The Asymmetry of Causal and Diagnostic Inferences –

A Challenge for the Study of Implicit Attitudes

Klaus Fiedler, University of Heidelberg

Sydney-Symposium March 2009

Imagine an intellectual sports fan, who is perfectly convinced to be free of any nationalist motives or attitudes and who has actually developed a super-national identity in his professional life as a scientist. Yet, when watching a soccer match involving his own national team, the same person feels a strong preference for his own national team, which elicits positive emotions as strong as hardly any other object in daily life. In contemporary research on social cognition, one might refer to this phenomenon as an implicit nationalist attitude that is not captured by the person's explicit anti-nationalist attitude. Likewise, a full-hearted vegetarian who never ate fish or meat over ten years discovers herself feeling pity for not being able to enjoy the delicious grilled monk fish that is being served in a tavern near the beach of a Pacific Ocean resort. Again, an implicit pro-fish attitude can be postulated that would not show up as an explicit attitude. Or, two extremely liberal parents, whose major educational goals are liberalism and tolerance toward minorities, such as homosexuals, are nevertheless shocked when learning that their son is developing a homosexual orientation, and they engage in all kinds of influence attempts to "cure" their son's upcoming sexual orientation. Their implicit attitude against homosexuals breeds discriminative behavior, which is not supported by any visible explicit attitude.

Introduction: From Traditional Attitudes to Modern Implicit Attitudes

All three examples are quite representative of the phenomenon that recent research (cf. Nosek, 2007; Wittenbrink & Schwarz, 2007) has termed implicit attitudes. The most prominent types pertain to ethnic and racist targets, elderly and handicapped people, political parties, and more recently also consumer attitudes towards brands and products (for an overview, see Nosek, Greenwald & Banaji, 2006). The notion of an implicit attitude has been adopted so readily and it has become so familiar and self-evident that the construct is widely considered indispensable to deal with the phenomena listed at the outset. Or how else should one account for covert reactions and affective tendencies that are obviously incompatible with the same person's overt questionnaire responses or introspective self reports.

Upon some reflection, indeed, one has to admit that the old attitude research program, prior to the advent of the implicit-attitude construct, already had a number of theoretical degrees of freedom. With regard to the soccer-fan example, one could have argued, with respect to Rosenberg and Hovland's (1966) three-component model, that the cognitive, affective, and conative attitude component may diverge. Somebody may, at the cognitive level, believe that nationality does not matter and try to act accordingly and, yet, the affective component may be biased toward one's own national soccer team, if only because the national ingroup is confounded with familiarity and affective auto-biographical experience. As to the parents of a homosexual son, it has always been recognized that attitudes and manifest behaviors may diverge under certain conditions (Ajzen, 1991; Snyder & Swann, 1976). Moreover, the vegetarian's mental short-term romance with fish would not constitute a problem at all as long as an attitude was conceived, by definition, as a relative stable, latent, long-term disposition that need not be visible in each and every short-term reaction.

So, at a meta-theoretical level, postulating a new explanatory construct, *implicit attitudes*, only makes sense if the old theoretical devices are not sufficient to deal with discrepancies between attitude components, or between attitudes and manifest behaviors. This is indeed hardly justified, because one could easily attribute the soccer fan example to a deviant affective component, or to a confound of a nationalist attitude and a familiarity effect. Or, one could attribute the vegetarian's desire for fish to an attitude-irrelevant short-term reaction, or the parents' problem with a gay son to an attitude-behavior gap.

Changing definition. It is interesting to note that, although this meta-theoretical issue was hardly ever spelled out, the new research program on implicit-attitudes came along with a break in the definition of the attitude concept. Diverging from the old definition of an attitude as an individual's stable behavioral disposition toward an attitude target, there is now a wide consensus among proponents of the new research program to define an attitude as a mere association between an attitude target and an evaluative reaction (Fazio, 2001; Fazio, Chen,

McDonel, & Sherman, 1982; Greenwald, Banaji, Rudman, Farnham, Nosek, & Mellot, 2002). This minimal definition, to be sure, implies an inflation of the attitude domain. If any affective or evaluative association constitutes an attitude, regardless of its temporal stability, person specificity, and its relation to other cognitive or conative components, this amounts to a dramatic increase in the number of attitudes in the social environment. The vegetarian no longer has one generic, pro-vegetarian attitude, but in addition, she now has an anti-vegetarian implicit attitude toward forbidden monk fish at the beach and maybe a number of other divergent implicit attitudes towards all kinds of food objects. Moreover, whenever her affective reactions to food stimuli change, as she participates in an evaluative conditioning experiment with food as CS, or as she imagines food objects associated with attractive persons or feelings in advertising, a new (implicit) attitude is created on the spot.

From this historical sketch of the recent transition of attitude research, and from our meta-theoretical reflection, it should be evident that the meaning of an "attitude" is reduced to a mere stimulus-valence association, and the base rate of occasions on which such minimal attitudes can be diagnosed is inflated. A number of theoretical constraints have been given up: (a) An attitude no longer requires a relatively consistent configuration of three components. (b) It need not represent a temporally stable personality disposition. (c) It is irrelevant whether an attitudinal association is specific to the individual's personality or to the eliciting stimulus, which may elicit similar responses in most other individuals. (d) And it does not matter if the association is genuine or spurious, that is, whether the association of the soccer fan's national team with positive valence is direct or indirectly mediated by other attributes, such as the own national team's familiarity or the normative belief that friends and colleagues also like the national team. An inflation of attitude phenomena is the logical consequence of the lowered threshold for assessing and diagnosing attitudes due to the omission of all these constraints.

Chapter preview. The remainder of this chapter is devoted to a critical discussion of theoretical and empirical consequences this basic inflation. Starting from a conceptual

clarification of the implicit-explicit distinction, and a cursory overview of attitude measurement procedures and their underlying assumptions, in the next section I will first suggest a taxonomy of pertinent methods. The main purpose for presenting this taxonomy is to demonstrate that the current world-wide concentration on a few latency-based measures reflects a highly selective sample from a much greater variety of potentially very useful methods, for which hardly any systematic comparative evidence is available. Because the vast majority of implicit attitude studies refer to the Implicit Association Test (IAT) and the evaluative priming (EP) procedure, I will focus only on these two prototypical measures, disregarding the fact that different measures may not converge (Brauer, Wasel & Niedenthal, 2000) and that the definition of an implicit attitude is far from being precise (cf. DeHouwer, 2006; Fazio, 2003). Thus, in the third major section, I will point out a fundamental logical and methodological problem of all diagnostic measurement, which is accentuated by the inflation of attitudes defined as mere associations and diagnosed through IAT and EP.

Crucial to this fundamental problem is the asymmetry of causal and diagnostic inferences. Even though an existing person disposition may have a regular causal impact on a diagnostic indicator, the reverse, diagnostic inference from a given diagnostic measure to a person disposition may be weak or even negligible. This problem is particularly strong when a weak attitude definition produces an inflation of persons having a significant test score. In the forth section, I will examine some origins of the asymmetry, pointing out that significant IAT and EP scores may reflect various attitude-independent influences. The last section will be devoted to a discussion of what conceptual and empirical improvements are required to overcome the current problems with implicit attitude measurement.

