Psychological Bulletin
2009, Vol. 135, No. 3, 369–372

0033-2909/09/$12.00 DOI: 10.1037/a0014820

Methodological Issues in the Validation of Implicit Measures: Comment on De Houwer, Teige-Mocigemba, Spruyt, and Moors (2009)

Bertram Gawronski, Etienne P. LeBel, and Kurt R. Peters
University of Western Ontario

Rainer Banse
University of Bonn

J. De Houwer, S. Teige-Mocigemba, A. Spruyt, and A. Moors's (2009) normative analysis of implicit measures provides an excellent clarification of several conceptual ambiguities surrounding the validation and use of implicit measures. The current comment discusses an important, yet unacknowledged, implication of J. De Houwer et al.'s analysis, namely, that investigations addressing the proposed implicitness criterion (i.e., does the relevant psychological attribute influence measurement outcomes in an automatic fashion?) will be susceptible to fundamental misinterpretations if they are conducted independently of the proposed what criterion (i.e., is the measurement outcome causally produced by the psychological attribute the measurement procedure was designed to assess?). As a solution, it is proposed that experimental validation studies should be combined with a correlational approach in order to determine whether a given manipulation influenced measurement scores via variations in the relevant psychological attribute or via secondary sources of systematic variance.

Keywords: automaticity, control, implicit measures, validity

Despite the overwhelming impact of implicit measures on almost all subdisciplines of psychology, there is still a lot of conceptual confusion associated with their use. Common questions raised in this context are as follows: What exactly is "implicit" about implicit measures? Is it something about the measurement procedure or something about the psychological constructs they assess? How do we know that a measure is implicit? And what exactly does it mean to say that a measure is implicit? De Houwer, Teige-Mocigemba, Spruyt, and Moors's (2009) normative analysis provides an excellent framework that clarifies the conceptual quagmire surrounding the meaning of the term *implicit measure*. In addition, their analysis gives useful directions for future research by highlighting several gaps in our current knowledge about implicit measures.

Nevertheless, we believe that De Houwer et al. (2009) overlooked an important implication of their own conceptual framework that can have serious consequences when it comes to establishing the construct validity of implicit measures. Specifically, we argue that their implicitness criterion (i.e., does the relevant psychological attribute influence measurement outcomes in an automatic fashion?) cannot be separated from an important aspect of their what criterion (i.e., is the measurement outcome causally produced by the psychological attribute the measurement procedure was designed to assess?). On the basis of the respective implications of each of the two criteria, we argue that investigations addressing the implicitness criterion can be susceptible to fundamental misinterpretations if they are conducted independently of the what criterion. As the two criteria are closely linked to theorizing about the psychological attributes implicit measures are designed to reflect (e.g., Gawronski & Bodenhausen, 2006), such misinterpretations have the potential to distort these theories in a significant manner. These claims should not be interpreted as a criticism of De Houwer et al.'s normative analysis. To the contrary, we believe that our discussion further highlights the significance of De Houwer et al.'s approach by identifying an important implication that has not been spelled out in their original analysis.

Three Validity Criteria of Implicit Measures

De Houwer et al. (2009) defined implicit measures as outcomes of measurement procedures that are caused in an automatic fashion by the psychological attribute or construct the measurement procedure was designed to assess. Drawing on Borsboom, Mellenbergh, and van Heerden's (2004) conceptualization of construct validity (see also Borsboom, 2006), De Houwer et al. further proposed three criteria to establish that a measurement outcome is indeed an implicit measure. First, the measurement outcome should be causally produced by the psychological attribute the measurement procedure was designed to assess (what criterion). Second, one needs to examine the processes by which the psychological attribute causes variations in the measurement outcome (how criterion). Third, one needs to examine whether these processes operate in an automatic fashion (implicitness criterion). The third criterion requires further specification of the particular sense of the term *automatic* in regard to which the psychological at-

Bertram Gawronski, Etienne P. LeBel, and Kurt R. Peters, Department of Psychology, University of Western Ontario, London, Ontario, Canada; Rainer Banse, Department of Psychology, University of Bonn, Bonn, Germany.

Preparation of this article has been supported by Canada Research Chairs Program Grant 202555 and Social Sciences and Humanities Research Council of Canada Grant 410-2008-2247.

Correspondence concerning this article should be addressed to Bertram Gawronski, Department of Psychology, University of Western Ontario, Social Science Centre, London, Ontario N6A5C2, Canada. E-mail: bgawrons@uwo.ca

tribute is thought to influence the measurement outcome. That is, are the relevant processes unintentional, unconscious, efficient, and/or uncontrollable (Bargh, 1994; Moors & De Houwer, 2006)?

