, even if it was there. The EPT for automatic evaluation assessment often

shows relatively low reliability, usually below $\alpha = .6$ (Bar-Anan & Nosek, 2014; Gawronski & De Houwer, 2014; Olson & Fazio, 2003). As a consequence, the observed difference could be a function of less reliable automatic versus deliberate measures, and not a difference in the measured evaluative associations (see Buchner & Wippich, 2000, for similar difficulty in research on implicit versus explicit memory).

In Gregg et al.'s study, the main research question pertained to attitude change, rather than to the effect of (in)validity information. Because of that, the attitudes were measured before and after providing the validity information. Participants completed the IAT and a self-report measure before knowing that the co-occurrence information was invalid. That experience could itself have strengthened the automatic evaluation making it more resistant to counterattitudinal information. The results could suggest that after people express evaluation deliberately and automatically, it is easier to reverse deliberate than automatic evaluation.

To summarize, the results found by Peters and Gawronski (2011) and Gregg et al. (2006) are the best existing evidence that explicit validity information can have distinct effects on automatic versus deliberate evaluation, at least when the validity information is not provided immediately with the co-occurrence information. However, perhaps because the main focus of the previous research was not the effect of delayed validity information on evaluation, there are plausible alternative explanations for each of those findings. Because our literature review reveals that this evidence is unique and important, a more definitive test of the effect of delayed validity information on evaluation is needed.

The Present Research

We combined the learning procedures used by Peters and Gawronski and by Gregg et al. to pursue a more definitive test of the discrepancy hypothesis. Like Peters and Gawronski, we asked participants to form impressions of four novel target men, each presented with verbal descriptions of positive or negative behaviors. Then, we used Gregg et al.'s *mix-up* manipulation, and informed participants that, accidentally, the behaviors attributed to one man were actually performed by the other, and vice versa (see also Petty et al., 2006). The advantage of a mix-up manipulation is that it informs participants that the mischaracterized men actually performed negative or positive behaviors—just like the two men who co-occurred with behaviors that they actually performed.

Additionally, we added a memory measure to make sure that participants understood the information regarding the validity of the previous pairing and remembered the valence of the behaviors that each target actually performed. And, we also tested the results after removing participants with inaccurate comprehension.

Finally, to improve the automatic evaluation measurement we tested the effect of the same learning procedure on three different automatic measures. In Experiments 1a-1b, we used the EPT, in Experiments 2a-2b we used the IAT, and in Experiments 3a-3b we used the Affect Misattribution Procedure (AMP; Payne, Cheng, Govorun, & Stewart, 2005). For each measure type, we report the results of the main experiment (Experiment a) and the results of a close replication (Experiment b). We conducted all experiments with large samples for high-powered designs to maximize sensitivity and precision of effect estimation[1].

Experiments 1a-1b

Method

Participants. Participants in all six experiments were volunteers at the Project Implicit research website (Nosek, 2005), who registered for research and were randomly assigned to the study from a pool of available studies[2]. Of the 393 participants who completed Experiment 1a, we excluded five participants who had more than 40% error trials in the EPT (Bar-Anan & Nosek, 2014). The final sample of Experiment 1a included 388 participants (67% women, M_{age} = 30.86, SD_{age} = 14.11). Of the 153 participants who completed the relevant conditions in Experiment 1b[3], we excluded five participants because they had more than 40% error trials in the EPT. The final sample of Experiment 1b included 148 participants (57% women, M_{age} = 30.54, SD_{age} = 13.14).

Materials and procedure.

Stimuli. The four targets were pictures of four males selected from an open database of facial stimuli (Minear & Park, 2004). We named them Chris, James,

[1] We report all data exclusions, manipulations, and measures, and how we determined our sample sizes. To see the materials and data of the whole project (Experiments 1-6) visit https://osf.io/fhvr9/. See the online supplement for extended details on the materials.

[2] In all the experiments, we did not have a specific target sample size other than a plan to collect data from a few hundred participants to guarantee high statistical power. Decisions to stop collecting data did not depend on the obtained results.

[3] 1,564 volunteer participants were randomly assigned to one of the three experiments (Experiments 1b, 2b and 3b). Most of the participants (N = 1,051) completed preference measures that were not relevant to the present research question (see more details in footnote 4).

Michael and David. The targets in the EPT were 14 positive (*Pleasure, Wonderful, Love, Paradise, Cheer, Friend, Splendid, Glee, Smile, Enjoy, Delight, Beautiful, Attractive* and *Likeable*) and 14 negative words (*Bomb, Abuse, Sadness, Pain, Poison, Grief, Ugly, Dirty, Stink, Noxious, Humiliate, Annoying, Disgusting* and *Offensive*; Bar-Anan, 2010).

Learning Procedure. At the beginning of the experiment, participants read these instructions:

> For this study, imagine that you started a new job, and you want to learn about your four co-workers. You find out that your friend Lisa, who used to work in the same job, still remembers a lot about those co-workers. Lisa tells you eight facts about each of them. Next, you will read the eight facts. Try to learn those facts and form an impression about each of your new co-workers. Later, we will test your general memory about these people, and also ask about your attitude toward them.

Each trial presented a behavior, with the image of the person who allegedly performed it. For instance, *Chris helped an elderly man who dropped some packages*. After the participants read four statements about each person (16 total trials), they were requested to indicate about each of them whether the person performed mostly negative or mostly positive behaviors. We used these questions to emphasize to the participants that they should attend to this information. After that test, participants read another four behaviors for each of the targets. In total, participants read eight behaviors attributed to each target person. In Experiment 1a, Chris and Michael were always presented with behaviors with the same valence (e.g., positive), and James and David always with behaviors of the opposite valence. For each participant, the valence of Chris and Michael was randomly chosen as either positive or negative. In Experiment1b, the roles of the four men were completely randomized.[4] The order of the behaviors was randomly chosen for each participant.

[4] In Experiments 1a, 2a and 3a, the assignment of men to roles was randomized with the constraint that David and Michael always shared validity condition. We applied that constraint because the automatic evaluation measure always measured attitudes only toward David and Michael, and we were not interested in comparisons of men that did not share the same validity condition. In Experiments 1b, 2b and 3b, the roles of the four men were completely randomized, leading to some conditions that did not compare men that shared the same validity condition. Because the conditions that did not compare men of the same validity condition were irrelevant to the present research question (and, in retrospect, were hardly informative), we omitted these results from this report (they appear in supplementary web materials, at https://osf.io/2buh7/).

After reading all 32 behaviors, participants read the validity information:

> A few minutes after Lisa tells you all those facts about your four co-workers, she **suddenly realizes that she mixed up James and Chris** [bold in the text]. Everything that she told you about James was really something that Chris did, and everything that she told you about Chris was in fact something that James did. So now you need to change the attitudes that you formed about each of them.
>
> Please take a moment to correct this mistake in your mind. Have another look at the four men [their pictures were displayed in this page again], and correct your impression of them:
>
> James performed all the behaviors that Lisa told you that Chris performed.
>
> Chris performed all the behaviors that Lisa told you that James performed.
>
> There were no mistakes about David and Michael. David and Michael did exactly what Lisa told you they did.

The reversed pair was randomly chosen for each participant (in Experiment 1a, the two men were either James and Chris or Michael and David).

Comprehension test. Next, we tested whether the participants comprehended the mix-up information. For each man, the participants selected one of seven response options to indicate what behaviors the man *actually* performed. The options ranged from *Mostly positive behaviors (certain)* to *Mostly negative behaviors (certain)*, with *Equal number of positive and negative behaviors* in the middle of the response scale, and with *(probably)* and *(guess)* between the middle of the response scale and the scale's extremes.

After the comprehension test, in Experiment 1a participants completed in a random order the measures of automatic and deliberate evaluations[5]. In Experiment 1b, the self-report measure always followed the automatic evaluation measure.

Automatic evaluation. Participants completed an EPT. The EPT always included only two of the four men: those who co-occurred with characteristic behaviors *or* those who co-occurred with uncharacteristic behaviors. Specifically, the EPT always used only two characters (Michael and David) as the prime categories. In

[5] Measure order did not moderate any of the results in Experiments 1a and 2a and therefore was removed from the analyses of these experiments.

the learning phase, these two characters always appeared with behaviors of opposite valence, and always shared the same validity condition.

Each trial in the EPT started with a prime (i.e., a photograph of Michael or David) displayed for 275 milliseconds; then a positive or negative target word appeared until participant categorized it as positive or negative using one of two response keys on the keyboard. If the response was incorrect, a red "X" appeared for 450 milliseconds. An interval of 250 milliseconds preceded the start of the next trial. The rationale behind this task is that liked primes facilitate the classification of positive targets, and impede the classification of negative targets; whereas disliked primes facilitate the classification of negative targets, and impede the classification of positive targets. Therefore, the automatic preference for one prime over another is inferred by comparing the facilitation in responses to positive targets versus negative targets after one prime versus the other prime. The EPT consisted of three blocks of 60 trials (15 trials for each prime-target combination). Participants were instructed to categorize the words as quickly and accurately as they can, and ignore the images.