Theoretical Foundations of Attitude Measurement

Scientists are obliged to specify and critically test the rationale underlying their measurement procedures. What rationale justifies the assumption that certain responses capture latent person attributes, such as attitudes? When diagnosing intelligence in an

intelligence test, the rationale is mainly based on the notion of content validity. Intelligence test items are essentially probes of intelligence. Each item can be considered a tiny sample of the very construct to be measured. When error variance is minimized by aggregating over many items, test performance affords a representative sample of intelligence proper.

When testing attitudes in a traditional Likert-scale questionnaire, the situation is similar but slightly different. Each item also refers to an elementary sample of content pertaining to the same attitude, which is assumed to express itself on the test. However, the individual's responses to questionnaire items depend on more than a pure reflection of the attitude itself, plus unsystematic error variance. Whether an individual endorses the statement "I never discriminate against members of foreign ethnic groups" is not fully determined by the attitude proper, but also depends on his or her willingness to reveal that attitude, on social-desirability concerns and self-presentation motives (Snyder & Swann, 1976). Apart from such motivational factors, the test rationale also relies on non-trivial assumptions about the verbal comprehension of test items, intact auto-biographical memory, and introspective capacity.

Rationales for attitude assessment. Let us use the term *auto-expression* to denote content-valid measurement devices that consist of a representative sample of attitude-related elementary behaviors (i.e., items), on which the attitude is assumed to express itself. This rationale applies to countless attitude questionnaires, which are now usually called explicit procedures, but the same category would also include attitude inferences from facial expressions or behavior observations. Common to all these procedures is the assumption that attitude-relevant stimuli (e.g., attitude items; exposure to attitude objects) have the power to elicit diagnostically useful expressive behaviors, the premise being that both verbal and nonverbal expressions tends to be consistent with the latent attitude. This consistency rule is deemed to be strong enough to override other motives and distracters, such as shame, distrust, privacy, or display rules that prevent individuals from revealing their attitudes. Nevertheless, auto-expression is a complex function of the attitude itself and a complex set of motives to

exhibit and admit that attitude, introspective abilities to retrieve information from memory as well as social and verbal intelligence.

A second rationale that has been widely accepted as a basis for the straightforward assessment of attitudes relies on *approach versus avoidance responses* (Brendl, Markman & Messner, 2005). It is commonly presupposed that organisms approach pleasant and avoid unpleasant stimuli and, by reversing this assumption, that positive and negative attitudes can be inferred from approach and avoidance responses, respectively. Note that this principle can be understood as a special case of auto-expression on a relative-distance scale, presupposing an automatic manifestation of attitudes in movements on the distance dimension.

Pertinent measures include behavioral observations of approach and avoidance behavior in real settings (Weaver, 2008), reflexive versus extractive arm movements (Förster & Strack, 1997), perceptual defense against unwanted stimuli (Gackenbach, 1978), immediacy and abstractness of language use (Mehrabian, 1966; Pennebaker, Meehl & Niederhoffer, 2003), eye-tracking assessment of attended stimuli (Balcetis & Dunning, 2006), preferences in dichotic listening (Schotte, McNally & Turner, 1990) or dichoptic viewing (Gumpper, 1972), or analyses of the implications of thematic associations in projective tests (Lilienfeld, Wood & Garp, 2000). Note that the approach-avoidance principle can also assimilate the notion of selective accessibility of attitude-related information as a special case of individuals' tendency to approach wanted and avoid unwanted information.

Unlike the auto-expressive measures, which are commonly classified as explicit, many approach-avoidance measures are considered implicit, although both rationales share the assumption that attitudes tend to express themselves in attitude-consistent behaviors, and both types of measures involve a complex mixture of conscious and unconscious processes, controlled and strategic influences, display rules, and regulatory motives, which might obscure the manifestation of the attitude. Nevertheless, it is commonly taken for granted that

the consistency of attitudes and behaviors is stronger for approach-avoidance movements than for verbal endorsement responses, although systematic evidence for this claim is missing.

A third rationale draws on the diagnostic value of *representations in associative memory*. According to this rationale, which rests on the new definition, measuring an attitude amounts to assessing the strength of the associative bonds that link an attitude object to evaluative responses in memory. Clearly, this rationale underlies the most popular IAT and EP procedures. The inference of an attitude from response latencies in an IAT or EP does not rely on auto-expression or approach-avoidance impulses. Rather, it is based on the interpretation of response speed as a reflection of functional proximity of attitude target to evaluation in associative memory. Both IAT and EP are consensually called implicit, although the IAT instruction rarely conceals the attitude target, and both IAT and EP are susceptible to strategic influences and explicit instructions to exhibit or suppress an attitude (DeHower, 2001; Fiedler, Bluemke & Unkelbach, in press).

Clarifying the explicit-implicit distinction. To a considerable extent, indeed, what is classified as implicit or explicit appears to be a matter of convention. All three classes of procedures – corresponding to the rationales of auto-expression, approach-avoidance and associative distance – involve a complex mix of conscious and unconscious mental processes, controlled and automatic responses, revealed and concealed attitude targets. Although these mixtures may vary in degree, no task or test can be pretended to represent a pure measure of an attitude, dissociated from strategic, motivational, and conscious volitional influences.

However, is it fair and justified to conclude that the labels “explicit” versus “implicit” reflect labelling conventions rather than a scientifically well-defined distinction? How do proponents of the distinction themselves define implicit versus explicit measures? – The pertinent literature reveals different attempts to define implicit and explicit attitudes, which lead however to divergent results. Analogous to the fundamental difference between implicit learning and implicit memory in cognitive psychology (Schacter, 1992), either the genesis or

the assessment procedure may be crucial for the definition. On the one hand, researchers have distinguished between affective attitudes learned through basic associative processes, such as conditioning or mere exposure, which may be dissociated from cognitive attitudes responsive to arguments, rules, norms, and logical thinking (Wilson, Lindsey & Schooler, 2000).

On the other hand, however, the vast majority of pertinent studies do not care about the original process of attitude acquisition, which is neither controlled nor even assessed retrospectively. Neither IAT experiments nor EP experiments are confined to attitudes acquired through conditioning and associative learning, and several studies from both paradigms have shown that IAT and EP can be influenced through arguments, instructions, imagination, and propositional information (Blair, 2002; Blair, Ma, & Lenton, 2001; Govan & Williams, 2004; Mitchell, 2004). Instead, regardless of the attitude origin, whether it is explicit or implicit is derived from the measurement operation. An attitude measure is called explicit if it is reasonable to assume that the respondent is consciously aware of the measurement purpose and the attitude target, that he or she can volitionally influence the measurement outcome, and that the task allows for strategic control. A measure is called implicit if the target and purpose of measurement unconscious or subtle, and if the outcome is automatic and not amenable to volitional or strategic control. To quote from Nosek's (2007, p.65), "variation in controllability, intentionality, awareness, or efficiency is thought to differentiate implicit and explicit attitudes". These are exactly the "four horsemen" of automaticity (Bargh, 1994). It is thus assumed that IAT or EP measures capture attitudes automatically, and whatever these measures capture is then called an implicit attitude.

