RUNNING HEAD: COMPARING MEASURES OF ATTITUDES

Comparing measures of attitudes at the functional and procedural level: Analysis and implications

Jan De Houwer

Ghent University, Ghent, Belgium

De Houwer, J. (in press). Comparing measures of attitudes at the functional level and procedural level: Analysis and implications. In R. Petty, R. H. Fazio, & P. Brinol (Eds.), *Attitudes: Insights from the new implicit measures*. Erlbaum.

mailing address: Jan De Houwer
Department of Psychology
Ghent University
Henri Dunantlaan 2
B-9000 Ghent
Belgium
email: Jan.DeHouwer@UGent.be

phone: 0032 9 264 64 45
fax: 0032 9 264 64 89

During the past decade, many new measures of attitudes have been proposed, several of which received the label "implicit measure". The most commonly known implicit measures are probably the affective priming task (e.g., Fazio, Jackson, Dunton, & Williams, 1995), the Implicit Association Test (IAT; e.g., Greenwald, McGhee, & Schwartz, 1998), the (extrinsic) affective Simon task (e.g., De Houwer, 2003a; De Houwer & Eelen, 1998), and the Go-Nogo-Association Task (GNAT; Nosek & Banaji, 2001) (see Fazio & Olson, 2003, for a review). Given this increase in the number of available measures, there is a risk of not being able to see the proverbial forest through the proverbial trees. Although each of the measures is assumed to reveal attitudes, it is often difficult to understand how the different measures are related to each other, whether they can be expected to converge, and if so, under which conditions. There is thus a need for criteria that can be used to compare and describe the different measures.

In this chapter, I will discuss two levels at which measures can be compared. As I recently pointed out in another chapter (De Houwer, 2006), it is crucial to realize that the term "measure" can be used in different ways. It can be used either to refer to the outcome of a measurement procedure (e.g., a particular score on a questionnaire or a particular pattern of reaction time performance such as an IAT effect) or to the objective measurement procedure itself (e.g., the questionnaire itself as consisting of certain instructions and certain questions or the exact instructions and stimuli that are presented during an IAT task). The outcome of a measurement procedure has certain functional properties, that is, it functions as an index of an attitude under certain conditions. In the first part of this chapter, I will argue that the distinction between "implicit" and "explicit" measures is situated at this functional level. I will also argue that one needs to specify and examine the functional properties of (implicit) measures in order to (a) reduce conceptual confusion regarding the concept "implicit

measure", (b) arrive at a better understanding of the processes that underlie the measure, and (c) have a better understanding of which measures might predict which types of behaviors.

The second and longest part of this chapter focuses on the measurement procedure rather than the functional properties of the measurement outcome. A first aspect of the procedure that can be used to differentiate measures is related to whether participants are asked to self-assess their attitudes. This aspect determines the distinction between direct and indirect measures. The second procedural aspect involves the structural properties of the measurement procedures. As I argued in an earlier chapter (De Houwer, 2003b), reaction time measures of attitudes can be characterized on the basis of the type of compatibility that is manipulated. In this chapter, I focus on two implications of the structural analysis, namely (a) the issue of whether measures (in particular the IAT) reflect the attitudes toward categories (e.g., FLOWERS) or toward the exemplars of those categories (e.g., TULIP) and (b) the issue of the validity and reliability of measures that are based on different types of compatibility.

The Functional Level

Many of the attitude measures that have been proposed during the past decade have been called "implicit measures". Despite the immense popularity of these measures, it is rarely spelled out what sets these implicit measures apart from more traditional (explicit) measures of attitudes such as questionnaires. When we look at the definitions of implicit measures that can be found in recent psychological literature, researchers often argue that implicit measures provide an index of a certain attitude or cognition even though participants (a) are not aware of the fact that the attitude or cognition is being measured (e.g., Brunel, Tietje, & Greenwald, 2004), (b) do not have conscious access to the attitude or cognition (e.g., Asendorpf, Banse, & Mücke, 2002), or (c) have no control over the measurement outcome (e.g., Fazio & Olson, 2003). What is clear from these definitions is that they do not

refer to objective properties of the measurement procedure itself. A procedure is merely a set of guidelines about what one should do as a researcher (e.g., present certain instructions and stimuli and record certain responses). Rather, the definitions of "implicit measure" that can be found in the literature refer to the conditions under which the outcome of the procedure functions as an index of the to-be-measured attitude or cognition. In line with the available definitions, one can therefore say that the term "implicit measure" refers to certain functional properties of measurement outcomes: The outcome functions as an index of an attitude or cognition under certain conditions, for instance, despite the fact that participants are unaware of the impact of the attitude or cognition on the outcome, are not aware of the attitude or outcome, or have no control over the outcome.

It is not entirely clear which functional properties can be considered as typical for implicit measures. Most often, the term "implicit" is used to refer to properties related to (un)awareness. However, much can be said for using it in the broader sense of "automatic" (see De Houwer, 2006). The concept "automatic" can be linked to a variety of functional properties of which properties related to (un)awareness are only a subset. Each property refers to certain conditions such as the presence or absence of (a) goals (i.e., [un]controlled, [un]intentional, goal-[in]dependent, autonomous, purely stimulus driven), (b) awareness of an input, output, goal, or process, (c) processing resources, or (d) time (i.e., speed). For instance, a process can be said to be automatic in the sense of unintentional if the process operates even when the person does not have the goal to start the process (see Moors & De Houwer, in press, for an in depth discussion of the exact meaning of these properties). In a similar manner, implicit measures can be defined as measurement outcomes that reflect to-be-measured attitude in an automatic manner, that is, even in the absence of certain goals, awareness of certain elements, processing resources, or time. For instance, a measurement

outcome can be described as implicit in the sense of uncontrolled if it reflects the attitude even when participants have the goal to avoid expressing their attitude (see De Houwer & Moors, in press, for a detailed discussion of how properties related to automaticity can be applied to the concept of implicit measures).

One can thus examine the functional properties of a measure by testing whether the measurement outcome provides a valid index of the to-be-measured outcome when certain conditions are or are not met. Many of these conditions refer to mental states of the person (e.g., goals, awareness) but others refer to objective properties of the situation (e.g., the time during which a stimulus is presented) or can be phrased both in terms of mental states (e.g., resources) or objective properties of the situation (e.g., the presence of demanding secondary tasks). The crucial issue is not whether functional properties refer to mental states or objective properties of the situation. I also do not want to make strong claims about which functional properties are central to the concept automatic or implicit. This is in large part a matter of convention. But what is crucial is the idea to define implicit measures as measurement outcomes with certain functional properties, that is, by specifying which conditions do or do not need to be fulfilled in order for the measurement outcome to function as a valid index of the to-be-measured construct.

There is also no one-to-one mapping between the functional properties of attitude measures and the functional properties of attitude activation. Just like one can examine the conditions under which a measurement outcome functions as a valid index of an attitudes, one can also examine the conditions under which the to-be-measured attitude can be activated (e.g., when participants do not have the goal to activate the attitude, are not aware of the attitude, are engaged in other demanding tasks, ...). [1] The conditions under which a measurement outcome reflects the to-be-measured attitude are not necessarily the same as the

conditions under which the to-be-measured attitude is activated. On the one hand, a measurement outcome can function as a valid index of an attitude only when the attitude has been activated. Hence, the functional properties of an attitude measure (i.e., the conditions under which a measure is valid) can provide important information about the functional properties of attitude activation (i.e., conditions under which the underlying attitude can be activated). On the other hand, situations in which a measure is not valid do not necessarily provide information about the conditions under which the underlying attitude is activated. This is because each measure depends not only on the activation of the crucial attitude but also on additional processes by which this attitude is translated into behavior. Hence, conclusions about the functional properties of measurement outcomes cannot be based solely on knowledge about the functional properties of attitude activation.