The reliability of the preference score between David and Michael (computed from four parcels) was α=.46, .40, in Experiments 1a and 1b respectively, when David and Michael were paired with behaviors that they actually performed (the characteristic behaviors condition), and α=.50, .51 when David and Michael were "erroneously" paired with each other's behaviors (the uncharacteristic behaviors condition).

Deliberate evaluation. In Experiment 1a, participants evaluated Michael and David in a random order. In Experiment 1b, participants evaluated all four men in a random order. The instructions were "*Based on your very first emotional response, how much do you like the person in the picture?"* The response scale consisted of 7 responses ranging from *dislike extremely* to *like extremely*.

Design. In all the experiments, the validity factor was manipulated between participants (the evaluation measures included either the men who co-occurred with characteristic behaviors *or* those who co-occurred with uncharacteristic behaviors) and the measure type factor was manipulated within participants (each participant completed the EPT and the self-report). Excluding the many counterbalancing procedural parameters, the design was 2 (valid or invalid condition; between participants) X 2 (automatic or deliberate measure; within participants).

Results and Discussion

Data processing. Following previous work with this data source (Bar-Anan, 2010), EPT analyses were based on log-transformed response latencies, excluding trials with incorrect responses (5.84%, 6.54% of the trials in Experiments 1a and 1b, respectively), responses faster than 300 ms (0.2%, 0.19% of the trials), and responses with latency more than 2.5 SDs away from the participant's average latency in each prime-target condition (3.17%, 2.96% of the trials) [6].

The average latency (in ms) for each condition in the EPT for each condition in the experiments are presented in Table 1. For each of the two primes, the mean reaction time to trials with positive words was subtracted from the mean reaction time to trials with negative words to create evaluation score of the prime target. However, that evaluation score is biased because people are usually faster to categorize positive than negative words (Wentura & Degner, 2010). Therefore, the important EPT score was the difference between the two evaluation scores, representing the preference between the two men.

Evaluation

The main hypothesis we tested in this research is that deliberate evaluation is more sensitive than automatic evaluation to (delayed) validity information. To overcome alternative accounts related to weaker reliability of the EPT, we identified a comparison that should show *stronger* effect for EPT than for self-report if validity information has a stronger impact on self-report. We used the preference for the truly positive man over the truly negative man as the dependent variable. In the invalid condition (when the men co-occurred with uncharacteristic behaviors), people must use the validity information (the information that the behaviors were mixed-up) to form a preference for the truly positive man over the truly negative man. Therefore, measures that are *less* sensitive to validity information would show *smaller* preference for the truly positive person over the truly negative person in the *invalid* condition than in the *valid* condition (i.e., would show a *stronger* effect of validity information). According to the discrepancy hypothesis, the self-report would successfully account for the validity information and show no (or little) difference between the two

[6] In all six experiments, when we repeated the main analysis without removing outlier participants and outlier trials we found the same pattern of results.

preferences. The EPT would be less sensitive to the validity information and show a larger difference between the two preferences.

We submitted the preference for the truly positive man over the truly negative man to a 2 (measure: self-report, EPT, within participants) x 2 (valid or invalid condition, between participants) mixed ANOVA. The discrepancy hypothesis predicts an interaction due to stronger effect of the validity factor on the EPT than on the self-report, reflecting stronger effect of invalid co-occurrence on the EPT than on the self-report. Figure 1 (panels A1 and A2) illustrates this analysis in Experiments 1a and 1b. The preference was stronger in the valid than invalid condition in the two experiments: Experiment 1a: $F(1, 386) = 19.46$, $p < .001$, $\eta_p^2 = .04$; Experiment 1b: $F(1, 146) = 5.85$, $p = .016$, $\eta_p^2 = .03$. Also, preferences were stronger when attitudes were measured with self-report compared to EPT: Experiment 1a: $F(1, 386) = 147.97$, $p < .001$, $\eta_p^2 = .27$; Experiment 1b: $F(1, 146) = 57.84$, $p < .0001$, $\eta_p^2 = .28$. This probably reflects the greater reliability of self-report than EPT leading to greater validity. More importantly, contrary to the discrepancy hypothesis tested in the present research, in both experiments, the effect of validity condition was *not* moderated by measurement type: Experiment 1a: $F(1, 386) = 0.02$, $p = .885$, $\eta_p^2 < .001$; Experiment 1b: $F(1, 146) = 0.58$, $p = .449$, $\eta_p^2 = .003$. In other words, there was no evidence that the automatic evaluation was less sensitive to validity information than the deliberate evaluation; only that it was less sensitive to any information overall.

Although we found no evidence that the two measures were different in their sensitivity to validity information, another interesting question is the influence of invalid co-occurrence on each measure separately. Table 1 details the average rating and average latency for each condition. The self-reported preferences for the truly positive man over the truly negative man were weaker in the *invalid* condition compared to the *valid* condition, Experiment 1a: $F(1, 386) = 21.91$, $p < .001$, $\eta_p^2 = .05$; Experiment 1b: $F(1, 146) = 8.79$, $p = .003$, $\eta_p^2 = .05$. The preference measured by the EPT was reliably weaker in the *invalid* than in the *valid* condition in Experiment 1a, $F(1, 386) = 5.57$, $p = .018$, $\eta_p^2 = .01$, but not in Experiment 1b: $F(1, 146) = 1.16$, $p = .283$, $\eta_p^2 = .007$. Therefore, we found strong evidence that *deliberate* evaluation is sensitive to invalid co-occurrence, and only weak evidence that the EPT is sensitive to invalid co-occurrence.

Table 1
EPT latencies and deliberate evaluation scores in Experiments 1a-1b as a function of man's valence and validity condition.

		Automatic evaluation measure				Deliberate evaluation measure	
Experiment		Valid		Invalid		Valid	Invalid
		Positive words	Negative words	Positive words	Negative words		
1a (EPT)	Positive man	710 (168)	744 (207)	739 (217)	768 (271)	6.28 (1.01)	5.79 (1.48)
All	Negative man	721 (170)	739 (197)	742 (218)	778 (324)	1.99 (1.21)	2.66 (1.74)
	Preference	**16* (87)**		**-6 (161)**		**4.29** (1.91)**	**3.13** (2.86)**
1a (EPT)	Positive man	701 (135)	728 (126)	708 (128)	731 (118)	6.39 (0.91)	6.13 (1.10)
Accurate participants	Negative man	711 (139)	725 (138)	710 (117)	729 (118)	1.82 (1.09)	2.16 (1.36)
	Preference	**13 (83)**		**4 (130)**		**4.56** (1.77)**	**3.97** (2.16)**
1b (EPT)	Positive man	720 (182)	738 (144)	714 (160)	738 (166)	6.41 (1.14)	5.42 (1.84)
All	Negative man	724 (159)	737 (135)	705 (147)	729 (157)	2.03 (1.48)	2.52 (1.83)
	Preference	**5 (97)**		**0 (103)**		**4.38** (2.32)**	**2.91** (3.41)**
1b (EPT)	Positive man	717 (164)	746 (152)	703 (151)	733 (162)	6.56 (1.11)	5.93 (1.31)
Accurate participants	Negative man	731 (167)	745 (142)	701 (144)	718 (156)	1.77 (1.40)	1.86 (1.30)
	Preference	**15 (65)**		**13 (109)**		**4.79** (2.28)**	**4.07** (2.31)**

Notes. Standard deviations are in bracket; For the deliberate evaluation, **preference** is the self-reported evaluation of the negative man subtracted from the self-reported evaluation of the positive man; For the EPT, although we present latency in ms for clarity, the analysis used log transformed latencies; **Preference** is the difference (the mean reaction time to trials with positive words subtracted from the mean reaction time to trials with negative words) of the negative man subtracted from the difference of the positive man; Preference scores significantly different from zero are presented with asterisks ($^{*}p < .05$, $^{**}p < .001$).

Comprehension of the validity information. We averaged the four comprehension items to compute a comprehension score regarding the information about each of the four target persons. The score ranged from 3 (correct response with certainty) to -3 (incorrect response with certainty). Participants' average accuracy regarding the actual behaviors that were performed by each of the two targets in the invalid condition (M = 2.22, 2.31, SD = 1.47, 1.46, in Experiments 1a and 1b) was only slightly lower than the average accuracy regarding the two targets in the valid condition (M = 2.38, 2.50, SD = 1.26, 1.02), $t(387) = 1.84$, $p = .066$, $d = 0.11$ in Experiment 1a, $t(147) = 1.47$, $p = .13$, $d = 0.15$ in Experiment 1b.