The automaticity criterion is indeed very strong and hard to justify. After all, performance on a typical IAT or EP task normally does not conceal the attitude target, nor is it protected from volition, faking, or self-instruction, nor is it purely automatic and not amenable to strategic control, as we shall see below. Conversely, "explicit" questionnaire

responses are not independent of automatic associations and retrieval processes, repression, and spontaneous approach-avoidance tendencies.

Given the circularity of implicit attitudes defined by procedures which are hardly defined by anything else but the goal to measure implicit attitudes, one may look out for other defining features that may help us to understand the potential assets of implicit attitude measurement. One possibility is to interpret the implicit-explicit distinction analogous to direct and indirect scaling procedures, as illustrated by the old comparison of Likert and Thurston scales. In Likert-scaled attitude questionnaires, the raw data (i.e., responses to pre-tested attitude statements) are directly obtained on the same dimension as the attitude itself. In contrast, indirect Thurstone scaling is derived from the statistical distribution of comparative judgments that have to be translated into attitude scale values on the basis of a testable psychometric model. This model does not always guarantee a solution; it may not be applicable when the premises are falsified. From a scientific point of view, transparency and falsifiability are obvious advantages over the direct scaling methods.[1]

Given that the raw data obtained in IAT and EP are latencies rather than responses on the attitude dimension proper, one might interpret them as indirect scaling methods, which might be superior to the uncritical methodology guiding direct questionnaire response. Unfortunately, this methodological asset does not apply to IAT and EP, because no testable and falsifiable psychometric model is offered to translate latency data into attitudes. Rather, an IAT or EP score is always assumed to reflect an attitude. No limiting conditions are identified under which the psychometric model fails.

Taxonomy of attitude measurement procedures. Table 1 presents a taxonomy of attitude measurement procedures, organized by the combination of the two types of underlying assumptions, the three rationales introduced before and the direct versus indirect distinction. This framework reveals that half a century of attitude research has generated a multitude of

[1] It should be noted, though, that Likert scaling also involves a test of one testable premise, internal consistency.

measures, distributed across all six cells. In general, indirect procedures are less prominent than direct scaling methods, which simply equate raw responses with attitudes. Note that by this criterion, IAT and EP are also classified as direct measures, because they assume that the latency differences they assess represent attitudes (defined as target-valence association), thereby circumventing the need to test the latency-attitude relation. Note also that many measurement procedures that might be called "implicit" have been met with little research interest. This is evident from the number of literature references encountered in the PsychInfo data base, using the italicized phrases in Table 1 search prompts.

What feature of the prominent IAT and EP procedures could explain this highly selective research focus? Why is it the case that other fascinating methods, like eye-tracking, linguistic content analysis, or binocular rivalry, received so much less attention than IAT and EP? – Although historical interpretations should be always considered with caution, I dare to provide three good reasons for the concentration on two paradigms. First, in accordance with the often-stated role of research instruments in the evolution of science (Gigerenzer, 1991), the availability and sharability of inexpensive hardware and software tools has rendered chronometric measurement extremely easy and convenient. Many other methods (e.g., eye-tracking) are more expensive, laborious, and difficult to handle. Second, the low threshold imposed on the identification of attitudes defined as mere associations makes IAT and EP experiments very likely to yield many new attitude findings. Moreover, interpreting this inflated reference set of IAT and EP findings as unconscious, hidden, repressed, or as implicit secrets, greatly enhances their surplus meaning. And third, the multi-trial structure of both speeded-classification tasks warrants highly reliable and significant results in almost every study, thus increasing the chances for publication and providing a strong incentive for young researchers whose existential task is to increase their publication record.

Asymmetry of Causal and Diagnostic Inferences: A Major Problem For Implicit Measures

To repeat, this is but a subjective interpretation of the selective interest in some methods, and the almost total neglect of other, equally promising methods, and it is not meant to be polemic at all. But whatever historical interpretation is correct, it is an empirical fact that IAT and EP research has greatly increased the baserate of responses that qualify as an attitude. Returning to the basic asymmetry noted at the outset, one primary function of implicit attitude research was to create and enhance the baserate asymmetry between p(D) and p(A) depicted in Figure 1. The inflation of significant attitude scores, reflecting the low threshold, is vividly evident in findings of 90% White Americans having a significant score in a race IAT, and a similarly inflated proportion of Germans turn out to exhibit prejudice against Turks according to German-Turk IAT results.

One could of course reify these test outcomes and conclude that if IAT latencies are biased that way, then we have to accept the truth and admit that the prevalence of implicit attitudes is actually that high. However, scientific standards prohibit such an uncritical equation of test results with latent attitudes. After all, it might be demonstrated that elevated test scores can be due to other causes but attitudes. Such extraneous influences, or diagnostic artefacts, might explain why the set of diagnosed attitudes is much larger than the set of actually existing attitudes. Accordingly, in the next section, we review evidence for false alarms in IAT and EP effects, which may account for the inflation of p(D) in Figure 1.

One might object that calling p(D) inflated is only justified when the true-attitude base rate, p(A), is known objectively. In fact, however, whether the true base rate of Americans hating holding attitudes against Blacks or Germans prejudiced against Turks is 5%, 10%, 20%, 30% or actually 80% depends on how an attitude is defined arbitrarily. To overcome this problem, though, let us take a pragmatic stands and operationalize p(A) in terms of the behavioral criteria used in validity studies. What is the prevalence of people conversing or interacting with Blacks, voting for Barak Abama, cheering and clapping for Black athletes, or engaging in close relationships with Blacks? The asymmetry between p(D) and p(A) in Figure

1 would then compare diagnosed attitudes no longer to true attitudes but to behavioral criteria. In any case, the asymmetry of causal and diagnostic inferences constitutes a major source of confusion and misunderstanding.

This problem has been recognized in various areas of medical diagnosis (Gigerenzer & Hoffrage, 1995), legal decisions (Wells & Olson, 2003), and risk assessment (Swets, Dawes & Monahan, 2000). Because costs and liability problems associated with false negatives (i.e., unidentified disease, guilt, or danger) are typically higher than the costs associated with false positives (i.e., misclassifications of healthy, innocent or harmless cases as dangerous), diagnostic assessment in all these domains is characterized by overly low thresholds, or liberal decision criteria. As the rate of p(positive HIV test) is roughly seven times higher than the actual p(HIV) rate, HIV diagnostic is very false-alarm prone. Likewise, the pragmatic context of eye-witnesses identifications produces an inflated number of false identifications, reflecting too liberal a threshold for eye-witnesses identification (Fiedler, Kaczor, Haarmann, Stegmüller & Maloney, in press; Wells & Olson, 2003). The same holds for high-tech safety systems such as airplane cockpit alarm systems, which generate a huge number of false alarms (cf. Swets et al., 2000).

Impressive research in all these domains has pointed out severe misinterpretations of validity statistics, due to the asymmetry of p(test result | cause) and p(cause | test result). Diagnostic inferences are inflated when the former probability is higher than the latter. Bayes theorem tells us that the probability p(cause | test result) of a correct diagnostic inferences is equal to the reverse, causal probability p(test result | cause) multiplied by the ratio of the two base rates, p(cause) / p(test result). If this ratio it 1/7 as in HIV testing, this means that although the rate of positive test results given HIV is close to 100%, the diagnostic inference from a positive test result to an existing HIV virus is only 15% (Swets et al., 2000).