In their discussion of the what criterion, De Houwer et al. (2009) point to an important ambiguity in the interpretation of this criterion. On the one hand, the criterion could be interpreted as a requirement that variations in the psychological attribute indeed produce corresponding variations in the measurement outcome. If variations in the psychological attribute do not produce corresponding variations in the measurement outcome, it would be ill-founded to call the measure a measure of this particular attribute. On the other hand, the what criterion could also be interpreted as requiring that any variation in the measurement outcome is uniquely produced by variations in the psychological attribute it was designed to reflect. This interpretation points to the significance of systematic error or confounded variance, namely, variance stemming from sources other than the psychological attribute the measurement procedure has been designed to assess (also described as contamination). Even though it seems desirable to have measures that do not include any error variance—systematic or nonsystematic—it seems uncontroversial that this request is a normative ideal that is virtually impossible to achieve for any measure that is currently used in psychology.

Still, as correctly pointed out by De Houwer et al. (2009), the role of systematic error can lead to problems in the interpretation of empirical results, if variations that are due to factors other than the to-be-measured psychological construct are misattributed to that construct. For instance, if implicit measures are used as independent variables (e.g., prediction of behavior), the implicit measure and the dependent measure may show significant correlations because they share sources of systematic error variance, and this may be true even if there is no relation between the dependent measure and the psychological attribute the measurement procedure was designed to assess (see Mierke & Klauer, 2003). In such cases, the shared variance between the implicit measure and the dependent variable reflects a contamination or confounding in the implicit measure, not the construct of interest. Moreover, if implicit measures are used as dependent variables in experimental designs (e.g., in studies of attitude change), the employed manipulations may influence measurement scores via sources of systematic error rather than the psychological attribute the measurement procedure was designed to assess (see Deutsch & Gawronski, 2009). Hence, whenever one uses an implicit measure (or any other measure) as an independent or a dependent measure, it is important to consider sources of systematic variance that do not reflect the to-be-measured psychological attribute (Gawronski, Deutsch, LeBel, & Peters, 2008). Needless to say, if effects that are driven by alternative sources of variance are misattributed to the psychological construct the measurement procedure was designed to assess, theorizing about that construct can be seriously distorted.

Studying the Implicitness of Implicit Measures

These issues have important implications for the study of automatic processes in implicit measures. As noted by De Houwer et al. (2009), the question of whether the psychological attribute a measurement procedure was designed to assess can indeed be described by features of automaticity is an empirical question that should be addressed as such (see also De Houwer, 2006; Gawronski, Hofmann, & Wilbur, 2006). Investigations of this question, including those reviewed by De Houwer et al., typically manipulate the processing conditions under which the measure is administered and then compare the resulting measurement scores under the respective conditions. For instance, to investigate whether evaluative responses assessed by sequential priming tasks (e.g., Fazio, Jackson, Dunton, & Williams, 1995) are independent of evaluative processing goals, researchers systematically investigated whether priming scores obtained with these measures differ as a function of whether an evaluative processing goal is present or absent. As discussed by De Houwer et al., the evidence regarding this particular question is somewhat mixed, in that some studies obtained evidence for goal dependence (e.g., De Houwer, Hermans, Rothermund, & Wentura, 2002) whereas others found evidence for goal independence (e.g., Bargh, Chaiken, Raymond, & Hymes, 1996). Similar studies have been conducted to test the resource dependency (e.g., Klauer & Teige-Mocigemba, 2007) and controllability (e.g., Degner, in press; Teige-Mocigemba & Klauer, 2008) of evaluative priming effects, as well as their dependency on attentional processes (e.g., Gawronski, Cunningham, LeBel, & Deutsch, 2008; Musch & Klauer, 2001; Simmons & Prentice, 2006).