Sixty-eight percent (N = 266) of the participants in Experiment 1a and seventy-one percent (N = 106) of the participants in Experiment 1b responded correctly with at least moderate confidence (i.e., indicated a "probably" or "certainly" confidence in the correct response) with regard to all four targets. To make sure that the results are not affected by people who misunderstood instructions or did not pay attention, we tested the results also with these sub-samples. Figure 2 (panels A1 and A2) illustrates the analysis for the accurate sub-samples. The results were similar to those found with the whole sample. The preference was stronger in the valid than invalid condition in Experiment 1a, $F(1, 264) = 3.67$, $p = .056$, $\eta_p^2 = .01$. Unlike the result of the entire sample, in Experiment 1b there was no main effect of validity, $F(1, 104) = 0.99$, $p = .322$, $\eta_p^2 = .009$. Also, preferences were stronger when attitudes were measured with self-report compared to EPT, Experiment 1a: $F(1, 264) = 129.57$, $p < .0001$, $\eta_p^2 = .32$; Experiment 1b: $F(1, 104) = 44.75$, $p < .0001$, $\eta_p^2 = .30$. More importantly, the effect of validity condition was not moderated by measurement type, Experiment 1a: $F(1, 264) = 0.00$, $p = .966$, $\eta_p^2 < .001$; Experiment 1b: $F(1, 104) = 0.19$, $p = .660$, $\eta_p^2 = .001$. The self-reported preferences for the truly positive man over the truly negative man were weaker in the *invalid* condition compared to the *valid* condition, Experiment 1a: $F(1, 246) = 6.02$, $p = .014$, $\eta_p^2 = .02$; Experiment 1b: $F(1, 104) = 2.61$, $p = .109$, $\eta_p^2 = .02$. Validity information did not have a reliable effect on the preference measured with the EPT, Experiment 1a: $F(1, 246) = 1.09$, $p = .298$, $\eta_p^2 = .004$; Experiment 1b: $F(1, 104) = 0.08$, $p = .772$, $\eta_p^2 < .001$.

In summary, in Experiments 1a-1b we did not find evidence that automatic evaluation is less sensitive to explicit validity information than the deliberate evaluation. However, that (null) finding can also be attributed to the low reliability of the EPT. With the present analysis strategy, we tested whether the automatic preference for the truly positive man over the truly negative man would be more sensitive to (in)validity information than the self-reported preference. Perhaps that difference exists, but is not easily detected with the EPT. As illustrated in Figures 1 and 2 (panels A1c and A2c), the EPT found no preference between the two men when the co-occurring behaviors were uncharacteristic of the target men. Perhaps the lack of preference reflects sensitivity to invalid co-occurrence but the EPT fails to detect it due to low reliability. To overcome the reliability obstacle, in Experiments 2a-2b we replaced the EPT with the Implicit Association Test (IAT; Greenwald, et al., 1998;

Nosek, et al., 2007), a more reliable automatic evaluation measure (Bar-Anan & Nosek, 2014; Gawronski & De Houwer, 2014).

Experiments 2a-2b

Method

Participants. Of the 385 participants who completed Experiment 2a, we excluded 11 participants who had more than 10% fast trials (RT< 300ms; Greenwald, Nosek, & Banaji, 2003) in the IAT, or had missing data in the critical blocks of the IAT. The final sample included 374 participants (65% women, M_{age} = 31.76, SD_{age} = 14.14). Of the 168 participants who completed the relevant conditions in Experiment 2b, we excluded six participants who had more than 10% fast trials in the IAT, or had missing data in the critical blocks of the IAT. The final sample of Experiment 2b included 162 participants (59% women, M_{age} = 30.04, SD_{age} = 12.97).

Materials and procedure. The material and procedure were identical to Experiments 1a-1b except that we used the IAT as the automatic measure.

Automatic evaluation. In the IAT, participants categorize stimuli using two computer keys. In the critical blocks, participants respond with the left key to stimuli of two categories (e.g., "Michael" and "Good words"), and with the right key to stimuli of two other categories (e.g., "David" and "Bad words"). In two of these blocks Michael and Good words shared the same response key, and in the other two critical blocks, David and Good words shared the same response key. When "Michael" and "Good words" shared the same key, people with more positive associations for Michael than David should respond more quickly. Therefore, the difference between the average response latencies in blocks that assigned Michael and Good words to the same key and blocks that assigned David and Good words to the same key is interpreted as automatic preference. The seven-block IAT followed the procedure described in Nosek, Greenwald, and Banaji (2005). Michael and positive words always shared the same response key first. We used eight positive words (*Pleasure, Wonderful, Love, Laughter, Happy, Glorious, Joy,* and *Peace*) and eight negative words (*Awful, Failure, Agony, Hurt, Horrible, Terrible, Nasty* and *Evil*) that were succuessfully used in many IATs in the Project Implicit website (Nosek et al., 2007). Based on the IAT scores from four parcels of the IAT critical blocks, internal consistency was α = .84, .75, in Experiment 2a and 2b respectively, in the valid condition, and α = .85, .83 in the invalid condition.

Design. Excluding counterbalancing procedural parameters, the design was 2 (valid or invalid condition; between participants) X 2 (automatic or deliberate measure; within participants).

Results and Discussion

Data processing. The average latency (in ms) for each condition in the IAT for each condition in the experiments are presented in Table 2. For each participant we computed an IAT *D* score (Greenwald et al., 2003) such that positive score indicated a preference for David over Michael. For the main analysis, we standardized those preference scores, and then re-coded them to reflect a preference for the man whose *true* valence was positive over the man whose *true* valence was negative.

Evaluation. Figure 1 (panels B1 and B2) illustrates the standardized and raw preference scores in each condition. We submitted the standardized preference scores to a 2 (measure) x 2 (validity) mixed ANOVA. Like in Experiments 1a-1b, we found a stronger preference in the *valid* than *invalid* condition, Experiment 2a: $F(1, 372) = 19.35$, $p < .001$, $\eta_p^2 = .04$; Experiment 2b: $F(1, 160) = 12.59$, $p = .0005$, $\eta_p^2 = .078$. Further, preferences were stronger with the self-report measure compared to the IAT, Experiment 2a: $F(1, 372) = 45.47$, $p < .0001$, $\eta_p^2 = .10$; Experiment 2b: $F(1, 160) = 17.24$, $p < .0001$, $\eta_p^2 = .09$. This may reflect a stronger overall sensitivity of the self-report measure. Most importantly, in line with the main hypothesis tested in the present research, the *measure* by *validity* interaction was significant, Experiment 2a: $F(1, 372) = 11.54$, $p = .0008$, $\eta_p^2 = .03$; Experiment 2b: $F(1, 160) = 10.31$, $p = .001$, $\eta_p^2 = .06$. The interaction reflected a stronger effect of validity condition on the IAT than on the self-report (Figure 1, panels B1 and B2). Importantly, in the present analysis strategy, a *stronger* effect of the validity condition reflects difficulty in reversing the preference in the mix-up condition based on the mix-up (validity) information. Difficulty in reversing the preference reflects *reduced* sensitivity to validity information. Therefore, the interaction reflected weaker sensitivity of the IAT to validity information in comparison to the self-report's sensitivity. As detailed in Table 2, self-reported preferences for the truly positive man over the truly negative man were only slightly weaker in the *invalid* condition compared to the *valid* condition, Experiment 2a: $F(1, 372) = 3.46$, $p = .063$, $\eta_p^2 = .009$; Experiment 2b: $F(1, 160) = 1.78$, $p = .183$, $\eta_p^2 = .01$. This small effect reflects only little difficulty in reversing the preference based on the mix-up information. In contrast, the preference measured by the IAT was considerably smaller in the *invalid* condition, when the

validity information was crucial for forming the preference, compared to the *valid* condition, when validity information was in line with the co-occurrence information, Experiment 2a: $F(1, 372) = 21.31$, $p < .001$, $\eta_p^2 = .05$; Experiment 2b: $F(1, 160) = 15.50$, $p < .001$, $\eta_p^2 = .08$. That effect reflects considerable difficulty in reversing the preference based on the mix-up (validity) information, larger than the difficulty in reversing the self-reported preference.

Table 2
IAT latencies and D scores and deliberate evaluation scores in Experiments 2a-2b as a function of man's valence and validity condition.