It is important to realize that it only makes sense to say that a measure is implicit if one is explicit about the functional properties of the measure and if one has evidence to back up that claim. One cannot simply say that a measure "is" implicit, because the different functional properties do not always co-occur (Bargh, 1992; Moors & De Houwer, in press). For instance, existing evidence suggests that participants have little intentional control over the outcome of the IAT (e.g., Steffens, 2004) but are often aware of what a certain IAT is meant to measure (e.g., Monteith, Voils, & Ashburn-Nardo, 2001). Therefore, an IAT effect can called an implicit measure in the sense that the size and direction of the IAT effect is difficult to control, but not in the sense that participants are typically unaware of the fact that the IAT effect measures the target attitude. In order to claim that a measure is implicit, it is thus not only necessary to demonstrate that the measure is valid and reliable (otherwise it is not a measure in the real sense of the word), one also needs to specify its functional properties and collect evidence to support these claims about functional properties (otherwise it cannot

be called implicit).

There are several reasons why it is important to examine the functional properties of (implicit) measures. First, as is clear from the previous paragraph, without empirical evidence regarding the functional properties of a measure, it is not possible to specify the sense in which the measure can be regarded as implicit. In fact, without such evidence, one cannot even claim that the measure is implicit. Hence, research about the functional properties of measures is necessary before one can reach an acceptable level of conceptual clarity.

Second, examining functional properties will also increase our understanding of the processes that underlie the measures. As I pointed out above, a measurement outcome can reflect a construct or entity (such as an attitude) only if it is (partially) produced or determined by the construct or entity. In other words, there are some underlying processes through which the construct or entity is activated and influences the outcome of the measurement procedure. When one says that an (implicit) measure has certain functional properties, that is, functions as an index of an attitude or cognition under certain conditions, it means that the processes that underlie the measure (e.g., the automatic activation of the attitude) operate under those conditions. Hence, research about the functional properties can help us to understand the processes the underlie the measure and provide the much needed measurement theory that is needed in order to consider a measure as valid (see Borsboom, Mellenbergh, & van Heerden, 2004, for an excellent discussion of this issue).

Finally, understanding the functional properties of a measure could also help us get an insight in the conditions under which the measure can be used to predict behavior. Let us consider the MODE model of Fazio (1990) that provides a useful framework for understanding the possible benefits of implicit measures in predicting behavior. As Fazio and Olson (2003, p. 301) pointed out, the MODE model “proposes that attitudes can exert

influence through relatively spontaneous or more deliberative processes. The former involve judgments of, or behavior toward, an object being influenced by one's construal of the object in the immediate situation - perceptions that themselves can be affected by individuals' attitudes having been automatically activated upon encountering the attitude object. In contrast, deliberative processing involves a more effortful, cost-benefit analysis of the utility of a particular behavior". Importantly, deliberative processing will take place only when participants have the opportunity and are motivated to engage in such processing.

Most often, people do not analyze their attitudes towards stimuli in a conscious and deliberate manner. Rather, their behavior is guided by a spontaneous, automatic affective appraisal of the environment (e.g., Zajonc, 1980). Whereas traditional questionnaires typically tap into the consciously constructed and expressed attitudes, implicit measures could index the spontaneous, automatic evaluation of stimuli. Hence, implicit measures could be particularly suited to predict spontaneous, uncontrolled behavior. Fazio and Olson (2003) reviewed evidence which suggests that implicit measures are indeed particularly helpful in predicting behavior that is intrinsically difficult to control or behavior in situations where people are not motivated or do not have the opportunity to control the impact of automatically activated attitudes on behavior.

From the perspective of the MODE model, implicit measures can be regarded as laboratory equivalents of the automatic influence of attitudes on real-life behavior. Hence, implicit measures can provide a unique perspective on real life behavior. This argument is closely related to the idea of transfer-appropriate processing (e.g., Roediger, 1990). That is, the closer the overlap between the processes that determine the measurement outcome and those that determine the actual behavior that one wants to predict, the more that the measurement outcome will be able to predict the behavior (also see Vargas, 2004). In fact,

one could say that both the measurement outcome and the real-life behavior have certain functional properties (i.e., the conditions under which the attitude influences the outcome or behavior). One could thus argue that the predictive value of the measurement outcome depends on the extent to which its functional properties overlap with the functional properties of the real-life behavior that one wants to explain. For instance, real-life attitude-driven behavior that occurs when people do not have the conscious goal to evaluate stimuli in the environment might be related most to measurement outcomes that occur in the absence of a conscious evaluation goal.

Despite the importance of knowing the functional properties of (implicit) measures, relatively little research has been conducted to examine these properties (see De Houwer, 2006, and De Houwer & Moors, in press, for a more detailed overview). There is some, albeit limited evidence that affective priming effects can capture attitudes even when participants do not have the conscious goal to evaluate the stimulus (uncontrolled, unintentional), when they do not have the general conscious goal to evaluate stimuli in the environment (partial goal-independence), when they are unaware of the presented stimuli (awareness), when resources are limited (efficient), and when there is little time to process the stimuli (fast) (see Klauer & Musch, 2003, for a review). Surprisingly, even less research has been conducted to examine the functional properties of IAT effects even though the IAT is currently the most popular implicit measure. There are some studies showing that it is difficult to fake IAT effects (e.g., Steffens, 2004, but see Fiedler & Bluemke, 2005). This suggests that IAT effects are relatively uncontrollable in that they provide a valid measure of attitudes even when participants have the goal to alter the impact of their attitude on behavior. There is one study showing that many participants are aware of what an IAT measures (Monteith et al., 2001), but little is known about whether this kind of awareness influences the validity of the IAT.

Because of the reasons given above, it is imperative that more research is conducted to test the functional properties of the various implicit measures.

The Procedural Level

Direct versus indirect measures

Whereas the terms "implicit measure" and "explicit measure" refer to functional properties of a measurement outcome, the terms "indirect measure" and "direct measure" refer to an objective property of a measurement procedure (see De Houwer, 2006). Direct measures are procedures (tasks) in which the participant is asked to self-assess the to-be-measured construct (e.g., an attitude) by selecting a certain response (e.g., given a rating) and in which the attitude is inferred on the basis of that response. Indirect measures, on the other hand, are procedures in which the participant is not asked to self-assess the to-be-measured construct or, if the participant is asked to self-assess the construct by selecting a certain response, the construct is not inferred on the basis of selected response. Whether a measure is direct or indirect is thus an objective property of the task or procedure. There is no need to do research about whether a measure is direct or indirect. It can be determined simply by looking at the procedure.

Consider the following example (see De Houwer, 2006). In a typical study on the name letter effect (Nuttin, 1985), participants are asked to express their liking of each letter of the alphabet using a Likert-type rating scale. This measurement procedure is a direct measure of attitudes towards letters because participants are asked to self-assess their attitude toward each letter by selecting a certain rating and the attitudes are inferred on the basis of the ratings that the participants select. However, on the basis of the letter ratings, researchers can indirectly infer self-esteem by comparing how much a person likes the letters of his/her name better than other letters. This procedure (i.e., asking participants to rate letters and then

comparing ratings for name letters with ratings for other letters) is an indirect measure of self-esteem because participants are not asked to self-assess their self-esteem. There is indeed evidence that this indirect procedure of assessing self-esteem results in valid estimates of self-esteem (e.g., Koole, Dijksterhuis, & van Knippenberg, 2001). [2] When the participant is asked to self-assess self-esteem, for instance, by giving a rating on a Likert scale but self-esteem is inferred from another behavior such as the activity of certain facial muscles or the degree of eye contact with the experimenter during the rating task, the procedure can still be classified as an indirect measure. This is because the to-be-measured construct is a assessed on the basis of a behavior or index that is different from the one that was said to be relevant for the self-assessment.

