

Implicit Prejudice: Pentimento or Inquisition?

A few years ago I attended a Person Memory Interest Group (PMIG) pre-conference at which the IAT featured prominently. At the end of PMIG Brian Nosek and Wil Cunningham stopped to ask my opinion of the IAT. As they put it, “Most people are either for or against it, but we can’t decide where you stand.” This comment reveals a striking uniqueness about the IAT in recent social psychological history. Rarely has a methodological tool garnered such strong adherents and detractors. Scientists are often emotional advocates or critics of new theoretical approaches, but methods are typically less divisive. Why has the IAT led to such polarization of the research community?

With regard to positive feelings, I think the reasons for the excitement are clear. The IAT provides a clever and simple procedure for measuring implicit attitudes that typically generates huge effect sizes (more on this issue later). Indeed the IAT effect is so strong that it is unique among implicit measures in that naïve participants can discern its purpose by virtue of the difficulty they experience during the “incompatible” trials (more on this later too). Like many social psychologists, I use a version of the IAT in which students tap on their desks as a classroom demonstration of implicit ageism, and nervous laughter erupts every year when students first attempt the incompatible trials and are suddenly and audibly slowed down. For these reasons, many of us are excited about the potential of the IAT to expand the study of implicit social cognition. We still do not really understand what it reveals and why it works (although Riki Conrey, Jeff Sherman, and their colleagues (2004) have some promising models that separate its component processes), but the IAT may yet prove to be one of the most important methodological advances in social cognition.

Negative attitudes toward the IAT are also pervasive, however, and Arkes and Tetlock have clearly articulated why. To label an unconscious response, and especially one that is still poorly understood, as *prejudice* strikes many as irresponsible (in this regard, I too am guilty,

as my colleagues and I have often used that label; e.g., von Hippel et al., 1997). The use of this label also suggests that political views are leaking into psychological research, as the agenda served by implicit prejudice research is largely a liberal one. So, with the proviso that I believe that Arkes and Tetlock have done the field a great service by opening debate on the meaning of *implicit prejudice*, let me turn now to the specifics of their argument. I focus on areas in their article where I think they have stretched an argument too far and on areas where I think they have not gone far enough.

Arkes and Tetlock are too bold ...

Rationality of prejudice: In their discussion of Expected Utility Theory, Arkes and Tetlock suggest that prejudice and discrimination can be rational if societal-level data suggest that members of different groups are likely to behave in different ways. They provide a compelling example of the nervous Reverend Jackson, and mathematically derive that he should be much more likely to flee when followed by a black than a white pedestrian at night. But their example is missing two ingredients that are common in real-life applications of this principle: individuating information, which can be much more diagnostic than category-level information, and cost to the target.

Regarding individuating information, let me relate an incident involving one of my students in a course on prejudice. After class one day she went to withdraw cash, but became nervous when a group of young black males was hanging around the ATM. Remembering my exhortations from lecture, she decided to be “unbiased” and withdrew her money, at which point she was robbed. I asked her to describe the men who were hanging out by the ATM, and then asked her if she would have withdrawn the money if they looked and behaved identically but were white. She said she would not, so I suggested that these individuating cues were the relevant data that should have influenced her behavior. The fact that the men had black skin was far less predictive of their probability of taking her money than the fact

that they dressed and acted like thugs. Bending over backward to be nonprejudiced in this circumstance was clearly a poor idea, but only because she relied on category-level data rather than the more diagnostic individuating cues that were available.

We also need to consider more fully the cost to targets of discriminatory behavior. There is virtually no cost in the “nervous pedestrian” example, but disutility rises when targets suffer (as Tetlock once pointed out to me in a discussion of liberty vs. responsibility, my right to swing my fist ends at your chin). For this reason, the rationality of discrimination is far less clear in the case of the “nervous employer”, the “nervous landlord”, or the “nervous banker”. Even if we place no utility on egalitarian concerns, rationality suggests that failure to employ or house a significant percentage of the population will create a self-fulfilling prophecy that will be costly to all members of society.

Discounting: Arkes and Tetlock point out that it is irrational not to discount ability attributions to recipients of affirmative action. They raise an excellent point, but they fail to note that affirmative action is a psychological sword that cuts only one way. Many members of society are beneficiaries of affirmative action, but only members of stigmatized groups suffer ability discounting as a consequence. Recipients of affirmative action who are not chronically stereotyped as inferior are generally unbothered when they receive affirmative action and their abilities are not discounted by others (see Crocker et al., 1998). Thus, I would agree that discounting of minority affirmative action recipients should be at least partially offset by augmenting for obstacles overcome, whereas discounting without augmentation should be applied in corporate cases of hiring personal or family connections, and university cases of legacy admits, student athletes, and all the other beneficiaries of preferential admission and hiring practices who are not visibly identifiable by virtue of their membership in chronically stigmatized groups. The fact that stigmatized recipients of affirmative action themselves tend to discount other ingroup recipients is greater testimony to the power of stigmatization than it

is evidence of appropriate discounting.