		Automatic evaluation measure			Deliberate evaluation measure	
Experiment		Valid	Invalid		Valid	Invalid
2a (IAT) All	Compatible Block	833 (211)	868 (255)	Positive man	6.24 (1.23)	5.87 (1.53)
	Incompatible block	1021 (315)	933 (256)	Negative man	2.20 (1.38)	2.32 (1.53)
	D score	**0.36** (0.41)**	**0.16** (0.43)**	**Preference**	**4.04** (2.31)**	**3.55** (2.74)**
2a (IAT) Accurate participants	Compatible Block	791 (167)	844 (261)	Positive man	6.31 (1.14)	6.18 (1.24)
	Incompatible block	981 (271)	936 (270)	Negative man	1.90 (1.11)	2.07 (1.31)
	D score	**0.40** (0.40)**	**0.22** (0.41)**	**Preference**	**4.41** (1.98)**	**4.12** (2.35)**
2b (IAT) All	Compatible Block	803 (222)	919 (314)	Positive man	6.26 (1.20)	5.92 (1.54)
	Incompatible block	1051 (534)	1019 (316)	Negative man	1.77 (1.11)	1.94 (1.43)
	D score	**0.44** (0.34)**	**0.20** (0.41)**	**Preference**	**4.49** (2.01)**	**3.99** (2.73)**
2b (IAT) Accurate participants	Compatible Block	785 (234)	891 (311)	Positive man	6.43 (0.97)	6.00 (1.54)
	Incompatible block	1049 (581)	1019 (333)	Negative man	1.62 (0.80)	1.77 (1.32)
	D score	**0.48** (0.33)**	**0.24** (0.43)**	**Preference**	**4.82** (1.32)**	**4.23** (2.60)**

Notes. Standard deviations are in bracket; For the deliberate evaluation, **preference** is the self-reported evaluation of the negative man subtracted from the self-reported evaluation of the positive man; For the IAT, the compatible block was the block in which the positive man shared response key with positive words; Preference scores and D scores that were significantly different from zero are presented with asterisks ($^{*}p < .05$, $^{**}p < .001$).

Comprehension of the validity information. Participants' average accuracy regarding the targets' actual behaviors in the *invalid* condition (M = 2.31, SD = 1.41) was not different from the *valid* condition (M = 2.42, SD = 1.17), $t(373) = 1.31$, p = .18, d = 0.08, in Experiment 2a. In Experiment 2b, participants' average accuracy was better in the *valid* condition (M = 2.63, SD = 0.90) than in the *invalid* condition (M = 2.38, SD = 1.44), $t(161) = 2.13$, p = .03, d = 0.20. Seventy percent (N = 264) of the participants in Experiment 2a and eighty percent (N = 130) of the participants in Experiment 2b responded correctly with at least moderate confidence with regard to all four targets. The results were replicated with these accurate subsamples (Figure 2, panels B1 and B2). The preference was stronger in the *valid* than *invalid* condition, Experiment 2a: $F(1, 262) = 11.72$, p = .0007, η_p^2 = .04; Experiment 2b: $F(1, 128)$ = 11.15, p = .001, η_p^2 = .08. The preference was stronger when measured with self-report than the IAT, Experiment 2a: $F(1, 262) = 29.15$, $p < .0001$, η_p^2 = .10; Experiment 2b: $F(1, 128) = 14.51$, $p < .001$, η_p^2 = .10. Most importantly, the *measure* by *validity* interaction was of similar size to the one observed in the whole sample, Experiment 2a: $F(1, 262) = 7.88$, p = .005, η_p^2 = .02; Experiment 2b: $F(1, 128)$ = 7.76, p = .006, η_p^2 = .05. Again, self-reported preferences were only slightly weaker in the *invalid* condition compared to the *valid* condition, Experiment 2a: $F(1,262)$ = 1.22, p = .270, η_p^2 = .004; Experiment 2b: $F(1,128) = 2.61$, p =.108, η_p^2 = .019. In contrast, the preference measured by the IAT was considerably smaller in the *invalid* condition compared to the *valid* condition, Experiment 2a: $F(1, 262) = 12.28$, p = .0005, η_p^2 = .04; Experiment 2b: $F(1, 128) = 12.94$, p = .0005, η_p^2 = .09.

Unlike Experiments 1a-1b, Experiments 2a-b's results were consistent with the hypothesis that deliberate evaluation is more sensitive than automatic evaluation to explicit validity information. Nonetheless, the IAT still showed some sensitivity to the validity information by showing a preference for the truly positive man over the truly negative man even when each man co-occurred with opposite valence (Table 2). This suggests that the difference is a matter of degree rather than complete insensitivity of automatic evaluation.

The inconsistency in results between Experiments 1a-b and 2a-b could be the result of the low reliability of the EPT compared to the IAT. Alternatively, it might reflect differences in the processes and mental constructs that influence each measure (Deutsch & Gawronski, 2009; Gawronski & De Houwer, 2014). In order to generalize the results further, Experiments 3a-b used the Affective Misattribution Procedure

(AMP; Payne et al., 2005), another indirect measure that usually shows high reliability (Bar-Anan & Nosek, 2014; Gawronski & De Houwer, 2014).

Experiments 3a-3b

Method

Participants. Of the 569 participants who completed Experiment 3a, we excluded 67 participants who had more than 95% or less than 5% *pleasant* responses on the AMP (Bar-Anan & Nosek, 2014). The final sample included 502 participants (60% women, M_{age} = 34.95, SD_{age} = 12.67). Of the 192 participants who completed the relevant conditions in Experiment 3b, we excluded 16 participants who had more than 95% or less than 5% *pleasant* responses on the AMP. The final sample included 176 participants (60% women, M_{age} = 30.99, SD_{age} = 13.54).

Materials and procedure. The materials and procedure were identical to Experiments 1a-1b and 2a-2b except that we used the AMP as the automatic measure.

Automatic evaluation. Each trial of the AMP displayed stimuli in the following sequence: A photograph of one of the two targets appeared for 100 ms, followed by a blank screen for 100 ms, followed by a Chinese pictograph for 100 ms, followed by a pattern mask of black-and-white noise that appeared until participants responded. Upon presentation of the mask the participants indicated if the Chinese pictograph is more pleasant or more unpleasant than the average Chinese pictograph using two response keys on the keyboard signifying *pleasant* and *unpleasant*. Following Payne et al. (2005), the instructions informed the participants that the pictures appearing before the Chinese pictographs may bias responses and that they should try not to let the pictures influence their judgments. The task started with three example trials and then participants completed three blocks of 40 trials, each with 20 primes for each man. We computed the internal consistency from four parcels of the task, and found α = .92, .95, in Experiment 2a and 2b respectively, in the valid condition and α = .92, .94, in the invalid condition.

Design. Excluding the counterbalanced procedural parameters, the design of Experiment 3 was 2 (valid or invalid condition; between participants) X 2 (automatic or deliberate measure; within participants).

Results and Discussion

Data processing. Table 3 presents the proportions of pleasant responses after David primes and after Michael primes, in each condition in the two expeirments. We computed the AMP score for the preference of David over Michael by subtracting the

proportion of *pleasant* responses after Michael primes from the proportion of *pleasant* responses after David primes. The preference scores were standardized and then recoded to reflect the preference of the *truly* positive man over the *truly* negative man.

Evaluation. Figure 1 (panels C1 and C2) illustrates the raw and standardized preference scores in each condition in Experiments 3a and 3b. Unlike Experiments 1a and 2a, measures order did moderate the results in Experiment 3a and therefore was included in the analysis of this experiment. In Experiment 3a we submitted the standardized preference scores to a 2 (measure) x 2 (validity) x 2 (measures order) mixed ANOVA. In Experiment 3b we submitted the standardized preference scores to a 2 (measure) x 2 (validity) mixed ANOVA. These ANOVAs found a significant effect of measure, Experiment 3a: $F(1, 498) = 104.76$, $p < .0001$, $\eta_p^2 = .17$; Experiment 3b: $F(1, 174) = 48.82$, $p < .0001$, $\eta_p^2 = .21$, reflecting stronger preferences when evaluation was measured with self-report than when it was measured with the AMP. That effect may reflect stronger overall sensitivity of the self-report measure. Unlike previous experiments, in Experiment 3a there was no main effect of validity, $F(1, 498) = 2.64$, $p = .105$, $\eta_p^2 = .005$. In Experiment 3b, the preference was stronger in the valid than invalid condition, $F(1, 174) = 7.40$, $p = .007$, $\eta_p^2 = .04$. Importantly, unlike Experiments 2a-2b but similar to Experiments 1a-1b, the effect of validity was not moderated by the measure used, Experiment 3a: $F(1, 498) = 1.01$, $p = .315$, $\eta_p^2 = .002$; Experiment 3b: $F(1, 174) = 0.01$, $p = .919$, $\eta_p^2 < .001$. In other words, despite using a highly reliable measure, Experiments 3a-3b found no evidence that automatic evaluation was less sensitive to co-occurrence validity than deliberate evaluation. In Experiment 3a, we found a small moderation effect of the measures order on the interaction between validity and measure type, $F(1, 498) = 6.24$, $p = .01$, $\eta_p^2 = .01$. When the deliberate measure was first, validity and measure type did not interact, $F(1, 250) = 1.06$, $p = .303$, $\eta_p^2 < .001$. However, when the automatic measure was first, validity had a stronger effect on deliberate evaluation, $F(1, 248) = 11.90$, $p = .0007$, $\eta_p^2 = .004$ than on the AMP, $F < 1$, $\eta_p^2 < .001$. Importantly, that interaction is the opposite to the discrepancy hypothesis: it means that in that particular measures order, the AMP was slightly better than the self-report measure in reversing the preference when validity information indicated that the co-occurrence was uncharacteristic of the men.