It should be clear that indirect measures are not a separate class of measures next to the class of implicit measures. The qualification "direct/indirect" refers to the measurement procedure whereas the qualification "implicit/explicit" refers to the functional properties of the outcome of the measurement procedure. Each direct and indirect measure produces an outcome with certain functional characteristics. Not all indirect measures produce outcomes that have (all the) functional properties typical of implicit measures. For instance, asking person A to self-report how much time (s)he spends with person B can provide an indirect way of assessing how much person A likes person B, but chances are that it does not provide an implicit measure of that attitude (in the sense of, e.g., uncontrollable, unintentional, or unaware). Likewise, the IAT task is an indirect measure because participants are not asked to self-assess the construct of interest. But evidence suggests that IAT effects are implicit only with regard to some functional features. For instance, IAT effects appear to be implicit measures in the sense that they are relatively difficult to control, but not in the sense that participants are unaware of what is being measured. Whether and in what sense indirect

measures are implicit is thus a matter of research. Direct measures also do not by definition provide explicit measures. For instance, one can ask participants to express their liking of a certain attitude object as quickly as possible or while performing a demanding secondary task. In such cases, participants might have little control over the expressed attitude (e.g., Wilson, Lindsey, & Schooler, 2000). But regardless of whether a measure is direct or indirect, one should always verify what the functional properties of the measurement outcome are before claiming that it is an implicit measure.

The fact that context variables can influence the functional features of a measurement outcome while leaving intact the indirect nature of the measurement procedure also shows that the concepts "indirect" and "implicit" do not overlap. Take the example of asking person A to report how much time he or she spend with person B. The extent to which person A is aware of the fact that the answer to this question is meant to reflect his or her attitude toward person B can depend on the context in which the question is given (e.g., when it follows after questions about the positive or negative traits of person B versus questions related to physical attributes of person B). But the context in which the question is given has no impact on the fact that person A is not asked to self-assess the attitude toward person B. Hence, context can change the sense in which a measurement outcome is implicit while leaving intact the indirect nature of the measurement procedure.

Structural properties

A taxonomy. Many of the so-called implicit measures that have been proposed during the past decade are reaction-time based measures. That is, attitudes are inferred on the basis of the speed or accuracy with which participants respond to presentation of certain stimuli. I previously proposed a heuristic framework or taxonomy for describing and comparing reaction time based tasks at the structural level (De Houwer, 2003b). The structure of a task

refers to procedural elements that remain constant in different implementations of the task. Hence, like the distinction between direct and indirect measures, a structural description of reaction time based measures is situated at the procedural level.

The framework that I proposed originated from an analysis of standard stimulus-response (S-R) compatibility tasks. These tasks can be characterized on the basis of which types of compatibility vary over trials: relevant S-R compatibility, irrelevant S-R compatibility, and stimulus-stimulus (S-S) compatibility. In traditional S-R compatibility tasks (e.g., Fitts & Seeger, 1953; Kornblum & Lee, 1995), on some trials, the relevant stimulus feature that determines the response is somehow related or similar to the correct response whereas on other trials, both elements differ or are less related (see Table 1). Consider a task in which participants are required to say "left" or "right" in response to the location of stimuli. In one part of the task, participants are instructed to give the corresponding responses (i.e., say "left" when a stimulus appears on the left side of a screen and say "right" when a stimulus appears on the right side; compatible task) whereas in the other part, they give the opposite responses (i.e., say "left" to right stimuli and "right" to left stimuli; incompatible task). The two parts of the task differ with regard to the compatibility between the relevant stimulus feature (left or right position) and the responses (say "left" or "right"). Hence, in this task, relevant S-R compatibility is manipulated: On some trials (i.e., all trials in the compatible task) the relevant stimulus feature and the to-be-emitted response are compatible (both refer to left or both refer to right) whereas on the other trials (i.e., all trials in the incompatible task) the relevant stimulus feature and the response are incompatible (one refers to left and the other to right). The typical outcome of these studies is that performance is better when the relevant stimulus feature and the response correspond than when they differ (i.e., faster performance in the compatible task than in the incompatible

task).

Irrelevant S-R compatibility and S-S compatibility can be illustrated on the basis of the well know Stroop color-word task. In this task, participants name the inkcolor of colorwords while ignoring their meaning. On congruent trials, the word refers to the inkcolor (e.g., say "BLUE" to the word BLUE in blue ink). On incongruent trials, the word refers to a different color than the inkcolor (e.g., say "GREEN" to the word BLUE in green ink). Unlike to what is the case in traditional S-R compatibility tasks, the compatibility between the relevant feature and the correct response does not vary over trials because response always corresponds to the inkcolor. Hence, relevant S-R compatibility is not manipulated. The compatibility between an irrelevant stimulus feature and the correct response, however, does vary over trials: On congruent trials, the correct response corresponds to the (task-irrelevant) meaning of the word, whereas on incongruent trials, both elements differ. Hence, the Stroop task can be characterized as a task in which irrelevant S-R compatibility is manipulated. In addition, the compatibility between an irrelevant stimulus feature (i.e., the meaning of the word) and the relevant stimulus feature (i.e., the color of the word) also varies over trials. On congruent but not incongruent trials, the irrelevant meaning of the word corresponds to the inkcolor. Hence, in Stroop tasks, S-S compatibility also varies over trials. Moreover, the manipulation of irrelevant S-R compatibility and S-S compatibility is confounded. That is, when the irrelevant meaning of the word corresponds to the correct response, it also corresponds to the inkcolor (congruent trials) and when word meaning differs from the correct response, it also differs from the inkcolor (incongruent trials). The Stroop task can therefore be described as a task in which irrelevant S-R compatibility and S-S correspondence are manipulated in a confounded manner (see Table 1).

Reaction-time based measures of attitudes can also be characterized according to this

taxonomy (see De Houwer, 2001, and 2003b, for a more detailed discussion). The IAT, for example, is a task in which both irrelevant S-R compatibility and relevant S-R compatibility are manipulated, most often in a confounded manner (see De Houwer, 2001, 2003b). Consider the flower-insect IAT as introduced by Greenwald et al. (1998) in their seminal paper. Participants were asked to categorize names of flowers, names of insects, positive words, and negative words by pressing one of two keys. Results showed that performance was better when flowers and positive words were assigned to one key and insects and negative words to the other key (FLOWER+POSITIVE task) than when the first response was assigned to insects and positive words and the second key to flowers and negative words (INSECT+POSITIVE task). Importantly, all exemplars of the category "flowers" had a positive valence (e.g., TULIP) whereas all exemplars of the category "insects" had a negative valence (e.g., COCKROACH).

Because of the task instructions, one response became extrinsically associated with positive valence (i.e., the response assigned to positive words) whereas the other response was extrinsically associated with negative valence (i.e., the response assigned to negative words; see Proctor & Lu, 2002; De Houwer, 2003a, 2004). Therefore, in the FLOWER+POSITIVE task, the valence of the correct response always corresponded to the valence of the relevant category ("flower" or "insect") whereas in the INSECT+POSITIVE task, the valence of the correct response always differed from the valence of the relevant category. Hence, the degree of compatibility between the correct response and the relevant stimulus feature (i.e., the category of the word) varied over trials. Put more formally, relevant S-R compatibility is manipulated. Because all flower exemplars were positive and all insect exemplars were negative, the task-irrelevant valence of the presented flower names and insect names also corresponded to the valence of the correct response in the FLOWER+POSITIVE

task but not in the INSECT+POSITIVE task. Hence, irrelevant S-R compatibility is also manipulated. Because there is a perfect confound between the valence of the category and the valence of the items in each category (i.e., all flower names are positive and all insect names are negative), the flower-insect IAT as implemented by Greenwald et al. (1998) is a task in which relevant S-R compatibility and irrelevant S-R compatibility are manipulated in a confounded manner (see Table 1).