There is no reason to hope that attitude assessment can circumvent this fundamental problem. Let us assume that the actual rate of racist attitudes (according to some validity

criterion) is 30% whereas the prevalence of IAT scores indicating racist attitudes is 90%, giving a base rate ratio of 1/3. Now imagine that a published IAT study – using a selected sample of racists, or an experimental manipulation of racist attitudes through conditioning – yields a causal probability of p(IAT | A) = 100%, suggesting perfect accuracy. The reverse, diagnostic inference that an individual with an equally strong IAT score is a racist would still be correct in only 33% of the cases. If the base rate ratio is more extreme than 1/3, the diagnostic accuracy rate may further shrink to 20%, 10% or below.

Thus, even when the causal hit rate p(IAT | A) of an IAT used as a research instrument in an experiment using selected or manipulated attitudes were perfect, the diagnostic accuracy p(A | IAT) of attitude inferences from IAT scores observed in an unselected sample may be very modest. Given typical baserates of 90% people with racist IAT scores but, say, only 9% real racists according to some behavioral criterion, the asymmetry may easily amount to 1/10.

It should be recognized that this scenario is not derived from a contestable theory. It is an analytical truth derived from Bayesian calculus. If something as mundane as an association is sufficient proof for an attitude, then it is very likely that p(D) will be inflated relative to p(A). This reflects a universal problem of all diagnostic inference; it is not peculiar to any inherent weakness of IAT or EP. What undermines diagnostic accuracy is neither Bayesian calculus nor the mode of measurement. The problem, rather, arises from asymmetric base rates, when an inflated p(D) exceeds p(A), due to a weak diagnostic threshold. Thus, quite independent of the fuzzy debate about implicit or explicit measurement, a major weakness of IAT and EP arises from the inflation of diagnosed attitudes and stereotypes.

A somewhat unusual, non-mainstream implication of the asymmetry problem is that correlation coefficients should not be used to quantify criterion validity. This conclusion cannot be denied with the conformist argument that correlation coefficients have always been used as a standard means of expressing validities. If the purpose of validation is to predict a criterion from a diagnostic symptom, then the asymmetry of causal and diagnostic (or

prognostic) inferences must be taken into account, not only in IAT or EP research but in general. Diagnostic inference in classical domains, such as intelligence testing, is not immune to asymmetry but actively avoids it. Intelligence scores, for instance, are calibrated such that roughly one half of the population receives a score above and below the average, respectively. Thus, the ratio of p(intelligent) and p(diagnosed to be intelligent) is balanced. If a new test of (hidden aspects of) intelligence were proposed, showing that in actuality 90% or so of the population are (implicitly) smart or stupid, a similar asymmetry problem would arise for intelligence assessment. In this regard, low diagnostic accuracy is a price to be paid for a low diagnostic threshold, or a weak definition of the phenomenon to be diagnosed.

Empirical Evidence For the Inflation of Implicit Attitudes

Thus far, we were only concerned with theoretical and methodological issues. These issues have been of little interest to implicit-attitude researchers, who mainly emphasize empirical findings. A common argument is that psychometric models are of little interest if only implicit measures do predict manifest behavior. Let us therefore look out for concrete evidence for empirical manifestations of the asymmetry problem. This evidence refers to spurious IAT and EP effects, or false alarms, reflecting extraneous causes different from attitudes, thus providing empirical evidence for why p(D) exceeds p(A).

Deliberate strategies mimicking implicit attitudes. Although independence of voluntary and strategic control is commonly praised as a major advantage of implicit measurement, there is compelling evidence that IAT and EP effects can be reduced, eliminated, reversed, simulated, or controlled strategically. In social cognition, the term “strategic” is often used in an unduly restrictive fashion, confined to conscious, deliberate self-control. However, in cognitive psychology, there is ample evidence for strategic effects both above or below the threshold of conscious awareness. Let us first examine the impact of overt intentions to produce certain IAT and EP outcomes before we turn to evidence on unconscious strategies.

A growing number of findings show that self instruction or mental imagination can undo or radically change the IAT or EP performance. Blair's (2002) insightful review describes several studies in which participants were asked to mentally generate of imagine positive experiences with members of ethnic groups (e.g., Blacks) before they engaged in an IAT supposed to measure prejudice against that group. Convergent results demonstrated that people could change their IAT-diagnosed attitudes on the spot. This evidence strongly suggests that there is considerable latitude for control and self-presentation on a test deemed to reveal attitudes automatically. Another way to interpret these findings is to conclude that self-induced short-term mental states can change or override the enduring trait-like person disposition we usually have in mind when we refer to attitudes. In a similar vein, it has been shown that both IAT-experienced and inexperienced respondents can fake their IAT scores. Successful faking is relatively independent of training and assisting instructions (Fiedler & Bluemke, 2005), the only condition being at least one prior exposure to an IAT task. Recent research by Teige-Mocigemba and Klauer (2008) shows that EP performance is subject to similar intentional or volitional influences.

Thus, an unknown proportion of the inflated set of implicitly diagnosed attitudes may reflect momentary short-term states rather than enduring attitudes. These local states must of course constitute a more inclusive set than more serious attitudes that generalize over time and situations. This leads us to the question of what it is that IAT or EP procedures measure. Do they measure a genuine aspect of the respondent's personality, or of the task situation, or do they even measure aspects of the particular stimuli used for the IAT or EP task?

With regard to the latter possibility, Bluemke and Friese (2006) have provided impressive evidence that subtle changes in the stimuli used for an IAT can induce dramatic changes. In a German-Turk IAT, they were concerned with cross-category associations between target labels and valence stimuli, varying the extent to which the words representing Germans and Turks carried slightly different valence, and to what extent the positive and

negative valence terms had slightly different affinity to Germans or Turks. By varying this semantic aspect of the stimulus materials – which is hardly ever controlled in IAT research (Fiedler, Messner & Bluemke, 2006) – over seven graded levels, they were able to systematically turn a "normal" IAT effect (i.e., roughly 80% Germans seemingly prejudiced against Turks) into a reverse effect (i.e., most Germans favoring Turks). A similar problem with EP as a measure of prejudice against Blacks was already noted by Lepore and Brown (1997), pointing out that the stimuli used to prime the concept of Black people in Devine's (1989) seminal priming study had been biased toward negative valence.

Given such a strong, consensual influence of particular stimuli, which generalizes across respondents rather than measuring individual differences, the question is indeed to what extent IAT outcomes should be attributed to attributes of the person or of the stimulus. To the extent that semantic or pragmatic stimulus meanings suffice to induce a generalized effect, this might explain the inflated p(D) obtained in so many studies.