As detailed in Table 3, like the IAT in Experiments 2a-2b, with the AMP we found reliable preference for the truly positive man over the truly negative man, even

in the invalid condition, in which validity information was essential for forming the preference. However, unlike the IAT, when measured with the AMP, that preference was not smaller than the preference formed in the valid condition, in which the validity information was not required for the preference formation, Experiment 3a: $F(1, 500) = 0.31$, $p = .580$, $\eta_p^2 < .001$; Experiment 3b: $F(1, 174) = 3.36$, $p = .068$, $\eta_p^2 = .01$. Numerically, even the self-report showed more sensitivity to the co-occurrence information, with smaller preference in the *invalid* condition than in the *valid* condition, Experiment 3a: $F(1, 500) = 7.63$, $p = .006$, $\eta_p^2 = .01$; Experiment 3b: $F(1, 174) = 10.50$, $p = .001$, $\eta_p^2 = .05$.

Table 3
AMP scores and deliberate evaluation scores in Experiments 3a-3b as a function of man's valence and validity condition.

		Automatic evaluation measure		Deliberate evaluation measure	
Experiment		Valid	Invalid	Valid	Invalid
3a (AMP) All	Positive man	0.64 (0.19)	0.63 (0.20)	6.18 (1.20)	5.79 (1.45)
	Negative man	0.51 (0.21)	0.51 (0.22)	2.08 (1.24)	2.29 (1.43)
	Preference	**0.13** (0.29)**	**0.12** (0.29)**	**4.10** (2.19)**	**3.50** (2.58)**
3a (AMP) Accurate participants	Positive man	0.64 (0.18)	0.63 (0.19)	6.45 (0.86)	6.01 (1.28)
	Negative man	0.53 (0.20)	0.52 (0.21)	1.86 (1.04)	2.05 (1.19)
	Preference	**0.12** (0.26)**	**0.12** (0.29)**	**4.60** (1.65)**	**3.97** (2.18)**
3b (AMP) All	Positive man	0.67 (0.19)	0.61 (0.23)	6.41 (1.12)	5.79 (1.59)
	Negative man	0.47 (0.24)	0.50 (0.25)	1.73 (0.98)	2.24 (1.59)
	Preference	**0.20** (0.34)**	**0.11** (0.35)**	**4.68** (1.73)**	**3.55 ** (2.76)**
3b (AMP) Accurate participants	Positive man	0.66 (0.19)	0.64 (0.22)	6.55 (0.89)	6.09 (1.31)
	Negative man	0.47 (0.24)	0.47 (0.25)	1.61 (0.81)	1.88 (1.36)
	Preference	**0.19** (0.33)**	**0.17** (0.35)**	**4.94** (1.54)**	**4.22** (2.13)**

Notes. Standard deviations are in bracket; For the deliberate evaluation, **preference** is the self-reported evaluation of the negative man subtracted from the self-reported evaluation of the positive man; For the AMP, we show proportions of *pleasant* responses for each target man in each condition, and preference scores (the difference between the proportions); Preference scores significantly different from zero are presented with asterisks ($^{*}p < .05$, $^{**}p < .001$).

Comprehension of the validity information. Participants' average accuracy regarding the actual behaviors that were performed by each of the two target men in the *invalid* condition (M = 2.52, 2.36, SD = 1.19, 1.38, in Experiments 3a and 3b respectively) was not different from the average accuracy in the *valid* condition (M = 2.44, 2.45, SD = 1.20, 1.21), $t(501) = -1.40$, $p = .106$, $d = -0.07$ in Experiment 3a, $t(175) = 0.81$, $p = .42$, $d = 0.06$ in Experiment 3b . The accurate sub-samples (Ns = 386, 128) showed the same results (Figure 2, panels C1 and C2). The analyses found a significant effect of measure, Experiment 3a: $F(1, 382) = 116.16$, $p < .0001$, $\eta_p^2 = .23$; Experiment 3b: $F(1, 126) = 35.43$, $p < .0001$, $\eta_p^2 = .21$, and, like with the entire sample, the effect of measure was not moderated by validity condition, Experiment 3a: $F(1, 382) = 0.02$, $p = .888$, $\eta_p^2 < .001$; Experiment 3b: $F(1, 126) = 0.55$, $p = .458$, $\eta_p^2 = .004$. Validity did not influence the preference measured with the AMP, Experiment 3a: $F(1, 384) = 0.00$, $p = .993$, $\eta_p^2 < .001$; Experiment 3b: $F(1, 126) = 0.04$, $p = .834$, $\eta_p^2 < .001$. The self-report showed more sensitivity to the co-occurrence information, with smaller preference in the *invalid* condition than in the *valid* condition, Experiment 3a: $F(1, 384) = 9.93$, $p = .001$, $\eta_p^2 = .02$; Experiment 3b: $F(1, 126) = 4.79$, $p = .030$, $\eta_p^2 = .03$.