Other reaction-time based implicit measures have a different structure. Consider the affective priming task. In a typical affective priming study, a prime stimulus is presented briefly and followed immediately by a target that needs to be classified on the basis of its valence (see Klauer & Musch, 2003, for a review). In this task, the valence of the prime is a task-irrelevant feature, the valence of the target is the task-relevant feature, and the responses are related to positive or negative valence. On congruent trials, the prime and the target have the same valence whereas on incongruent trials, they have a different valence. Hence, S-S compatibility is manipulated. The valence of the prime is also compatible with the valence of the response on congruent trials but not on incongruent trials. Hence, irrelevant S-R compatibility is also manipulated. Because the valence of the target is always identical to the valence of the response, just as in the Stroop task, S-S compatibility and irrelevant S-R compatibility are manipulated in a confounded manner (see Table 1).

Finally, the affective Simon task has yet another underlying structure. In this task, participants give a valenced response on the basis of a non-affective feature of valenced stimuli (De Houwer & Eelen, 1998). For instance, participants can be asked to say “GOOD” when they see an adjective and “BAD” when they see a noun, irrespective of the valence of the word. The irrelevant valence of the word and the response match on congruent trials (e.g., say “GOOD” to the word HAPPY because it is an adjective) but differ on incongruent trials

(e.g., say "GOOD" to the word SAD because it is adjective). Hence, the (affective) Simon task is a task in which irrelevant S-R compatibility is manipulated. Relevant S-R compatibility is not manipulated because the responses are always unrelated to the relevant feature. S-S compatibility is also kept constant because the irrelevant valence of the words is unrelated to their relevant grammatical category (see Table 1).

Characterizing implicit measures at the structural level offers a way to see important commonalities and differences between different measures and thus has a clear heuristic value. In the remainder of this chapter, I will argue that it can also have implications for understanding how these measures work. First, I will discuss evidence regarding the role of exemplars and categories in the IAT and the affective priming task. Next, I will look at the implications of the structural analysis for the validity and reliability of different measures.

On the role of exemplars and categories in the IAT. The effect in a reaction-time task is always based on a comparison of performance on certain types of trials. The effect (i.e., the difference in performance on certain types of trials) can arise because of the structural differences between the trial types that are compared. Therefore, when there is a confound between the valence of the categories and the valence of the exemplars in the IAT, the IAT effect can arise either because the trials in the two IAT tasks differ with regard to relevant S-R compatibility or because those trials differ with regard to irrelevant S-R compatibility. In the former case, IAT effects would depend on the properties (e.g., valence) of the categories; in the latter case, IAT effects will reflect the properties of the exemplars. [3] Consider the flower-insect IAT as introduced by Greenwald et al. (1998). If the IAT effect is driven by relevant S-R compatibility, than the typical flower-insect IAT effect is due to the fact that participants have a more positive attitude toward to concept "flowers" than toward the concept "insects". If the IAT depends on variations in irrelevant S-R compatibility, the

flower-insect IAT effect reflects more positive attitudes towards flower exemplars such as "tulip" than toward insect exemplars such as "cockroach".

I examined the relative contribution of variations in relevant and irrelevant S-R compatibility by designing an IAT in which the confound between the valence of the categories and the valence of the exemplars was removed (De Houwer, 2001). British participants were asked to classify names of British persons, names of foreign persons, positive words, and negative words by pressing one of two keys on the basis of the nationality of the persons (British or foreign) or the valence of the words (positive or negative). Importantly, half of the British and Foreign persons were liked by the participants (e.g., Princes Diana; Mahatma Gandhi) whereas the other British and foreign persons were disliked (e.g., Harold Shipman, a well known British mass murderer; Adolf Hitler). Results showed that the British participants were faster in the task where British names and positive words were assigned to one key and foreign names and negative words to the other key than when the foreign and positive words were assigned to the first key and British and negative words to second key. This effect was not influenced by the valence of the exemplars. On the basis of these results, I concluded that IAT effects are driven primarily by the properties of the categories (i.e., British participants have a more positive attitude toward the concept "British" than toward the concept "foreign") whereas properties of the exemplars seem to have little or no effect. Similar results were found in subsequent studies (e.g., Mitchell, Nosek, & Banaji, 2003, Experiment 1; Rothermund & Wentura, 2004, Experiment 4)

The conclusion that properties of the exemplars have little influence on IAT effects was, however, based on a null finding. Moreover, in a footnote of the paper, I described an experiment in which the categories had a more or less neutral valence (i.e., "person" and "animal") but the exemplars were positive or negative (e.g., FRIEND, ENEMY, SWAN,

SNAKE). In this experiment (which actually was the first Extrinsic Affective Simon experiment ever conducted, see De Houwer, 2003a), I did observe a significant effect of exemplar valence (see De Houwer, 2001, Footnote 4). The discrepancy between the results of the two studies can be explained as follows: When the categories are clearly positive or negative, category valence might be much more salient than the valence of the exemplars. Hence, exemplar valence might not have much effect on performance. But when the categories are fairly neutral, exemplar valence might be salient and have an effect on performance (see below for a discussion of data showing that salience of stimulus features can matter).

Research suggests that there are also other ways in which exemplar properties might influence IAT performance. For instance, Steffens and Plewe (2001) presented names of women, names of men, positive words, and negative words. When the positive words referred to stereotypic positive attributes of women (e.g., emphatic) and the negative words to stereotypic negative attributes of men (e.g., brutal), the IAT revealed more positive attitudes toward women than toward men in female participants. This IAT effect was, however, significantly reduced when the positive attribute words were associated with men (e.g., independent) and the negative attribute words associated with women (e.g., bitchy). Likewise, Mitchell et al. (2003, Experiment 2; also see Govan & Williams, 2004, Experiment 1b) presented names of well known White persons, well known Black persons, positive words, and negative words. When all White persons were liked and Black persons disliked, the IAT revealed a strong preference for white people. When all White persons were disliked and all Black persons liked, the IAT effect disappeared. Finally, Govan and Williams (2004, Experiment 1a) found a normal flower-insect IAT effect (faster in the FLOWER+POSITIVE task than in the INSECT+POSITIVE task) when all flower exemplars had a positive valence

and all insect exemplars a negative valence (e.g., daffodil, cockroach). But this effect reversed when the flower exemplars were negative and the insect exemplars positive (e.g., poison ivy, butterfly).

It is important to note, however, that the impact of exemplars on IAT performance does not necessarily provide evidence for the hypothesis that IAT effects are partially determined by irrelevant S-R compatibility. An alternative explanation for these effects is that the nature of the exemplars has an impact on how the categories are conceptualized or on the attitude toward the categories. For instance, when participants are repeatedly exposed to (unusual) names of negative flowers (e.g., poison ivy) and positive insects (e.g., butterfly), it is possible that participants recode the categories in terms of "nasty plants" and "nice animals" or that their attitude toward the concept "flower" temporarily becomes more negative and the attitude toward "insects" more positive. If this is true, then the impact of exemplars on IAT performance is mediated by relevant S-R compatibility effects (i.e., effects of manipulations of relevant S-R compatibility). That is, performance on an IAT trial would be determined by the match between the relevant stimulus feature (i.e., the valence of the categories) and the responses on that trial rather than by the match between an irrelevant stimulus feature (i.e., the valence of the presented category exemplar).