A third possibility, besides person and stimulus attribution, is to attribute IAT and EP findings to the test situation. Thus, even when the stimuli used to represent Germans and Turks, or positive and negative valence, are unbiased and not confounded, they may take on different meaning as soon as they are juxtaposed in a speeded discrimination task that apparently involves pitting Germans against Turks. In such a "minimal-group" setting (Tajfel, Flament, Billig, & Bundy, 1971), the meaning of neutral Turkish concepts may move in the direction of an outgroup, whereas the meaning of originally neutral German concepts may become valenced ingroup labels. Again, such an emergent meaning shift – which has been hardly ever tested or controlled – may account for an inflated p(D), reflecting situation-specific states rather than enduring person attributes. The same individuals who can give an emergent negative meaning to Turkish labels in this minimal intergroup context, may be able to generate a positive emergent meaning for Turks in another comparative context. Again, the ability to turn Turkish stimuli into negative meaning must not be interpreted as an internal

person attribute. It could as well reflect an aspect of the task situation that can influence the majority of all participants in the same fashion (cf. Bluemke & Friese, 2006).

An intriguing question in this regard is whether IAT effects may be prone to stereotype threat (Steele, 1997). Just as a Black respondent's performance on an intelligence test may deteriorate when he or she is reminded of the existence of the stereotype Black is associated with low intelligence, reminding an IAT participant of the fact that his or her prejudice is being tested can worsen the IAT score (Frantz, Cuddy, Burnett, Ray, & Hart, 2004). Any compulsive attempt to speed up on incompatible trials (i.e., when Black and positive stimuli have to be mapped onto the same response keys) may only amplify the impairment on such trials, reflecting an unsuccessful attempt to suppress an apparent prejudice.

Han, Olson, and Fazio (2006) have pointed out that IAT effects may be due to "extra-personal" associations, as distinguished from genuine attitudes. This somewhat misleading term refers to consensual knowledge that most people, or the society as a whole, associate an attitude target (e.g., Blacks) with negative evaluation (e.g., low intelligence), as distinguished from the respondent's own evaluation. If the valence-target association reflects only vicarious rather than self-referent knowledge, it can hardly be an attitude. Otherwise, a lonely lover and admirer of Blacks, who lives among a majority of racists, could be called a racist.

In summary, there are various ways in which deliberate goals, self-instructions, and malleable stimulus meanings can induce and moderate IAT and EP effects, in the absence of a genuine attitude. However, the full spectrum of false alarms or attitude-independent IAT and EP effects becomes only visible when more primitive, unconscious response strategies are taken into account. Elaborating on this idea will disclose an extended class of strategic influences on response latencies, which must not be mistaken for attitudes proper.

Simplifying response strategies. Many primitive response strategies need not be consciously chosen or planned deliberately; they may reflect the brain's built-in capacity to exploit redundancy in the stimulus world and to simplify the task at hand. For an illustration,

think of the notions of maximizing and probability matching (cf. Shanks, Tunney & McCarthy, 2002). When an organism (e.g., a pigeon motivated to find food) is exposed to a discrimination learning task offering a choice between two paths, A and B, leading to food with reinforcement rates of 70% and 50%, respectively, the organisms may always choose the better alternative A (maximizing), or choose A with a probability that matches the reinforcement rate (i.e., 70%). A primitive organism need not "know" its strategy; it may just follow a very simple rule (i.e., continue doing what led to success). Countless experiments demonstrate that such simple response strategies can be utilized without awareness.

It is even more likely that intelligent human organisms follow similar strategies in speeded discrimination tasks, developing a bias toward more likely stimuli. Imagine an IAT participant who is to classify, under speed instructions, an extended series of stimuli as either White or Black. If the prevalence of White stimuli is markedly higher, a simple but useful strategy would be to expect, by default, a White stimulus on every trial (or on most trials). The resulting bias to respond "White" allows for quick and mostly correct guessing on most trials when full identification is hindered or retarded (e.g., due to fluctuation of attention). Now imagine that positive stimuli are also more frequent than negative stimuli. This will induce a similar bias to respond "Positive".

Given such a double-response bias, toward the more prevalent target and valence labels, what happens when the respondent is then exposed to a double classification task, as in an IAT, alternating between classifying target and valence stimuli? On a congruent trial block, when the two frequent categories, White and positive, have to be mapped onto the same response key, the co-existence of two biases toward the same motor response, must be helpful. The alignment of two simultaneous guessing strategies onto the same motor responses will facilitate the IAT performance on congruent trials. In contrast, on incongruent trials, when White and negative stimuli have to be mapped onto one response key while Black and positive stimuli call for the other key, the two response tendencies will be in conflict and

latencies will be retarded. Thus, the mere coexistence of two primitive response biases, which help to render speeded classification performance efficient, can produce an IAT effect. No real attitude at all is involved in this scenario. It is sufficient to assume that organisms find ways of making a straining task easier and more efficient.

To demonstrate empirically that simple response biases, induced by the unequal stimulus frequencies, can produce false IAT effects, we (Bluemke & Fiedler, 2009) have recently manipulated the base rates of IAT stimuli systematically. When the test included 75% West German (along with 25% East German) stimuli along with 75% positive (and 25% negative) stimuli, a normal IAT effect was obtained, suggesting a more positive attitude of West German participants toward the West German majority than toward the East German minority. The same effect was obtained when most targets were East Germans and most valence stimuli were negative, for the two response biases resulting from this condition should also support the joint mapping of East Germans and negative stimuli (and by complement, of West Germans and positive stimuli) onto the same response keys. However, the other two base rate conditions, involving 75% White and 75% negative, or 75% Black and 75% positive, resulted in a significant reduction of the apparently prejudiced attitude of West-German participants, drawn from the same population as those who produced strong IAT effects with different stimulus base rates.

That the base rate manipulation only reduced but did not reverse the “normal” IAT effect might suggest that frequency-driven response biases do not seem to account for the entire IAT effect. One might argue that even when the prevalent categories driving the response strategies were West German and negative, or East German and positive, there was still a trend to respond faster on congruent than on incongruent trials. However, while it was not intended here to pretend that response biases account for all IAT variance, the strategic account can be easily extended to assimilate this residual bias. After all, the virtual stimulus base rates, which drive strategic responding, depend not merely on the frequency of stimuli in

the experiment but also of stimuli in the participants' social environment. As there are many more West Germans than East Germans in their environment, and positive stimuli can be expected to be more frequent than negative ones (Parducci, 1968; Unkelbach, Fiedler, Bayer, Stegmüller & Danner, 2007), it is not surprising to find a basic West German advantage, due to the co-existence of two environmental or cultural biases, towards West Germans and toward positive stimuli.

Extending the response-bias notion, we can now easily understand that biases may reflect many other factors, besides intra-experimental and extra-experimental stimulus frequencies. Rothermund and Wentura's (2004) figure-ground approach assumes that the two poles of dichotomous stimulus dimensions are often asymmetric, bearing a figure-ground relation. From a West German's point of view, an East German is a figure against the ground of a West German environment. On the valence dimension, negative stimuli are the figure while positive stimuli are the ground. This structural aspect can be used to simplify speeded-classification performance. Rather than switching the dimension (targets or valence) on every trial, participants can sort all stimuli on the same dimension, as either "figure" or "ground". Given such a simplifying strategy, it is obvious that congruent IAT trials (e.g., mapping West Germans and positive words) are easier than incongruent trials (e.g., East Germans and negative words), because the response keys in the former are aligned with the figure-ground strategy. Across eight experiments, Rothermund and Wentura (2004) demonstrated that this strategic principle can account for a plethora of different IAT effects.