As may be evident from our preceding discussion, and as explicitly noted by De Houwer et al. (2009), the data of such experimental studies should be interpreted with caution, as experimental manipulations may influence measurement outcomes not only via the psychological attribute a measurement procedure was designed to assess but also via alternative sources of systematic variance. For instance, if variations in the outcome of a measurement procedure are caused by the relevant attribute in an automatic fashion, but variations driven by contaminating confounds are caused in a controlled fashion (see Conrey, Sherman, Gawronski, Hugenberg, & Groom, 2005), experiments designed to test features of automaticity will likely produce differences between experimental conditions. However, in such cases, the obtained differences in measurement scores reflect the controlled nature of the contaminating confounds, not the controlled nature of the psychological attribute (for a review of examples, see Sherman et al., 2008). In other words, when one investigates the implicitness of implicit measures, it does not suffice to simply manipulate the conditions of automatic and controlled processing and then infer that the psychological attribute influenced the measurement outcome in an automatic or controlled fashion. Rather, any such investigation must take the lack of process purity of implicit measures into account (see Conrey et al., 2005; Klauer, Voss, Schmitz, & Teige-Mocigemba, 2007; Payne, 2008) when inferences are being made about automatic and controlled features of the psychological attribute a measurement procedure was designed to assess.

A useful example to illustrate the significance of this issue is an unpublished study by Schmitz, Teige, Voss, and Klauer (2005) reviewed by De Houwer et al. (2009). Testing the effects of working memory load on measurement scores of the Implicit Association Test (IAT; Greenwald, McGhee, & Schwartz, 1998), Schmitz et al. found that working memory load influenced the mean values of IAT scores across experimental conditions. However, the size of external correlations to a criterion measure (i.e., self-reported attitudes) was unaffected by the working memory

load manipulation. These results suggest that working memory load most likely influenced a source of systematic variance that is unrelated to the psychological attribute the measurement procedure was designed to assess (in this case, attitudes). A somewhat different result was obtained by Simmons and Prentice (2006), who investigated the impact of attentional processes on affective priming effects (see Fazio et al., 1995). These researchers found that lack of attention to the prime stimuli influenced not only the mean values of measurement scores but also correlations to a criterion measure (i.e., self-reported attitudes). Thus, whereas Schmitz et al.'s findings most likely reflect an effect on secondary sources of systematic variance in the IAT (see also Conrey et al., 2005), the findings by Simmons and Prentice (2006) seem to reflect a genuine effect that is driven by the psychological attribute the measurement procedure was designed to assess (but see Gawronski et al., 2008, for an alternative interpretation). Taken together, these considerations suggest that experimentally produced differences in the mean values of measurement outcomes do not necessarily indicate that the relevant attribute has influenced measurement scores in an automatic or controlled fashion, as long as effects on alternative sources of variance cannot be ruled out.

Experimental Versus Correlational Approaches to Validation

These insights are not only relevant for studies on the implicitness of implicit measures; they also highlight a broader issue in the validation of implicit measures. Borsboom et al. (2004) correctly pointed out that in order to establish the construct validity of a new measure, one should investigate the causal mechanisms of how variations in the relevant attribute produce variations in the measurement outcome. In emphasizing their experimental approach to construct validity, Borsboom et al. criticized correlational approaches as being incapable of establishing the causal link between psychological attributes and measurement outcomes and cited well-known shortcomings of correlational designs, such as the third-variable problem. In this spirit, Borsboom et al. rejected Cronbach and Meehl's (1955) notion of nomological networks, in which construct validity is determined by means of conceptual and empirical links to established theories and measures, typically in the form of correlations to other measures and outcomes. Instead, construct validity should be established with experimental designs by investigating the causal link between psychological attributes and measurement outcomes. Even though De Houwer et al. (2009) seem to agree with Borsboom et al. (2004) about the limitations of the correlational approach, they are more cautious in their evaluation and emphasize that both experimental and correlational approaches provide insights into the validity of a given measure.

We argue that the most fruitful approach to the validation of implicit measures is to combine experimental and correlational approaches within the same study. As implied by the abovementioned examples of working memory capacity (Schmitz et al., 2005) and attentional influences (Simmons & Prentice, 2006), a purely experimental approach may be susceptible to misinterpretations as long as it cannot establish whether an obtained experimental effect reflects variations in the relevant psychological attribute or variations in contaminating sources of systematic variance that are confounded in the measurement score. This ambiguity could be resolved by investigating correlations to a criterion measure in the different conditions of an experimental design.[1] If an experimental manipulation influences not only the mean values of the measurement score but also its correlation to an accepted criterion variable (e.g., Simmons & Prentice, 2006), there is reason to believe that the experimental manipulation indeed affected the relevant psychological attribute.[2] If, however, correlations to the criterion variable are unaffected by the experimental manipulation (e.g., Schmitz et al., 2005), it seems more plausible that the experimental manipulation influenced measurement scores via alternative sources of variance rather than the psychological attribute the measure was designed to reflect.[3]

Such a combined approach to validation captures the main argument by Borsboom et al. (2004) that construct validity should be established by experimentally investigating how variations in a psychological attribute causally produce variations in measurement outcomes. However, it goes beyond the limitations of Borsboom et al.'s approach, in that it links such investigations to the overall nomological network of theoretical and empirical assumptions about the psychological attribute one aims to assess. We believe that De Houwer et al.'s (2009) normative approach, in conjunction with the combined approach proposed in the present comment, provides a more fine-grained validation of implicit measures that reduces the risk of conceptual and empirical misinterpretations.