Problem of relativity: Arkes and Tetlock suggest that we ought not use the label of *prejudice* if we cannot distinguish on the IAT whether performance reflects liking of both groups but preference for one's own group vs. liking of one's own group and disliking of the outgroup. Although disliking an outgroup is more clearly prejudicial than simply liking one's own group more, there is a long tradition of treating relative preferences as prejudicial. Allport (1954) derives from Spinoza the suggestion that "love prejudice" (ingroup positivity) is not only more prevalent than "hate prejudice" (outgroup negativity), but is also the source from which hate prejudice springs. More recently a number of social psychologists have suggested that various types of discrimination previously thought to arise from outgroup negativity may instead be a function of relative ingroup positivity (Brewer, 2001). Thus, the inability of the IAT to distinguish between outgroup negativity and ingroup positivity limits its utility, but does not mean that it is not tapping prejudice. Additionally, it should be noted that Karpinski and Steinman (under review) have adapted the IAT into the Single Category Association Test (or SCAT), which appears to have promise for separating ingroup and outgroup components of implicit prejudice.

Arkes and Tetlock are too timid ...

Measures ≠ Constructs: In order to study a construct it is, of course, necessary to measure it. But intense focus on a measure carries with it the risk of unintentionally conflating the measure with the construct. A great deal of evidence suggests that implicit attitudes do exist, but that does not mean that factors that influence a particular measure of implicit attitudes necessarily influence the attitude itself. Researchers are well aware of this distinction in the case of explicit attitudes, as few would argue that an explicit attitude has been changed if people report different attitudes when under duress. Nevertheless, many accept that an implicit attitude has changed if people show movement on the IAT. For example, consider the

following experiments: 1) Dasgupta and Greenwald (2001) found that exposure to admired blacks and disliked whites led to a reduction in the typical race IAT effect for up to 24 hours. 2) Lowery et al. (2001) demonstrated reduced implicit prejudice when white participants were tested by a black rather than a white experimenter. 3) Blair and Lenton (2001) demonstrated that implicit stereotyping is reduced when people imagine counter-stereotypic targets prior to the implicit stereotyping measurement. Blair and Lenton's research is noteworthy in that they measured implicit stereotyping via the IAT, the GNAT, and the Deese/ Roediger-McDermott (DRM) false memory paradigm, with similar results across all three measures.

While these findings are provocative and incredibly interesting, it is unclear what they mean. They show that the IAT and other implicit measures are malleable to manipulations of accessibility, but whether that indicates that implicit attitudes are similarly malleable is a separate question. My bet is that the intuition many of us had when this work began – that implicit attitudes are much harder to change than explicit attitudes – may indeed be true, and that these data may not directly implicate change in implicit attitudes themselves. Rather, these data may demonstrate that implicit attitude measures are as easy to move around as explicit ones (although different procedures are required), and that movement on the measure does not necessarily indicate movement in the underlying attitude.

An effect that is too big to be true probably isn't: Why do most Whites and Blacks show anti-Black bias on the IAT? Arkes and Tetlock expand on Karpinski and Hilton's (2001) suggestion that the race IAT taps familiarity with or exposure to cultural stereotypes, rather than an implicit attitude toward Blacks. An alternative, although not mutually exclusive, perspective can be found in a recent paper by Kinoshita and Peek-O'Leary (under review). Kinoshita and Peek-O'Leary suggest that compatibility effects in the IAT between *pleasant* and a target category such as *white* could arise in part from the default nature of the category *white* relative to the contrasting category *black*, rather than reflecting a conceptual association

between the target category and pleasantness. In support of such an account, they replicate the insect/flower IAT effect (albeit with a reduced effect size) when the intervening *pleasant/unpleasant* judgment is replaced with a *word/not word* judgment.

These data suggest that caution should be exercised in assuming that a race IAT effect that is greater than zero reflects differential implicit preference for Whites vs. Blacks, as figure-ground asymmetries appear to produce a reliable IAT effect when no evaluation is involved. An important advantage of the IAT was that it seemed to be a ratio scale, with a true zero value, but it may be the case that recalibration is necessary to recover this true zero value. In the meantime, the finding that a majority of respondents show race bias in the IAT is open to alternative interpretation.

The location of the true zero value of the IAT is further clouded by research that suggests that the IAT can itself induce stereotype threat in white students. Specifically, Frantz et al. (in press) showed that white students find the IAT threatening if they are told that it is a measure of prejudice, or if they ascertain that themselves as they take the test. In this case, stereotype threat emerges from participants' concern that they will be perceived as prejudiced because they are white. This threat disrupts performance on the IAT, much as it does with other groups in other performance domains, by causing people to provide even larger IAT effects than they would otherwise. Again, these findings call into question the interpretation of effect size and the location of the true zero value on the IAT.