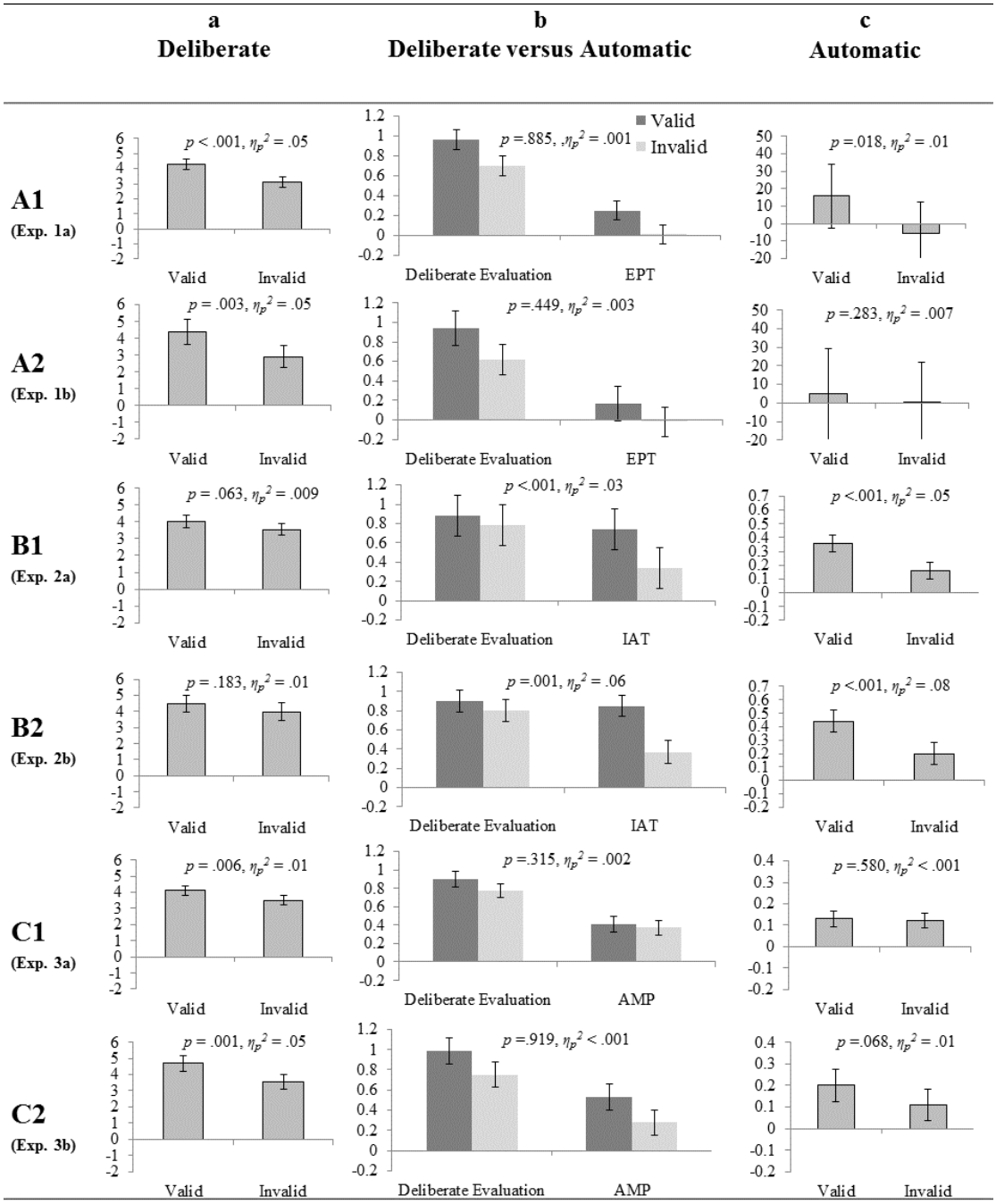

Figure 1. Experiments 1-3 (panels A-C, respectively): Preference for the truly positive man over the truly negative man, as a function of validity condition (valid versus invalid), and measure type (deliberate versus automatic). Graphs (1) present the results of the main experiments; graphs (2) present the results of the replications. Graphs (a) present deliberate evaluation, graphs (c) present automatic evaluation, and graphs (b) present comparison of standardized preferences in the deliberate versus automatic measures. p values and effects sizes in graphs (a) and (c), represent the simple effect of validity. p values and effects sizes in graphs (b), represent the validity by measure type interaction. Error-bars are 95% confidence intervals (calculated based on Jarmasz and Hollands, 2009).

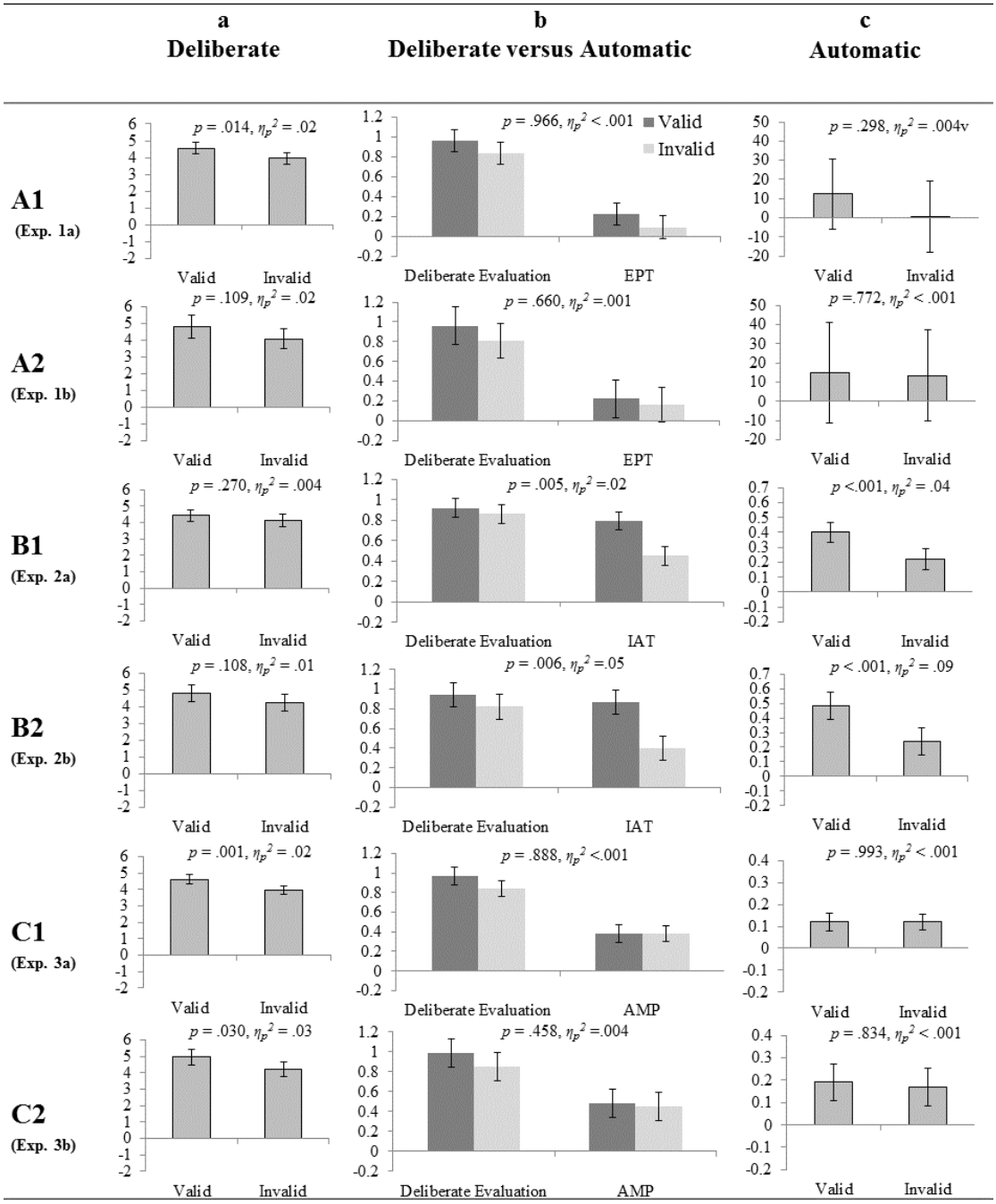

Figure 2. Results for accurate sub-samples in Experiments 1-3 (panels A-C, respectively): Preference for the truly positive man over the truly negative man, as a function of validity condition (valid versus invalid), and measure type (deliberate versus automatic). Graphs (1) present the results of the main experiments; graphs (2) present the results of the replications. Graphs (a) present deliberate evaluation, graphs (c) present automatic evaluation, and graphs (b) present comparison of standardized preferences in the deliberate versus automatic measures. p values and effects sizes in graphs (a) and (c), represent the simple effect of validity. p values and effects sizes in graphs (b), represent the validity by measure type interaction. Error-bars are 95% confidence intervals.

General Discussion

The hypothesis that deliberate evaluation is more sensitive than automatic evaluation to the validity of co-occurrence information is central to most contemporary theories of evaluation (Gawronski & Bodenhausen, 2006, 2011; Petty & Briñol, 2006; Petty et al., 2007; Rydell & McConnell, 2006; Rydell et al., 2006). Surprisingly, direct support of that hypothesis is scarce with the best existing contribution coming from experiments that provided the validity information well after all the co-occurrence information (Gregg, et al., 2006; Peters & Gawronski, 2011).

Given the importance of this hypothesis, we conducted six experiments that combined procedures from the previous experiments to manipulate co-occurrence and validity independently: four men were paired with positive or negative behaviors. Then, participants learned that the pairing was not always valid evidence that the men and the behaviors paired with each of them were the same valence. Participants learned that a man paired with positive behaviors and a man paired with negative behaviors actually performed each other's behaviors. In all six experiments, we tested whether deliberate evaluation would be more sensitive than automatic evaluation to the information about the co-occurrence validity. The key difference between the experiments was the automatic evaluation measure: EPT in Experiments 1a-1b, IAT in Experiments 2a-2b, and AMP in Experiments 3a-3b.

We observed that automatic evaluation was less sensitive than deliberate evaluation to delayed validity information with only one measure, the IAT. When automatic evaluation was measured with the EPT or with the AMP, automatic and deliberate evaluations were similarly sensitive to validity information. These results suggest both evidence consistent and inconsistent with the hypothesis depending on the measure used[7].

[7] The failure of the EPT and the AMP to support the discrepancy assumption (a null result) is probably not for lack of statistical power. The interaction effect found with the IAT was $\eta_p^2 = .03$ in Experiment 2a, and $\eta_p^2 = .06$ in Experiment 2b. As detailed in Table 4, for an effect size of $\eta_p^2 = .02$, the probability of stumbling on a Type II error in both the first experiment *and* its replication was 1% for the EPT experiments and for the AMP experiments. Therefore, it is unlikely that lack of statistical power was the reason for our failure to find reliable evidence for the discrepancy hypothesis in any of the four experiments that did not use the IAT.

Table 4
Power to detect a significant moderation effect in Experiments 1a-1b and 3a-3b

Measure	Experiment		Replication		Both	
	$\eta_p^2 = .01$	$\eta_p^2 = .02$	$\eta_p^2 = .01$	$\eta_p^2 = .02$	$\eta_p^2 = .01$	$\eta_p^2 = .02$
EPT	79%	97%	46%	75%	88%	99%
AMP	94%	99%	64%	90%	97%	99%

Notes. The replications were Experiments 1b and 3b. The *Both* column presents the chances to detect the moderation effect in at least one of the two experiments.