Govan and Williams (2004) obtained some evidence to support this alternative explanation. Their participants first completed an IAT in which names of plants, names of animals, positive words, and negative words were presented. In one condition, all animals had a positive valence (e.g., swan) and all plants a negative valence (e.g., poison ivy) whereas the reverse was true in the other condition (e.g., crocodile, daffodil). Afterwards, they completed a second IAT that was identical to the first except that only the words ANIMAL and PLANT

were used as exemplars for the category “animals” and the category “plants” respectively. This category IAT removed any possible impact of exemplars (simply because only the category labels were presented) and could therefore provide an estimate of the attitude towards the concepts “animals” and “plants”. Results showed that performance in both IATs was influenced in the same way by the nature of the exemplars in the first IAT. That is, both IATs revealed a preference for plants over animals when plant exemplars were positive and animal exemplars negative but a preference for animals over plants when plant exemplars were negative and animal exemplars positive. Because the second IAT could be based only on the properties of the categories, these results suggest that the nature of the exemplars in the first IAT changed the way in which the categories were conceptualized or resulted in a change in the attitude toward these categories.

One way to prevent these changes at the category level and thus to test whether irrelevant S-R compatibility can have a direct effect on IAT performance, is to manipulate the valence of the exemplars on a within-subjects basis rather than a between-subjects basis. This is exactly the approach that was followed in the British-foreign experiment described above (De Houwer, 2001). Because half of the British and foreign names were positive and half were negative, it is unlikely that participants recoded the categories or that the attitudes toward the categories would have changed. The fact that no impact of exemplar valence was found therefore raises doubts about whether irrelevant S-R compatibility can have a direct effect. [4]

Regardless of whether irrelevant S-R compatibility does contribute to IAT effects, it is clear that relevant S-R compatibility is an important source of IAT effects. This conclusion has important implications. First, it provides at least a partial explanation for why an IAT measure of attitudes often does not correlate with other measures of the same attitudes. For

instance, Bosson, Swann, and Pennebaker (2000) found little or no correlation between an IAT measure of self-esteem and self-esteem as measured in an affective priming task. This lack of correspondence could be partially due to the fact that IAT effects are predominantly driven by relevant S-R compatibility whereas affective priming effects are due mainly to effects of irrelevant S-R compatibility (see Klauer & Musch, 2003, for a review). This means that the IAT will generally reflect the attitudes toward the categories whereas affective priming will reflect the attitude toward the exemplars that are used to instantiate the categories (De Houwer, 2003b; Fazio & Olson, 2003).

Olson and Fazio (2003) obtained strong support for this hypothesis. Their participants completed an affective priming task and an IAT that were both directed at measuring attitudes toward White and Black people. When participants were asked to merely pay attention to the primes (faces of persons from different racial groups), there was no correspondence between the affective priming measure and the IAT measure. This is probably due to the fact that the IAT measured attitudes towards the concepts “Whites” and “Blacks” whereas the priming measure reflected the attitude toward the individual exemplars (e.g., how attractive each individual face was). When participants were asked to process the race of the primes (in order to keep a mental tally of the number of faces from each group), the IAT and priming measures did converge. By making race salient, it is likely that the priming effect reflected the racial category of the exemplars rather than other properties of the primes such as attractiveness.

Olson and Fazio (2004) pointed to a second implication of the fact that IAT effects are driven by relevant S-R compatibility. In most IATs, the words POSITIVE and NEGATIVE (or related words such as PLEASENT/UNPLEASENT or GOOD/BAD) are used to label the category of positive words and the category of negative words, respectively. Moreover, when

participants categorize a positive or negative word in a manner that is inconsistent with the categorization intended by the experimenter, they receive error feedback. Because of these task characteristics, it is possible that participants conceptualize the categories "positive" and "negative" in the sense of "normatively positive" and "normatively negative". Because of this, the IAT might reflect not the personal attitudes of the participants (e.g., how much someone likes to smoke) but rather knowledge that they have about the normative societal views regarding the attitude object (e.g., the fact that most people in Western societies nowadays strongly disapprove of smoking). Olson and Fazio therefore developed a personalized version of the IAT in which the labels "I like" and "I dislike" are used for the category of positive words and the category of negative words, respectively, and in which error feedback is no longer given. They reported the results of four experiments that support the hypothesis that such a personalized IAT is influenced less by societal views than a standard IAT.

It is interesting to note that both the studies on the effects of exemplars in the IAT (e.g., Govan & Williams, 2004) and the studies of Olson and Fazio (2004) suggest that IAT effects are susceptible to the manner in which participants conceptualize the categories in the IAT. On the one hand, this is a drawback because it introduces a potential source of error variance that the experimenter cannot control completely. But on the other hand, it also offers possibilities. For instance, in some cases it might be difficult to find a label that unequivocally represents the category or concept of interest. In that case, one can select a label that approaches the intended concept as closely as possible and explain to the participant what the exact meaning of the label is. For instance, in a recent study conducted at our lab (Dewitte, De Houwer, & Buysse, 2005), we used an IAT to measure the attachment dimension of anxiety (Brennan, Clark, & Shaver, 1998). Because it is difficult to represent

this complex concept in just one or two words, we used the labels "relationally worthful" and "relationally worthless" and explained to our participants that the labels and items did not refer to general self-esteem but only to feelings of worth in the context of close relationships. Under these conditions, we found that our IAT measure of relational anxiety did correlate with several questionnaire measures of this attachment dimension (but not other attachment dimensions).

Implications for the validity and reliability of implicit measures. It has been reported repeatedly that affective priming and affective Simon measures of interindividual differences have a lower split-half and test-retest reliability than IAT measures of the same constructs (e.g., Bosson et al., 2000; Banse, Seise, & Zerbes, 2001; Teige, Schnabel, Banse, & Asendorpf, 2004). In this section, I will argue that this could be due to the fact that both affective priming and affective Simon measures are based on irrelevant S-R compatibility effects (i.e., effects of the manipulation of irrelevant S-R compatibility; see De Houwer, 2003b; Klauer & Musch, 2003) whereas IAT measures primarily rely on effects of relevant S-R compatibility (see above). When a measure is based on irrelevant S-R compatibility effects, the target concept is implemented at the level of the irrelevant stimulus feature. For instance, in affective priming studies, one can measure the attitude toward the concept "smoking" by presenting the word SMOKING as the task-irrelevant prime stimulus and examining whether this facilitates positive or negative responses. Likewise, an affective Simon measure of attitudes toward smoking could entail that one asks participants to say "GOOD" or "BAD" on the basis of the grammatical category of the words whose task-irrelevant meaning is related to smoking (e.g., CIGARETTE, SMOKING; De Houwer, 2003a; De Houwer & Eelen, 1998). Importantly, because the concept that one wants to measure (i.e., the target concept) is implemented at the level of an irrelevant stimulus or stimulus feature, participants do not

need to process it or take it into account when selecting their response. But the measure can work only if the target concept *is* processed and if its associated properties *have* an impact on the selection of the responses. If these conditions are not met, than logically the target concept cannot have an effect on performance and thus cannot be revealed by the measure. It is likely that a variety of factors will influence whether, when, and to which extent these conditions are fulfilled and thus whether, when, and to which extent the measure is valid and reliable.

First, certain aspects of the procedure can determine the likelihood that the target concepts are processed or have an impact on performance. For instance, Musch and Klauer (2001) demonstrated that affective priming effects are smaller when the prime is consistently presented at a different location than the target. They argued that this effect is due to the fact that automatic processing of valence depends on the allocation of spatial attention. Likewise, De Houwer, Crombez, Baeyens, and Hermans (2001) showed that the magnitude of the affective Simon effect depends on the nature of the relevant feature. This is probably due to the fact that the relevant feature determines the extent to which participants process the irrelevant target feature (and thus the target concepts). Both findings suggest that procedural parameters can influence the likelihood that the target concept is processed. Hence, one should make sure to use a procedure that is known to allow for a sufficient processing of the target concepts. Otherwise, irrelevant S-R tasks such as the affective priming and affective Simon task cannot provide a valid measure of the properties of the target concepts.