Proctor and Cho (2006) have shown the figure-ground principle to be but a special case of a broader class of response strategies. They use the term polarity correspondence to denote a variety of compatibility relations between dichotomous distinctions. For instance, it is easier to pair "Yes" responses with unmarked variable poles and "No " responses with marked poles (e.g., using the suffix "un"; Clark, 1969) than vice versa, or it is easier to align "Go" than "No go" responses with marked variables, negative valence, or figure levels on the figure ground

dimension. All these structural compatibility effects afford candidates for non-attitudinal influences on speeded-classification tasks.

Suffice it to mention briefly that very similar response biases can mimic false attitudes on EP tasks as well. In a recent experiment, we (Fiedler, Freytag & Bluemke, 2009) orthogonally manipulated the base rates of positive and negative primes and of positive and negative targets in an EP experiment. As in the IAT experiments outlined above, it was expected that a double bias toward the more frequent prime valence and toward the more frequent target valence can facilitate the speeded-classification task. Guessing the more prevalent target valence will produce many quick and correct responses on those uncertain trials when the correct response does not come to mind fast enough. This adaptive bias should be further enhanced when participants are encouraged to respond to the prime stimuli too, and when the more prevalent prime valence is the same as the more prevalent target valence. In this condition (i.e., when there are 75% positive primes and 75% positive targets), the double response bias should support the usual congruity effect obtained in EP experiments, as the responses to primes and targets should be biased toward the same response keys. In contrast, when there are 75% negative primes but 75% positive targets, the double response bias should support incongruent trials (i.e., positive targets after negative primes). As expected on theoretical ground, the seemingly normal congruity effect turned into a reverse incongruity effect (i.e., faster responding to targets following primes of the opposite valence) when the prime and target base rates supported opposite response strategies. This pattern was particularly pronounced when participants were explicitly instructed to evaluate not only target stimuli (as usual in the EP paradigm) but also the primes. The sensitivity of EP to stimulus frequencies was also demonstrated by Chan, Ybarra, and Schwarz (2006).

In other EP experiments (Fiedler, Bluemke & Unkelbach, 2009), we manipulated the correlation of prime valence and target valence across all trials, as distinguished from the base rates. In different conditions, the likelihood of positive targets was higher either when the

preceding prime was positive or when the prime was negative, the correlation varying across five levels from r = –.50 to r = +.50. A normal congruity effect was obtained when the prime-target correlation was positive, but a reversal was obtained when the correlation was negative, especially when participants were instructed to respond to both primes and targets. Analogous findings were reported by Spruyt, Hermans, De Houwer, Vandromme, & Eelen (2007) with pictorial rather than verbal primes, and by Klauer, Roßnagel, and Musch (1997) for very short prime-target intervals.

Altogether, these findings highlight the fact that strategic adaptation to the stimulus context, and simplifying response strategies, can moderate, eliminate and even reverse the allegedly automatic functions used in IAT and EP tasks to measure attitudes. Measuring attitudes through EP procedure presupposes a robust congruity effect (Fazio, 2001); when responses to negative positive targets are slower after priming a Black than a White face, the inference is that Black must be associated with a less positive attitude. If, however, the congruity rule can be undone or reversed strategically, such an attitudinal inference may not be valid. The EP effect may reflect a sort-lived state, or the impact of a strategy, rather than an attitude-like person attribute, thus inflating the base rate p(D).

In summary, there are many reasons why IAT and EP tasks can produce significant test scores, which mimic an attitude, although the causal origin may not be an attitude. There are multiple ways in which response strategies, deliberate attempts to fake or modify one's test result, self-induced states, and attempts to simplify the speeded-classification task can produce false attitude scores.

How to Overcome the Shortcoming of Implicit Attitude Research

It must be acknowledged that no singular cause of non-attitudinal IAT and EP effects that we have considered in the preceding section must explain all available evidence. The crucial point of interested is not which specific type of false alarm is responsible for the inflation of implicit attitudes, or what proportion of variance can be explained by each

particular cause. The purpose of the preceding section was only to point out that empirical research has discovered diverse origins of attitude-independent IAT and EP effects, which together account for considerable inflation of p(D).

By emphasizing the attitude inflation problem and the resulting asymmetry of causal and diagnostic inferences, we can also predict and explain under what conditions IAT or EP scores should bear strong correlations with criterion behaviors. This should be the case when the base rate of a criterion behavior is not too low. Thus, when a study sample includes a reasonable proportion of racists, or psychopaths, or gender-stereotyped people, because half of the participants have been selected accordingly, the base rate p(A) of the "hot" attitude should be high. As a consequence, the ratio p(A) / p(D) should not be too low, and diagnostic accuracy when inferring attitude from conspicuous test scores should be rather high. Consistent with this argument, a relatively high validity has been found for IAT and EP in study designs that warrant a reasonably high p(A) rate, by selecting sufficient attitude holders or inducing the critical attitude in half of the participants (cf. Fiedler et al., 2006). In contrast, in many diagnostic settings in reality, involving rare attributes like criminal or psychopathic personality with a p(A) of 10%, 5%, 1% or less, will produce a large number of false alarms whenever p(D) > p(A).

Our discussion of the asymmetry problem has a number of implications for future research, suggesting theoretical and empirical issues that have to be tackled in order to improve the diagnostic assessment of implicit attitudes.

(1) The first and foremost implication and recommendation, which is the focus of this chapter, is that studies of predictive validity must not ignore the asymmetry of causal and diagnostic inference. However familiar and customary the practice is to use correlation coefficients as indices of predictive validity, it has to be acknowledged that the predictive probability p(attitude | test) can differ radically from p(test | attitude). To the extent that an attitude is more likely diagnosed in a test than encountered in reality, the causal hit rate

p(test | attitude) will greatly over-estimate the diagnostic probability p(attitude | test) that a test score really reflects an attitude. As long as this asymmetry is ignored, the inferences from IAT or EP scores to latent attitudes does not rest on solid scientific ground.

(2) It should be obvious that asymmetry arises as a consequence of the way in which attitudes are defined. As in every scientific discipline, the definitions must be taken serious. Definitions cannot be wrong like empirical statements, but they can differ in usefulness and coherence with other established definitions. The minimal definition that is widely adopted in the recent literature – considering a mere target-valence association as a sufficient condition of an implicit attitude – is not coherent with the old definition of an attitude that was traditionally accepted and used in social psychology. The altered definition is much weaker and much less restrictive, covering many associations that would not have been qualified as an attitude by the old definition. This weak criterion for diagnosing an attitude creates the asymmetry problem by inflating the prevalence of diagnosed attitudes relative to the prevalence of corresponding criterion behaviors. Personally, I find both implications unfortunate in a cumulative science striving for coherence and precise communication. If the new "implicit attitudes" differ fundamentally from the old "explicit attitudes" not only in its implicitness but also in referring to a new attitude concept, this creates an unwanted source of confusion that complicates the interpretation of empirical findings.

(3) Even when researchers consider it useful and necessary to add a second attitude concepts, the viability of the new definition must be examined critically. Rather than uncritically adopting and reifying the definition, it should be put to empirical test. For instance, a glance at the cognitive association literature suggests a rather complex and tricky architecture of associative memory, as expressed in the following quotation from Maki and Buchanan (2008, p.): "Semantic similarity determined from lexicographic measures is shown to be separable from the associative strength determined from word association norms, and these semantic and associative measures are in turn separable from abstract representations

derived from computational analyses of large bodies of text. The three-factor structure is at odds with traditional views of word knowledge."