---

[1] Alternative approaches include the use of mathematical procedures designed to quantify the relative contributions of qualitatively distinct processes within a given measure (e.g., Conrey et al., 2005; Klauer et al., 2007; Payne, 2008). However, these procedures are limited in their applicability, in that they are typically designed for only a particular class of measurement procedures. In this context, it is also important to note that the use of these procedures does not guarantee that the processes reflected by their parameters are automatic or controlled. Instead, these parameters reflect qualitatively distinct processes that may or may not be characterized by features of automaticity (see Sherman et al., 2008). The latter issue is an empirical question that needs to be tested in the same manner described by De Houwer et al. (2009) in the context of their implicitness criterion.

[2] Note that there are at least three cases under which correlations to a criterion measure may differ across experimental conditions even when the employed manipulation influenced measurement scores via secondary sources of variance. First, correlations may differ across conditions when the experimental manipulation influenced the reliability or sensitivity of the measure with respect to the psychological attribute it is purported to reflect (e.g., Gawronski et al., 2008; see also Olson & Fazio, 2003). Second, correlations may differ when the manipulation differentially influenced the within-condition variability of measurement scores across experimental conditions (i.e., lower variance in one condition than another). Third, correlations should differ across conditions when the experimental manipulation influenced measurement scores via a secondary source of variance in the implicit measure that is shared by the criterion variable. These possibilities should be ruled out with supplementary statistical and conceptual analyses.

[3] As a caveat, it is important to note that a lack of difference between correlations may be obtained despite a genuine impact on to-be-measured attribute and sufficient statistical power if (a) within-condition variability is very small (e.g., if all participants can be expected to show either high or low scores on either the implicit or the criterion measure) or (b) the reliability of the measure is low to begin with. Again, these possibilities should be ruled out with supplementary analyses.

References

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

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.

Borsboom, D. (2006). The attack of the psychometricians. *Psychometrika, 71,* 425–440.

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review, 111,* 1061–1071.

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.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52,* 281–302.

Degner, J. (in press). On the (un)controllability of affective priming: Strategic manipulation is feasible but can possibly be prevented. *Cognition and Emotion.*

De Houwer, J. (2006). What are implicit measures and why are we using them? In R. W. Wiers & A. W. Stacy (Eds.), *The handbook of implicit cognition and addiction* (pp. 11–28). Thousand Oaks, CA: Sage.

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

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

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

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.

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. (2008). *Affective priming as a measure of implicit preferences: Reliability depends on attention to relevant features in response interference tasks.* Manuscript submitted for publication.

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

Gawronski, B., Hofmann, W., & Wilbur, C. J. (2006). Are "implicit" attitudes unconscious? *Consciousness and Cognition, 15,* 485–499.

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.

Klauer, K. C., & Teige-Mocigemba, S. (2007). Controllability and resource dependence in automatic evaluation. *Journal of Experimental Social Psychology, 43,* 648–655.

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

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

Moors, A., & De Houwer, J. (2006). Automaticity: A theoretical and conceptual analysis. *Psychological Bulletin, 132,* 297–326.

Musch, J., & Klauer, K. C. (2001). Locational uncertainty moderates affective priming effects in the evaluative decision task. *Cognition and Emotion, 15,* 167–188.

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

Payne, B. K. (2008). What mistakes disclose: A process dissociation approach to automatic and controlled processes in social psychology. *Social and Personality Psychology Compass, 2,* 1073–1092.

Schmitz, F., Teige, S., Voss, A., & Klauer, K. C. (2005, June). *Working memory load in the IAT.* Paper presented at the Fifth Workshop on Implicit Representations and Personality, Berlin, Germany.

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

Simmons, J. P., & Prentice, D. A. (2006). Pay attention! Attention to the primes increases attitude assessment accuracy. *Journal of Experimental Social Psychology, 42,* 784–791.

Teige-Mocigemba, S., & Klauer, K. C. (2008). "Automatic" evaluation? Strategic effects on affective priming. *Journal of Experimental Social Psychology, 44,* 1414–1417.

Received October 3, 2008
Accepted October 15, 2008 ■