Second, whether the target concept is processed and has an impact on response selection will also depend on the salience of the concept. Consider the study of Olson and Fazio (2003) that I also discussed above. They found a higher correspondence between an affective priming measure of racial attitudes and an IAT measure of those same attitudes when participants were asked to determine the race of the prime stimuli in the priming task

than when participants were instructed to merely pay attention to the primes. Olson and Fazio attributed this difference to differences in the salience of the racial features of the primes compared to other features such as attractiveness or gender. Note, however, that until now there are hardly any studies in which the valence of multiple features of the same stimulus have been manipulated independently. It is thus unclear how the valence associated with different features and concepts interacts and how this interaction is influenced by salience or instructions. Nevertheless, the results of Olson and Fazio are at least in line with the possibility that the impact of the target concept on priming effects depends on the salience of that concept.

Olson and Fazio (2003) argued that the degree of salience of the target concept could also influences the reliability of the measure. For instance, in their priming measure of racial attitudes, the measure corresponds to the differences in responses on trials with a black face as prime compared to trials with a white face as prime. But those participants who pay more attention to the physical attractiveness of the faces will add noise to the measure, which will reduce reliability. Noise can also be added by variations in the type of feature that participants pay attention to over the course of the task or during different administrations of the task.

Based on these considerations, it seems important to ensure that the target concept is salient. This can be achieved through instructions (e.g., Olson & Fazio, 2003), but one should also not neglect the selection of the stimuli. For instance, rather then presenting exemplars of the target category (e.g., pictures of different black men or different names typical of black people), it could be better to present the label of the category itself (e.g., the word BLACKS) or at least to present the label as one of the stimuli. Whereas exemplars are rarely exemplars of a single category, labels often do represent a single category or concept. This reduces the probability that participants start paying attention to features other than the target feature (also

see Livingston & Brewer, 2002). Note, however, that repeated presentation of category labels might generate awareness about the purpose of the task.

A third factor that influences the whether the target concept is processed and influences performance concerns interindividual differences in the capacity to ignore irrelevant information. It is likely that participants will often try to ignore the target concept either because it distracts them from their task or because they are explicitly instructed to ignore it. There is clear evidence from Stroop studies that participants who differ in working memory capacity also differ in their ability to ignore irrelevant information (see Long & Prat, 2002). It is therefore likely that these interindividual differences will reduce the validity and reliability of measures that are based on irrelevant S-R compatibility such as the affective priming task and the affective Simon task. But as far as I know, this hypothesis has not yet been tested empirically.

The fourth factor relates to effects of the order in which trials are presented. Studies on the spatial Simon effect have demonstrated strong effects of trial order on the magnitude of irrelevant S-R compatibility effects. For instance, it has now been demonstrated repeatedly that the spatial Simon effect is much stronger after a congruent trial (e.g., press left because of the color of a stimulus on the left) than after an incongruent trial (e.g., press left because of the color of a stimulus on the right; e.g., Hommel, Proctor, & Vu, 2004). In many studies, the Simon effect even disappeared completely after an incongruent trial. Provided that these findings generalize to other tasks in which irrelevant S-R compatibility is manipulated, such order effects could have a profound effect on the reliability of affective priming and affective Simon measures of attitudes and other constructs. One way to reduce the adverse impact of these order effects on reliability is to control rather than randomize the order of the trials, for instance, by keeping the order of the trials fixed. Another option would be to ensure that each

type of trial is presented a large number of times so that the noise due to order effects is leveled out by averaging across trials. But the effects of trial order could interact in complex ways with interindividual differences in the ability to ignore irrelevant information and other factors such as the duration of the test. Therefore, there is no guarantee that such solutions will be successful.

All the factors that I have discussed above jeopardize the validity and reliability of reaction-time based measures such as the affective priming and affective Simon task that are based on a manipulation of irrelevant S-R compatibility. In contrast, the validity and reliability of measures that are based on a manipulation of relevant S-R compatibility such as the IAT (Greenwald et al., 1998; also see the approach-avoid task as implemented by Mogg, Bradley, Field, & De Houwer, 2003, and the Implicit Association Procedure by Schnabel, Banse, & Asendorpf, in press) is not endangered by factors that determine whether the target concept is processed. The reason is simple: By definition, in relevant S-R measures, participants must process the target concept in order to select the correct response. For instance, in a flower-insect IAT, participants are instructed to respond on the basis of whether a presented name refers to a flower or an insect. [5]

The hypothesis that the task-relevance of the target concepts is a crucial determinant of the validity and reliability of implicit measures is also supported by a recent series of studies on the extrinsic affective Simon task (EAST; De Houwer, 2003a) that was conducted at our lab (De Houwer & De Bruycker, 2005a, 2005b). In a typical EAST study, participants see words that are presented in white, blue, or green. They are asked to categorize white words on the basis of the valence and colored words on the basis of color. By assigning one key to positive words and the other key to negative words, the keys become associated with positive and negative valence, respectively. As such, the trials with colored words are

equivalent to affective Simon trials: Participants give valenced responses on the basis of a non-affective stimulus feature (color) while ignoring the valence of the words. Results typically show that participants respond more quickly and accurately when the irrelevant valence of the word corresponds to the (extrinsic) valence of the response than when the stimulus and response have a different valence (e.g., De Houwer, 2003a).

In principle, the EAST can be used to measure individual differences in attitudes. For instance, if a person needs less time to respond to a certain colored stimulus (e.g., the word SMOKING) by pressing the positive key than by pressing the negative key, one can infer that that person has a positive attitude toward the stimulus. Although some studies have found evidence for the validity of the EAST as a measure of individual differences in attitudes (e.g., Huijding & de Jong, in press; Ellwart, Becker, & Rinck, 2005), in a recent series of experiments, we consistently failed to find any evidence for the reliability and validity of the (extrinsic) affective Simon task as a measure of attitudes toward food items, political parties, and homosexuality. These failures were even more striking because in the same studies we did obtain evidence for the validity and reliability of IAT measures of these attitudes (De Houwer & De Bruycker, 2005a).

In light of these disappointing findings and the arguments presented above, we decided to create a variant of the EAST in which participants were forced to identify the target concept before they could select the correct response (De Houwer & De Bruycker, 2005b). For instance, in order to measure the attitude towards the concepts “beer” and “sprouts” (a cabbage-like vegetable), we presented the words BEER and SPROUTS intermixed with positive and negative adjectives. All words were sometimes presented in uppercase letters and sometimes in lowercase letters. Participants were instructed to evaluate all adjectives by pressing one key for positive adjectives and the other key for negative

adjectives, irrespective of the letter case in which the adjective was presented. The function of these trials was to link the responses with positive or negative valence. Participants were also told that there were two special words, namely the word BEER and the word SPROUTS (both nouns) for which the task would be different. When the word BEER or the word SPROUTS was presented, participants were asked to respond not on the basis of valence but on the basis of the letter case in which the word was presented. (e.g., press the positive key when the word BEER is presented in uppercase letters but press the negative key when it is presented in lowercase letters). Hence, in order to decide whether they should respond on the basis of stimulus valence while ignoring letter case (as was the case for all adjectives) or on the basis of letter case while ignoring valence (as was the case for the target words BEER and SPROUTS), participants first needed to identify the word. If the word BEER or SPROUTS was presented, then letter case was relevant. If it was another word, valence was relevant. Whereas the standard EAST failed to provide a reliable (split-half reliability) or valid (correlations with explicit ratings) measure of these attitudes, the EAST that required identification (we therefore call it the identification or ID-EAST) did provide scores that were fairly reliable (split-half correlations of about $r = .55$) and valid (correlations with explicit ratings and expected differences between heavy and light drinkers for “beer” but not “sprouts”). In fact, the ID-EAST performed at a level close to that of the IAT while overcoming some of the limitations of the IAT (e.g., the ID-EAST provides a measure of single attitudes; see De Houwer, 2003a). The fact that making the target concept relevant in an EAST task seems to improve the psychometric qualities of the EAST supports the idea that irrelevant S-R measures (such as the EAST) are often inferior to relevant S-R measures (such as the IAT) because the target concepts are typically not relevant in irrelevant S-R measures but are relevant in relevant S-R measures.