Explanations for the Results

To improve the ability to generalize from our results, we used multiple indirect measures. Similar results across the three measures could have served as reliable evidence regarding the effect of delayed validity information on automatic evaluation—the theoretical construct that presumably influences all these measures. However, we found different results from different measures. These results suggest no strong evidence that automatic evaluation is less sensitive to validity information than deliberate evaluation. The present research raises doubts about a central hypothesis in contemporary evaluation theories, and points to critical limitations of the current understanding of indirect evaluation measures. In the following analysis we consider alternative explanations for the results, each with important theoretical and methodological implications.

No discrepant sensitivity. Previous research found that when explicit information about the validity of the co-occurring affective stimuli is presented before or immediately after presenting the co-occurrence, validity information has the same effect on automatic and deliberate evaluation (Boucher & Rydell, 2012; Peters & Gawronski, 2011; Siegel, et al., 2012). If delayed validity information also has the same influence on automatic and deliberate evaluation, the only relevant evidence left for different effects of evaluative learning on automatic versus deliberate evaluation would come from studies that did not explicitly inform participants that the co-occurrence information is not valid (e.g., Moran & Bar-Anan, 2013).

If the present EPT and AMP results are a better reflection of automatic evaluation than the IAT results, then most contemporary evaluation models require revision to explain why explicit validity information has the same effect on deliberate and automatic evaluation. One possibility could be that people are generally good in

using validity information to form new evaluative associations. Support for that possibility comes from the finding that both automatic and deliberate evaluation are reversed after people receive new information that changes the meaning of earlier information (Mann & Ferguson, 2015; Wyer, 2010). Perhaps only when the validity information is not explicit (e.g., Moran & Bar-Anan, 2013; Rydell et al., 2006), validity is not used immediately for revising evaluations, new evaluative associations are not formed, and automatic/deliberate discrepancy emerges.

Confirmation of the discrepant sensitivity assumption. A second alternative is that the IAT results reflect true discrepancy between deliberate and automatic evaluative processes in their sensitivity to invalid co-occurrence. In that case, contemporary evaluation theories are correct in their hypothesis regarding automatic/deliberate discrepancies, but better theories about attitude measurements are required to explain why the EPT and the AMP failed to show the discrepancy.

Previous research had already suggested that task-specific mechanisms can influence the effects revealed by different indirect measures of evaluation (Deutsch et al., 2009; Deutsch & Gawronski, 2009; Gawronski et al., 2010). For example, different effects may be observed between indirect measures that are based on response-interference (RI) mechanism (like the EPT and the IAT) versus indirect measures that are based on misattribution (like the AMP; Deutsch & Gawronski, 2009). Gawronski et al. (2010) found that effects on the EPT were influenced from participants' attention to the category membership of the primes, while the effects on the AMP were not. Gawronski et al. argued that unlike measures that are based on RI, measures that are based on misattribution might integrate evaluative information from multiple sources. In the context of the present study, this might explain why the automatic evaluations measured with the AMP showed sensitivity to validity information more than the automatic evaluations measured with the IAT. On the other hand, the EPT that is also based on RI did not show similar pattern as the IAT. The difference between the IAT and EPT results could be due to the IAT's superior reliability.

Another possibility is that the AMP is more sensitive than the IAT to deliberate evaluative processes. In a comparison of seven indirect measures, the AMP was the only indirect measure not related to any of the other indirect measures more strongly than to self-report measures (Bar-Anan & Nosek, 2014). The AMP also showed evidence that it is sensitive to evaluative processes only when participants

reported that the priming effect was intentional (Bar-Anan & Nosek, 2012). On the other hand, these effects are not direct evidence that the AMP is more sensitive to deliberate evaluation than the IAT. First, Bar-Anan and Nosek (2012) did not investigate whether the IAT shows similar pattern of results. Second, people's retrospective reports that they intentionally influenced the AMP score do not prove that this indeed was the case (Bar-Anan & Nosek, 2012; Payne et al., 2013). Finally, there is evidence that the AMP is sensitive to automatic rather than deliberate evaluation (e.g., Gawronski & Ye, 2015; for a review, Cameron, Brown-Iannuzzi, & Payne, 2012).

Nuanced discrepant sensitivity. Another possible reason for discrepancy between the different indirect measures could be that these measures tap different aspects of automatic evaluation (Gawronski & De Houwer, 2014). Perhaps all three measures are sensitive to automatic evaluative processes, but not to the same processes.

One relevant difference between the three measures is that the IAT's different conditions are manipulated between blocks while they are intermixed within blocks in the EPT and the AMP. Unlike the EPT and the AMP, the IAT compares performance under task rules that are constant in different blocks of the task. Therefore, participants' performance in the IAT can improve if they focus on associations that fit the task rules in each block (De Houwer et al., 2005; Rothermund et al., 2005). In the present case, participants could recode the man categories according to co-occurrence (*Michael* is recoded to *good* because he co-occurred with positive behaviors) or true valence (*Michael* is recoded to *bad* because the validity information suggested that he is a negative person). In other words, perhaps both co-occurring valence and true valence contribute to automatic evaluation, but in cases that the two are in conflict, only the IAT (but not the AMP or the EPT) can capture the contribution of co-occurrence to automatic evaluation.

Sensitivity of Evaluation to Validity Information

Another important finding of the present investigation is that nearly all the measures in all the experiments and all samples showed a weaker preference when participants had to reverse the evaluation implied by the co-occurrence (e.g., the invalid condition) than when the co-occurrence was valid evaluative information. Importantly, deliberate evaluation consistently showed sensitivity to the invalid co-occurrence. Further, the IAT and the AMP showed sensitivity to validity information:

a reliable preference for the man who co-occurrence with uncharacteristic negative behaviors over the man who co-occurred with uncharacteristic positive behavior.

Thus, although the present investigation leaves open questions regarding the effect of validity on automatic versus deliberate evaluation, the results clearly show that deliberate evaluation is sensitive to invalid co-occurrence, whereas automatic evaluation is sensitive to validity information. In light of these findings, evaluation theories can no longer assume that deliberate evaluation *completely* ignores associations if they are based on co-occurrence with valence that is explicitly known to be the opposite of the target's valence (for more on that issue, see Moran, Bar-Anan, & Nosek, in press). And, evaluation theories also can no longer assume that validity information does not influence automatic evaluation. Our research suggests that if there is a difference between the sensitivity of automatic and deliberate evaluation to invalid co-occurrence, it is probably a matter of degree rather than an all-or-none relationship.

Future Directions

The present results identify productive new questions to investigate regarding the effect of validity information on automatic and deliberate evaluation. One is to test this question with other learning procedures to clarify moderators of sensitivity of automatic and deliberate evaluation to validity information and to co-occurrence. One factor might be participants' processing goals during learning. In the present research, the instructions guided the participants to memorize the information *and* to form impressions of the men. If participants have only one of these goals, the effect of invalid co-occurrence might change. For example, Moran, Bar-Anan, and Nosek (2015) directly manipulated processing goals and used a different learning procedure. They found that an impression formation goal increased the sensitivity of evaluation to validity information, whereas a memorization goal decreased the sensitivity of evaluation to validity information. This occurred for both automatic and deliberate evaluation. There was no discrepancy between automatic and deliberate evaluation in their sensitivity to the processing goal moderator. However, these results were obtained in a learning procedure that did not provide *explicit* validity information. Therefore, whether that pattern replicates when validity information is provided explicitly, after a delay, is still unknown.

Another possible moderator is the explicitness or directness of the validity information. The present research focused on explicit information whether the co-

occurrence is valid evidence of similarity between the co-occurring stimuli. Previous research that found discrepant sensitivity of automatic and deliberate evaluation to invalid co-occurrence did not use such explicit instructions (e.g., Moran & Bar-Anan, 2013; Rydell et al., 2006). Therefore, an important next step is to manipulate the explicitness of the validity information. This will show whether the sensitivity of deliberate versus automatic evaluation becomes more discrepant when the instructions are less explicit.

An unresolved mystery of the present research is why the indirect measures elicited different results. If, for example, the reason that the IAT showed more sensitivity than the other measures to invalid co-occurrence is that the IAT depends on recoding of categories, then other measures that are insensitive to recoding, such as the Sorting Paired Feature task (Bar-Anan, Nosek & Vianello, 2009) and the recoding-free IAT (Rothermund et al., 2009) should show results similar to those found in the present research with the EPT and the AMP. And, other measures that are sensitive to recoding, like the Brief IAT (Sriram & Greenwald, 2009), and the Go-No-Go Association Test (Nosek & Banaji, 2001), should show results similar to those we found with the IAT.