Attitude research must not ignore the state of the arts in the cognitive psychology of associations. It is well known that strong associates not only include synonyms and words of related and similarly valenced meaning but also antonyms and concepts of opposite meaning. How can we exclude that an association of Black and negative valence suggests represents an antonym rather than a synonym? How do we know which of the three dimensions distinguished by Maki and Buchanan (2008) is tapped by an associative measure? How do we know that an association reflects an individual's own valuation rather than cultural knowledge about others' valuation, or semantic knowledge?

(4) This reminder of the state of the art in fundamental cognitive psychology leads us immediately to another desideratum, namely, the need to formulate clearly spelled out psychometric models, the assumption of which can be falsified, rather than assuming uncritically that every instance of an IAT or EP effect will automatically yield an attitude (Blanton, Jaccard, Gonzales, & Christie, 2006). What precise cognitive algorithm can translate a latency difference into an attitude? What psychometric model implies that an average latency difference obtained on an IAT or EP task must reflect an association between attitude target and valence?

In the preceding section, we have seen that latency differences may originate in completely different processes, such as separate response biases for targets and the valence dimension. An IAT model must explain why a target category and a valence category, which can be easily sorted onto the same response keys, implies an association between target and valence (i.e., why does T → Key and V → Key imply T → V?). Similarly, an EP model would have to rule out the possibility that other processes than associative influences from a prime to a target can account for a priming effect. In particular, the model would need to rule out Ratcliff and McKoon's (1988) alternative account that a compound retrieval cue that

incorporates elements of both prime and target is responsible for the facilitation effect on priming tasks. Such a priming effect would not reflect a bilateral association of an attitude-target prime and a valence reaction but the effectiveness of a retrieval cue composed of a prime and a target.

Last but not least, an informed analysis of attitude assessment should not be confined to a few chronometric measures, which happen to be applied most easily. Rather, the research program should be open to all kinds of measurement (as shown in Table 1), and it should allow for a non-trivial test of the possibility that an attitude construct be falsified. Maybe social behavior is often not determined by stable internal dispositions within persons, but by the external influence of culture, social ecologies, task affordances, or the power of eliciting stimuli. It would be a category mistake to call such external determinants attitudes. Ironically, the most widely used instruments that have been proposed to measure (implicit) attitudes, IAT and EP, may turn out to play a major role in such a theoretical re-attribution of the origins of behavior.

References

Ajzen, I. (1991). The theory of planned behavior. *Organizational Behavior and Human Decision Processes, 50*, 179–211.

Balcetis, E., & Dunning, D. (2006). See what you want to see: Motivational influences on visual perception. *Journal of Personality and Social Psychology, 91,* 612-625.

Bargh, J.A. (1994). The Four Horsemen of automaticity. In R.S. Wyer & T.K. Srull (Eds.), *Handbook of Social Cognition* (pp. 1-40). Hillsdale, NJ: Erlbaum.

Blair, I. V. (2002). The malleability of automatic stereotypes and prejudice. *Personality and Social Psychology Review, 6*(3), 242-261.

Blair, I. V., Ma, J. E., & Lenton, A. P. (2001). Imagining stereotypes away: The moderation of implicit stereotypes through mental imagery. *Journal of Personality and Social Psychology, 81,* 828-841.

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

Bluemke, M., & Fiedler, K. (2009). *Base rate effects on the IAT*. Manuscript submitted for publication.

Bluemke, M., & Friese, M. (2006). Do features of stimuli influence IAT effects? *Journal of Experimental Social Psychology, 42,* 163-176.

Brauer, M., Wasel, W., & Niedenthal, P. (2000). Implicit and explicit components of prejudice. *Review of General Psychology, 4,* 79-101.

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

Chan, E., Ybarra, O., & Schwarz, N. (2006). Reversing the affective congruency effect: The role of target word frequency of occurrence. *Journal of Experimental Social Psychology,*

42, 365-372.

Clark, H. H. (1969). Linguistic processes in deductive reasoning. *Psychological Review, 76*, 387–404.

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

Devine, P. G. (1989). Stereotypes and prejudice: Their automatic and controlled components. *Journal of Personality and Social Psychology, 56*, 5-18.

Fazio, R. H. (2001). On the automatic activation of associated evaluations: An overview. *Cognition and Emotion, 15*, 115-141.

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

Fazio, R.H., Chen, J., McDonel, E.C., & Sherman, S.J. (1982). Attitude accessibility, attitude-behavior consistency, and the strength of the object-evaluation association. *Journal of Experimental Social Psychology, 18*, 339-357.

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.

Fiedler, K., Bluemke, M., & Unkelbach, C. (in press). Exerting Control over Allegedly Automatic Associative Processes. In Forgas, J.P., Baumeister, R., & Tice, D. *The psychology of self-regulation: Cognitive, affective, and motivational processes*. Cambridge University Press.

Fiedler, K., Bluemke, M., & Unkelbach, C. (2009). *Flexible cue utilization in evaluative priming*. Manuscript submitted for publication.

Fiedler, K., Freytag, P., & Bluemke, M. (2009). *Pseudocontingencies in evaluative priming*. Manuscript in preparation, University of Heidelberg.

Fiedler, K., Kaczor, K., Haarmann, S., Stegmüller, M., & Maloney, J. (in press). Impression Formation Advantage in Memory for Faces: When Eyewitnesses are Interested in Targets' Likeability, Rather than their Identity. *European Journal of Social Psychology*.

Fiedler, K., Messner, C., & Bluemke, M. (2006). Unresolved problems with the 'I', the 'A', and the 'T': A logical and psychometric critique of the Implicit Association Test (IAT). *European Review of Social Psychology, 17*, 74-147.

Förster, J., & Strack, F. (1997). Motor actions in retrieval of valenced information: A motor congruence effect. *Perceptual and Motor Skills, 85*, 1419–1427.

Frantz, C.M., Cuddy, A.J.C., Burnett, M., Ray, H., & Hart, A. (2004). A threat in the computer: The race Implicit Association Test as a stereotype threat experience. *Personality and Social Psychology Bulletin, 30,* 1611-1624.

Gackenbach, J. I. (1978). A perceptual defense approach to the study of gender sex related traits, stereotypes, and attitudes. *Journal of Personality, 46*, 645-676.

Gigerenzer, G. (1991). From tools to theories: A heuristic of discovery in cognitive psychology. *Psychological Review, 98*, 254-267.

Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instructions: Frequency Formats. *Psychological Review, 102*, 684-704.

Govan, C. L., & Williams, K. D. (2004). Reversing or eliminating IAT effects by changing the affective valence of the stimulus items. *Journal of Experimental Social Psychology, 40*, 357-365.

Greenwald, A. G., Banaji, M.R., Rudman, L.A., Farnham, S. D., Nosek, B.A., & Mellot, D. S. (2002). A unified theory of implicit attitudes, stereotypes, self-esteem, and self-concept. *Psychological Review*, *109*(1), 3-25.

Gumpper, D. (1972). Convergence between own-categories and binocular-rivalry measures of attitudinal direction. *Psychological Reports, 31,* 111-117.