This does not necessarily imply, however, that relevant S-R measures are always better than measures based on irrelevant S-R compatibility effects. Relevant S-R measures have their own limitations and potential weaknesses. For instance, most often, relevant S-R compatibility is varied between tasks. That is, in one task (e.g., the FLOWER+POSITIVE task in a flower-insect IAT), the mapping between the responses and the relevant target feature is compatible whereas in a separate second task (e.g., the FLOWER+NEGATIVE task) the mapping is incompatible. The measure is therefore derived from a comparison of performance in different tasks. This allows for the possibility that performance in the two tasks does not rely on the same processes and that the difference between the tasks therefore does not (only) reflect the construct that one want to measure. For instance, participants might succeed in finding a shortcut to simplify one of the tasks (e.g., by finding a feature such as valence or salience that is common to all stimuli assigned to the same response regardless of category) but could fail to find such a shortcut in the other task (e.g., De Houwer, 2003a; Mierke & Klauer, 2003). Another potential weakness is that the target concepts need to be made explicit in relevant S-R measures. This could increase the probability that participants are aware of what is being assessed (e.g., Monteith, Voils, & Ashburn-Nardo, 2001). Hence, while relevant S-R measures might often be more valid and reliable that irrelevant S-R measures, they might be less implicit in the sense that fewer of the functional characteristics of implicit processes apply to relevant S-R measures than to irrelevant S-R measures. Given current the lack of research on the functional properties of implicit measures, this hypothesis is, however, still speculative.

Summary and Conclusion

In this chapter, I have described two levels at which measures of attitudes can be described and compared. The first level is that of the functional properties of the outcome of a

measurement procedure. These properties relate to the conditions under which the outcome of a measurement procedure provides an index of the to-be-measured construct. So far, little research has been conducted about these functional properties. Nevertheless, it could help (a) reduce confusion about the concept “implicit measures”, (b) clarify the processes underlying the measures, and (c) make clear the conditions under which a measure can be used to predict behavior. The second level is that of the procedure. There are at least two ways in which measurement procedures can be classified. The first is related to whether participants are asked to self-assess the construct that is measured. If the answer to that question is affirmative, the measurement procedure can be labeled as direct. If the answer is negative, the measure is indirect. The second aspect refers to the structural properties of the task. Here I made a distinction between measures that are based on a manipulation of relevant S-R compatibility and measures based on a manipulation of irrelevant S-R compatibility. I then argued that this distinction has important implications with regard to what is measured and with regard to the reliability and validity of the measures.

It should be clear that there are undoubtedly other ways of characterizing and comparing measures. Likewise, our characterization of measures at the functional and procedural level might have implications that were not yet recognized. Nevertheless, I hope that the analysis and discussion presented in this chapter goes at least some way in clarifying the nature of and relation between different measures of attitudes. The analysis also led to the identification of some important gaps in our knowledge about (implicit) measures of attitudes. Hopefully, this chapter will provide an impetus for addressing these unresolved questions.

References

Asendorpf, J. B., Banse, R., Mücke, D. (2002). Double dissociation between implicit and explicit personality self-concept: The case of shy behavior. *Journal of Personality and Social Psychology, 83,* 380–393.

Banse, R., Seise, J., & Zerbes, N. (2001). Implicit attitudes towards homosexuality: Reliability, validity, and controllability of the IAT. *Zeitschrift für Experimentelle Psychologie, 48*, 145-160.

Bargh, J.A. (1992). The ecology of automaticity. Toward establishing the conditions needed to produce automatic processing effects. *American Journal of Psychology, 105*, 181-199.

Bluemke, M., & Friese, M. (in press). Do features of stimuli influence IAT effects. *Journal of Experimental Social Psychology*.

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

Brennan, K. A., Clark, C. L., & Shaver, P. R. (1998). Self-report measurement of adult romantic attachment: An integrative overview. In J. A. Simpson & W.S. Rholes (Eds.), *Attachment theory and close relationships* (pp.46-76). New York: The Guilford Press.

Brunel, F. F., Tietje, B.C., & Greenwald, A.G. (2004). Is the Implicit Association Test a valid and valuable measure of implicit consumer social cognition? *Journal of Consumer Psychology, 14*, 385-404.

De Houwer, J. (2001). A structural and process analysis of the Implicit Association Test. *Journal of Experimental Social Psychology, 37*, 443-451.

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

De Houwer, J. (2003b). A structural analysis of indirect measures of attitudes. In J. Musch & K.C. Klauer (Eds.), *The Psychology of Evaluation: Affective Processes in Cognition and Emotion* (pp. 219-244). Mahwah, NJ: Lawrence Erlbaum.

De Houwer, J. (2004). Spatial Simon effects with non-spatial responses. *Psychonomic Bulletin & Review, 11*, 49-53.

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 Publishers.

De Houwer, J., Crombez, G., Baeyens, F., & Hermans, D. (2001). On the generality of the affective Simon effect. *Cognition and Emotion, 15*, 189-206.

De Houwer, J., & De Bruycker, E. (2005a). The IAT outperforms the EAST as a measure of attitudes. Manuscript in preparation.

De Houwer, J. & De Bruycker, E. (2005b). The ID-EAST offers a valid and reliable measure of attitudes. Manuscript in preparation.

De Houwer, J., & Eelen, P. (1998). An affective variant of the Simon paradigm. *Cognition and Emotion, 12*, 45-61.

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*, 228-239.

De Houwer, J., & Moors, A. (in press). How to define and examine the implicitness of implicit measures. In B. Wittenbrink & N. Schwarz (Eds.). Implicit measures of attitudes: Procedures and controversies. Guilford Press.

Dewitte, M., De Houwer, J., & Buysse, A. (2005). An external validation of the anxiety/model of self attachment dimensions using the Implicit Association Test. Manuscript

submitted for publication.

Ellwart, T., Becker, E. S., & Rinck, M. (2005). Activation and measurement of threat associations in fear of spiders: An application of the Extrinsic Affective Simon Task. *Journal of Behavior Therapy and Experimental Psychiatry, 36*, 281-299.

Fazio, R. H. (1990). Multiple processes by which *attitudes guide behavior: The MODE model as an integrative framework. In M. P. Zanna (Ed.), Advances in experimental social psychology* (Vol. 23, pp. 75-109). San Francisco, CA: Academic Press.

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

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

Fiedler, K., & Bluemke, M. (2005). Faking the IAT: Aided and unaided response control on the Implicit Association Tests. *Basic and Applied Social Psychology, 27*, 307-316.

Fitts, P. M., & Seeger, C. M. (1953). SR compatibility: Spatial characteristics of stimulus and response codes. *Journal of Experimental Psychology, 46*, 199-210.

Govan, C. L., & Williams, K. D. (2004). Changing the affective valence of the stimulus items influences the IAT by re-defining the category labels. *Journal of Experimental Social Psychology, 40*, 357-365.

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

Hommel, B., Proctor, R. W., & Vu, K.-P. (2004). A feature-integration account of sequential effects in the Simon task. *Psychological Research, 68*, 1-17.

Huijding, J., & de Jong, P. J. (in press). Specific predictive power of implicit associations for automatic fear behavior. *Behavior Research and Therapy*.

Klauer, K. C., & Musch, J. (2003). Affective Priming: Findings and Theories. In J. Musch & K. C. Klauer (Eds.) *The Psychology of Evaluation: Affective Processes in Cognition and Emotion.* Mahwah, NJ: Lawrence Erlbaum.