Finally, perhaps the most important future research direction is to measure behavioral outcomes that are known to reflect automatic evaluation. If co-occurrence without validity shows stronger influence on a behavior known to reflect automatic evaluation than on behavior known to reflect deliberate evaluation, that would provide strong support for the theoretical assumption tested in the present research. Although many studies have found that some behaviors are related to automatic evaluation measures more than to self-reported measures (e.g., Dovidio et al., 2002; Friese et al., 2009), there are hardly any evaluative learning studies that used such behaviors as a measure for discrepancy between the formation of automatic versus deliberate evaluation (for a rare exception, see Rydell and McConnell, 2006, Experiment 4). From our experience, behavioral measures are difficult to administrate and often suffer from low reliability and validity. Nevertheless, such measures are essential for corroborating claims of discrepant formation of deliberate versus automatic evaluation (Gawronski & De Houwer, 2014).

Summary

The present research tested the assumption that deliberate evaluation is more sensitive than automatic evaluation to the validity of co-occurrence information. We found support for this assumption only when automatic evaluation was measured with the Implicit Association Test. The present results cast doubt on the discrepant sensitivity hypothesis, and call for more research to confirm (or dispute) this central hypothesis. The results emphasize the importance of advancing current knowledge about the specific evaluative constructs that influence each indirect measure. In fact, there is hardly any evaluative learning research that compared the sensitivity of different indirect measures to specific evaluative learning histories. The results of the present research also emphasize the need of using multiple automatic evaluation and behavioral measures in the investigation of the formation of deliberate versus automatic evaluation.

References

Bar-Anan, Y. (2010). Strategic modification of the evaluative priming effect does not reduce its sensitivity to uncontrolled evaluations. *Journal of Experimental Social Psychology*, *46*(6), 1101-1104.

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*(9), 1194-1208.

Bar-Anan, Y., & Nosek, B. A. (2014). A comparative investigation of seven implicit measures of social cognition. *Behavior Research Methods, 46,* 668-688.

Bar-Anan, Y., Nosek, B. A., & Vianello, M. (2009). The sorting paired features task: A measure of association strengths. *Experimental psychology*, *56*(5), 329-343.

Boucher, K. L., & Rydell, R. J. (2012). Impact of negation salience and cognitive resources on negation during attitude formation. *Personality and Social Psychology Bulletin, 38*(10), 1329-1342.

Buchner, A., & Wippich, W. (2000). On the reliability of implicit and explicit memory measures. *Cognitive Psychology*, *40*(3), 227-259.

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*(4), 330-350.

De Houwer, J., Geldof, T., & De Bruycker, E. (2005).The Implicit Association Test as a general measure of similarity. *Canadian Journal of Experimental Psychology, 59*(4), 228–239.

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*(1), 101-114.

Deutsch, R., Kordts-Freudinger, R., Gawronski, B., & Strack, F. (2009). Fast and fragile: A new look at the automaticity of negation processing.*Experimental psychology*, *56*(6), 434-446.

Dovidio, J. F., Kawakami, K., & Gaertner, S. L. (2002). Implicit and explicit prejudice and interracial interaction. *Journal of personality and social psychology, 82*(1), 62-68.

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*(6), 1013-1027.

Friese, M., Hofmann, W., & Schmitt, M. (2009). When and why do implicit measures predict behaviour? Empirical evidence for the moderating role of opportunity, motivation, and process reliance. *European Review of Social Psychology, 19*(1), 285-338

Gawronski, B., & Bodenhausen, G. V. (2006). Associative and prepositional processes in evaluation: An integrative review of implicit and explicit attitude change. *Psychological Bulletin, 132*(5), 692-731.

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., Cunningham, W. A., LeBel, E. P., & Deutsch, R. (2010). Attentional influences on affective priming: Does categorisation influence spontaneous evaluations of multiply categorisable objects?. *Cognition and Emotion*, *24*(6), 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). New York, NY: Cambridge University Press.

Gawronski, B., & Ye, Y. (2015). Prevention of intention invention in the affect misattribution procedure. *Social Psychological and Personality Science, 6*(1), 101-108.

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), 1-20.

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*(6), 1464-1480.

Greenwald, A. G., Nosek, B. A., & Banaji, M. R. (2003). Understanding and using the Implicit Association Test: I. An improved scoring algorithm. *Journal of Personality and Social Psychology, 85*(2), 197-216

Jarmasz, J., & Hollands, J. G. (2009). Confidence intervals in repeated-measures designs: The number of observations principle. *Canadian Journal of Experimental Psychology, 63*(2), 124-138.

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

Minear, M., & Park, D. C. (2004). A lifespan database of adult facial stimuli. *Behavior Research Methods, Instruments, & Computers, 36*(4), 630-633.

Moran, T., & Bar-Anan, Y. (2013). The effect of object–valence relations on automatic evaluation. *Cognition & emotion, 27*(4), 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.

Moran, T., Bar Anan, Y., Nosek, B. A., (in press). The Assimilative Effect of Co-occurrence on Evaluation Above and Beyond the Effect of Relational Qualifiers. *Social Cognition*.

Nosek, B. A. (2005). Moderators of the relationship between implicit and explicit evaluation. *Journal of Experimental Psychology: General, 134*(4), 565-584.

Nosek, B. A., & Banaji, M. R. (2001). The go/no-go association task. *Social cognition, 19*(6), 625-666.

Nosek, B. A., Greenwald, A. G., & Banaji, M. R. (2005). Understanding and using the Implicit Association Test: II. Method variables and construct validity. *Personality and Social Psychology Bulletin, 31*(2), 166-180.

Nosek, B. A., Greenwald, A. G., & Banaji, M. R. (2007). The Implicit Association Test at age 7: A methodological and conceptual review. In J. A. Bargh (Ed.), *Automatic processes in social thinking and behavior* (pp. 265–292). New York: Psychology Press.

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*(1), 36-88.

Olson, M. A., & Fazio, R. H. (2003). Relations Between Implicit Measures of Prejudice What Are We Measuring? *Psychological Science, 14*(6), 636-639.

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*(3), 375-386.

Payne, B. K., Cheng, C. M., Govorun, O., & Stewart, B. D. (2005). An inkblot for attitudes: affect misattribution as implicit measurement. *Journal of personality and social psychology*, *89*(3), 277-293.

Peters, K. R., & Gawronski, B. (2011). Are We Puppets on a String? Comparing the Impact of Contingency and Validity on Implicit and Explicit Evaluations. *Personality and Social Psychology Bulletin*, *37*(4), 557-569.

Petty, R. E., & Briñol, P. (2006). A meta-cognitive approach to "implicit" and "explicit" evaluations: Comment on Gawronski and Bodenhausen (2006). *Psychological Bulletin*, *132*, 740-744.

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*(5), 657-686.

Petty, R. E., Tormala, Z. L., Briñol, P., & Jarvis, W. B. G. (2006). Implicit ambivalence from attitude change: An exploration of the PAST model. *Journal of Personality and Social Psychology*, *90*(1), 21–41.

Prestwich, A., Perugini, M., Hurling, R., & Richetin, J. (2010). Using the self to change implicit attitudes. *European Journal of Social Psychology*, *40*(1), 61-71.

Ranganath, K. A. & Nosek, B. A. (2008). Implicit attitude generalization occurs immediately, explicit attitude generalization takes time. *Psychological Science*, *19*(3), 249-254.

Ratliff, K. A., & Nosek, B. A. (2010). Creating distinct implicit and explicit attitudes with an illusory correlation paradigm. *Journal of Experimental Social Psychology*, *46*(5), 721-728.

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*(1), 84-98.

Rothermund, K., Wentura, D., & De Houwer, J. (2005).Validity of the salience asymmetry account of the Implicit Association Test: Reply to Greenwald,

Nosek, Banaji, and Klauer (2005). *Journal of Experimental Psychology: General, 134*, 426–430.

Rydell, R. J., & McConnell, A. R. (2006). Understanding implicit and explicit attitude change: A systems of reasoning analysis. *Journal of Personality and Social Psychology, 91*(6), 995–1008.

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*(11), 954-958.

Siegel, E., Sigall, H., & Huber, D. E. (2012). The IAT is sensitive to the perceived accuracy of newly learned associations. *European Journal of Social Psychology*, *42*(2), 189-199.

Sriram, N., & Greenwald, A. G. (2009). The brief implicit association test. *Experimental psychology*, *56*(4), 283-294

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. 95-115). New York, NY: Guilford.

Whitfield, M., & Jordan, C. H., (2009). Mutual Influence of implicit and explicit attitudes. *Journal of experimental social psychology*, *45*(4), 748-759.

Wyer, N. A. (2010). You never get a second chance to make a first (implicit) impression: The role of elaboration in the formation and revision of implicit impressions. *Social Cognition, 28*, 1–19.