Han, H. A., Olson, M. A., & Fazio, R. H. (2006). The influence of experimentally created extrapersonal associations on the Implicit Association Test. *Journal of Experimental Social Psychology, 42,* 259-272.

Klauer, K. C., Roßnagel, C., & Musch, J. (1997). List-context effects in evaluative primimg. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23,* 246-255.

Lepore, L., & Brown, R. (1997). Category and stereotype activation: Is prejudice inevitable? *Journal of Personality and Social Psychology, 72,* 275-287.

Lilienfeld, S. O., Wood, J. M., & Garp H. O. (2000). The scientific status of projective techniques. *Psychological Science in the Public Interest,* 1, 27-66.

Maki, W.S., & Buchanan, E. (2008). Latent structure in measures of associative, semantic, and thematic knowledge. *Psychonomic Bulletin & Review, 15,* 598-603.

Mehrabian, A. (1966). Immediacy: An indicator of attitudes in linguistic communication. *Journal of Personality, 34,* 26-34.

Mitchell, C. J. (2004). Mere acceptance produces apparent attitude in the Implicit Association Test. *Journal of Experimental Social Psychology, 40,* 366-373.

Nosek, B. A. (2007). Implicit-explicit relations. *Current Directions in Psychological Science, 16,* 65-69.

Nosek, B. A., Greenwald, A. G., & Banaji, M. R. (2006). The Implicit Association Test at age 7: A methodological and conceptual review. In J. A. Bargh (Ed.), *Automatic Processes in Social Thinking and Behavior* (pp. 265-292). New York: Psychology Press.

Parducci, A. (1968). The relativism of absolute judgment. *Scientific American, 19,* 84–90.

Pennebaker, J. W., Mehl, M. R., & Niederhoffer, K. (2003). Psychological aspects of

natural language use: Our words, our selves. *Annual Review of Psychology, 54,* 547-577.

Proctor, R.W., & Cho, Y.S. (2006). Polarity correspondence: A general principle for performance of speeded binary classification tasks. *Psychological Bulletin, 132,* 416-442.

Ratcliff, R., & McKoon, G. (1988). A retrieval theory of priming in memory. *Psychological Review,* 95, 385-408.

Rosenberg, M.J., & Hovland, C.I. (1966). *Attitude organization and change: An analysis of consistency among attitude components.* Oxford, England: Yale U. Press.

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.

Schacter**,** D.L. (1992). Understanding implicit memory: A cognitive neuroscience approach. *American Psychologist, 47,* 559-569.

Schotte, D.E., McNally, R.J., & Turner, M.L. (1990). A dichotic listening analysis of body weight concern in bulimia nervosa. *International Journal of Eating Disorders, 9,* 109-113.

Shanks, D. R., Tunney, R. J., & McCarthy, J. D. (2002). A reexamination of probability matching and rational choice. *Journal of Behavioral Decision Making, 15,* 233-250.

Snyder, M., & Swann, W.B. (1976). When actions reflect attitudes: the politics of impression management. *Journal of Personality and Social Psychology, 34,* 1034–1042.

Spruyt, A., Hermans, D., De Houwer, J., Vandromme, H., & Eelen, P. (2007). On the nature of the affective priming effect: Effects of stimulus onset asynchrony and congruency proportion in naming and evaluative categorization. *Memory and Cognition, 35,* 95-106.

Swets, J. A., Dawes, R. M., & Monahan, J. (2000). Psychological science can improve diagnostic decisions. *Psychological Science in the Public Interest, 1,* 1-26 [whole issue].

Steele, C.M. (1997). A threat in the air: How stereotypes shape intellectual identity and performance. *American Psychologist, 52,* 613-629.

Tajfel, H., Flament, C., Billig, M. G., & Bundy, R. P. (1971). Social categorization and intergroup behaviour. *European Journal of Social Psychology, 1*, 149–178.

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

Unkelbach, C., Fiedler, K., Bayer, M., Danner, D., & Stegmüller, M. (2007). Why positive information is processed faster: The density hypothesis. *Journal of Personality and Social Psychology, 95*, 36-49.

Weaver, C.N. (2008). Social distance as a measure of prejudice among ethnic groups in the United States. *Journal of Applied Social Psychology, 38*, 779-795.

Wells, G.L. & Olsen, E.A. (2003). Eyewitness testimony. *Annual Review of Psychology, 54*, 277-295.

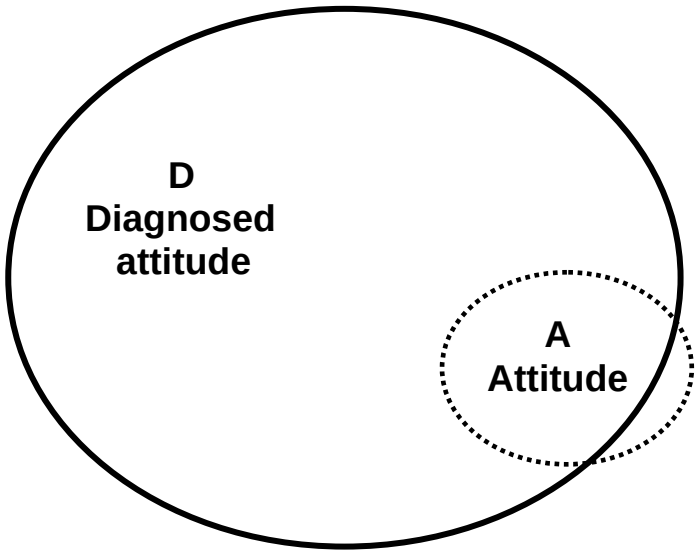
Wilson, T. D., Lindsey, S., & Schooler, T. Y. (2000). A model of dual attitudes. *Psychological Review, 107*, 101-126.

Wittenbrink**,** B., & Schwarz, N. (2007). *Implicit measures of attitudes*. New York, NY, US: Guilford Press.

Table 1: Taxonomy of existing procedures for attitude assessment. The number of references found in PsychInfo (after 2000) for each procedure is indicated in brackets

Rationale for attitude inference	Scaling method	
	Direct	Indirect
Auto-expression	Likert-scaled attitude *questionnaires* [15312] *Observation techniques* [82] Analysis of *facial expression* [94]	*Thurstone*-scaled measurement [13] *Guttman* scalogram analysis [30] *Rasch* scaling [96]
Approach-avoidance	*Arm movement* [11] *Immediacy* [72] Linguistic *abstractness* [9] *Eye-tracking* [3] *Binocular rivalry* & attitude [3] *Dichotic listening* [8] *Polygraph* [29]	*Perceptual defense* [1]
Representation in associative memory	*Projective test*s [15] *Incomplete sentence blank* [3] *Implicit Association Test* [224] *Evaluative* or *affective Priming* [52]	

Note: PsychInfo inquiries were always based on the italicized keywords in conjunction with the term "attitude".



According to Bayes theorem:

$$p(A \mid D) = p(D \mid A) \cdot p(A) / p(D)$$

Figure 1: If the base rate p(D) of a diagnosed attitude is higher than the base rate p(A) of the attitude proper, diagnostic inferences of A from D are less likely to be true than causal inferences of D from A.