Koole, S.L., Dijksterhuis, A., & van Knippenberg, A. (2001). What's in a name: Implicit self-esteem and the automatic self. *Journal of Personality and Social Psychology, 80,* 669-685.

Kornblum, S. & Lee, J.-W. (1995). Stimulus-Response compatibility with relevant and irrelevant stimulus dimensions that do and do not overlap with the response. *Journal of Experimental Psychology: Human Perception and Performance, 21*, 855-875.

Livingston, R. W., & Brewer, M. B. (2002). What are we really priming? Cue-based versus category-based processing of facial stimuli. *Journal of Personality and Social Psychology, 82*, 5-18.

Long, D. L., & Prat, C. S. (2002). Working memory and Stroop interference: An individual differences investigation. *Memory & Cognition, 30*, 294-301.

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

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

Mogg, K., Bradley, B. P., Field, M., & De Houwer, J. (2003). Eye movements to smoking-related pictures in smokers: Relationship between attentional biases and implicit and explicit measures of stimulus valence. *Addiction, 98*, 825-836.

Monteith, M. J., Voils, C. I., & Ashburn-Nardo, L. (2001). Taking a look

underground: Detecting, interpreting, and reacting to implicit racial bias. *Social Cognition, 19*, 395-417.

Moors, A., & De Houwer, J. (in press). Automaticity: A conceptual and theoretical analysis. *Psychological Bulletin*.

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

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

Nuttin, J.M. (1985). Narcissism beyond Gestalt awareness: The name letter effect. *European Journal of Social Psychology, 15,* 353-361.

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

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

Proctor, R. W., & Vu, K.-P. L. (2002). Eliminating, magnifying, and reversing spatial compatibility effects with mixed location-relevant and irrelevant trials. In W. Prinz & B. Hommel (Eds.), *Common mechanisms in perception and action: Attention and performance XIX* (pp. 443-473). Oxford, UK: Oxford University Press.

Roediger, H. L. (1990). Implicit memory: Retention without remembering. *American Psychologist, 45*, 1043-1056.

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

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

Steffens, M. (2004). Is the Implicit Association Test immune to faking? *Experimental Psychology, 51*, 165-179.

Steffens, M., & Plewe, I. (2001). Items' cross-category associations as a confounding factor in the Implicit Association Test. *Zeitschrift für Experimentelle Psychologie, 48*, 123-134.

Teige, S., Schnabel, K., Banse, R., & Asendorpf, J. B. (2004). Assessment of multiple implicit self-concept dimensions using the Extrinsic Affective Simon Task (EAST). *European Journal of Personality, 18*, 495-520.

Vargas, P. T. (2004). On the relationship between implicit attitudes and behavior: Some lessons from the past, and directions for the future. In G. Haddock & G. R. Maio (Eds.), Contemporary Perspectives on the Psychology of Attitudes. NY: Psychology Press.

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

Zajonc, R. B. (1980). Feeling and thinking. Preferences need no inferences. *American Psychologist, 35*, 151-175.

Footnotes

1. Just like implicit measures can be seen as measures that capture the attitude in an automatic manner, one can define implicit attitudes as attitudes that influence behavior in an automatic manner. One could object that automaticity and awareness are to a certain extent orthogonal. For instance, an attitude can in principle be activated automatically (e.g., in the sense of unintentionally or efficiently) regardless of whether the participant is aware of the attitude. But the same is true for any other functional feature of automaticity. For instance, an attitude might be activated automatically in the sense of unintentionally regardless of whether resources are available. Hence, awareness of the attitude is just one of several functional properties that are related to automaticity. I see no reason to assign a special status to the property "awareness" or to reserve the term "implicit attitudes" for attitudes that are unaware. First, awareness of the attitude is not central in existing definitions of implicit attitudes (e.g., Greenwald & Banaji, 1995; Wilson et al., 2000; also see De Houwer, 2006). Second, there are important problems in assessing whether people are unaware of an attitude. Third, defining implicit attitudes as unaware attitudes implies that these attitudes are somehow fundamentally different from other attitudes, a claim for which there are neither sound arguments nor empirical data. I thus prefer to define "implicit attitude" as "automatically activated attitude". But regardless of this personal preference, it would be good if researchers would clarify their use of the term "implicit measure".

2. Note that the procedure of indirectly measuring self-esteem on the basis of the ratings of letters is not identical to the procedure of directly measuring attitudes towards letters. In the former but not latter case, the procedure entails that one compares the ratings for name letters with rating of letters that are not in the name. This illustrates that also the calculation of the

dependent variable is part of the procedure.

3. It is possible that the IAT (sometimes) does not reflect the valence of the categories or exemplars, but other features such as salience (see Rothermund & Wentura, 2004; De Houwer, Geldof, & De Bruycker, 2005). To simplify the discussion, I will assume that IAT effects do reflect valence.

4. In a recent study, Bluemke and Friese (in press; Experiment 2) did find an effect of the nature of the exemplars when this was manipulated on a within-subjects basis. The presented words related to East-Germany, West-Germany, positive words, and negative words. Unlike to what was the case in the British-foreign experiment, they manipulated not only the valence of the East-German (e.g., Weimar, surveillance) and West-German items (e.g., freedom, greed), but also the extent to which the positive words (e.g., modest, optimistic) and negative words (e.g., unproductive, greedy) were related to the concept "East-German" and "West-German". Note, however, that because of these manipulations, participants probably found it difficult to decide whether an item should be classified according to region (East- or West-German) or valence (positive or negative). This was evidenced by high percentage of errors. It would therefore be interesting to see whether this pattern of results can be replicated with other categories and stimuli (see De Houwer et al., for related evidence).

5. Note that measures based on relevant S-R compatibility could to some extent be influenced by interindividual differences in working memory capacity and by trial order effects. First, although participants need to process the target concept in order to select the correct response (e.g., TULIP is a flower, therefore push left button), they do not need to process the property

of the target concept that one wants to measure (e.g., that the concept “flower” has a positive valence). On incompatible trials, participants probably try to ignore the target property in active manner because it is associated with the incorrect response. For instance, in the FLOWER+NEGATIVE task of a flower-insect IAT, the word TULIP has a positive valence and thus activates the incorrect positive response. Hence, performance on incompatible trials (and thus the measurement outcome) will depend on the ability to ignore the task-irrelevant property of the target concept. This ability is probably determined by working memory capacity. But the role of working memory capacity will most likely be larger in irrelevant S-R measures than in relevant S-R measures because the former depend not only on the capacity to ignore the target property but also on the capacity to ignore the irrelevant feature that contains the target concept. Second, the extent to which the target property can be ignored probably also depends on task order, for instance, whether on the previous trial, the target property was compatible or incompatible with the correct response. Note, however, that in most relevant S-R tasks, all compatible trials are grouped in one block and all incompatible trials in another block whereas in most irrelevant S-R tasks, compatible and incompatible trials are presented in a random order. Hence, the noise introduced by order effects is probably larger in irrelevant S-R tasks than in relevant S-R tasks.

Author Note

Jan De Houwer, Department of Psychology, Ghent University. Preparation of this chapter was supported by Grant G.0356.03 of the Fund for Scientific Research (Flanders, Belgium). Correspondence should be addressed to Jan De Houwer, Department of Psychology, Ghent University, Henri Dunantlaan 2, B-9000 Ghent, Belgium. Electronic mail can be sent to Jan.DeHouwer@UGent.be .

Table 1.

A Taxonomy of the Stimulus Response Compatibility Tasks and Indirect Measures.

	Is there a manipulation of		
Task	S-S Compatibility?	Irrelevant S-R Compatibility?	Relevant S-R Compatibility?
Traditional S-R Compatibility	No	No	Yes
Stroop	Yes	Yes	No
IAT	No	Yes	Yes
Affective Priming	Yes	Yes	No
Affective Simon	No	Yes	No