A shameless plug for my own research

In contrast to the dominant methods for studying implicit attitudes in social psychology, our approach to the study of implicit attitudes has its intellectual roots in the work of Roediger (1990), who suggested that the implicit/explicit memory distinction might be best understood by focusing on different processes rather than different systems. Roediger and his colleagues suggested that one way to understand the dissociations that emerged between meas-

ures in this literature was to focus on the type of information processing that the measures involved. This work clearly demonstrated that some (but not all) of the variance in the measures was a function not of their “implicitness”, but of the nature of the task demands required by the measures.

Similar to this perspective, we have proposed that one important issue is whether a measure taps into biased information processing. In our research we have found that people who show a linguistic bias with regard to African Americans also evaluate a black but not a white male as more threatening than people who do not show the linguistic bias (von Hippel et al., 1997). In that same paper we found that people who show the linguistic bias are also likely to show an attributional bias, which we now refer to as the Stereotypic Explanatory Bias (SEB; Sekaquaptewa et al., 2003; Sekaquaptewa & Espinoza, 2004). Sekaquaptewa and her colleagues, in the aforementioned papers, have demonstrated that the SEB predicts whether people choose to ask stereotypic questions of Blacks but not Whites in a mock job interview and whether they have negative interactions with Blacks but not Whites in an unstructured setting. It is worth noting, in light of Arkes and Tetlock’s criticisms, that the “negative interactions” were a blend of liking by the confederate and nonverbal behavior of the white participant (as reported by the confederate). This combined score had high reliability, but because it was based only on confederate ratings, it does not really address the issue raised by Arkes and Tetlock regarding the ambiguity of nonverbal behavior as an indicator of animosity.

It is our belief that these measures, and similarly derived ones we have developed in the area of attitudes (Vargas et al., in press) and the self (von Hippel et al., 2004), have the potential to supplement the IAT and affective priming procedures to provide a more complete picture of people’s unintended cognitive and behavioral responses to others. Undoubtedly our measures are also rife with interpretive ambiguities, but because they rely on a very different set of procedures and assumptions, they could prove to be a useful addition to the current im-

PLICIT measurement menagerie. Additionally, because our procedures are much more deliberative than the IAT and affective priming, they also have the potential to broaden the scope of investigation from automatic stereotyping and prejudice to more thoughtful if not fully conscious processes.

Conclusions

Arkes and Tetlock clearly describe the pitfalls and consequences of the label *implicit prejudice*. They then contrast survey research, which shows great declines in prejudice, with implicit prejudice research, which shows a preponderance of prejudice. Readers are asked to choose which indicator is more valid. This concluding question brings me to my final point as well: just because we are not sure what the IAT measures does not mean we should accept people's survey responses at face value. First, despite nearly universal self-reported egalitarianism, behavioral measures gathered by testing institutes (such as those sponsored by the Department of Housing and Urban Development) continue to document discrimination in critical areas such as housing, hiring, and bank loans. This tension between self-report and behavior can be seen in the ABC News program "True Colors" when a landlord denies prejudicial motives moments after refusing to show an apartment to an African American Yuppie that he had just shown to a similar white applicant. As social psychologists, we have a long tradition of believing behavior when it contradicts self-report. Second, a number of studies suggest that among people who typically show no sign of prejudice in self-report or behavior, prejudice can easily rise to the surface when they feel threatened or insecure (Fein & Spencer, 1997). The fact that stereotyping and prejudice automatically manifest themselves among otherwise unprejudiced people in such circumstances (Spencer et al., 1998) suggests that implicit prejudices may indeed be lurking in the hearts and minds of many if not all of us, and may indeed be appropriately labeled *prejudice*.

Note: Correspondence should be addressed to William von Hippel, School of Psychology,

University of New South Wales, Sydney, 2052, Australia (w.vonhippel@unsw.edu.au).

References

- Allport, G. (1954). *The nature of prejudice*. Cambridge, MA: Addison-Wesley.
- 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.
- Brewer, M. B. (2001). Ingroup Identification and Intergroup Conflict: When Does Ingroup Love Become Outgroup Hate? In R. Ashmore, L. Jussim, & D. Wilder (Eds.), *Social Identity, Intergroup Conflict, and Conflict Reduction* (pp. 17-41). Oxford University Press.
- Conrey, F. R., Sherman, J. W., Gawronski, B., Hugenberg, K., & Groom, C. (2004). *Separating multiple processes in implicit social cognition: The Quad-Model of implicit task performance*. Unpublished manuscript, Northwestern University, Chicago, IL.
- Crocker, J., Major, B., & Steele, C. (1998). Social stigma. In D. T. Gilbert, S. T. Fiske, & G. Lindzey (Eds.), *Handbook of social psychology*. (pp. 504-553), New York: McGraw-Hill.
- Dasgupta, N., & Greenwald, A. G. (2001). On the malleability of automatic attitudes: Combating automatic prejudice with images of admired and disliked individuals. *Journal of Personality and Social Psychology*, *81*, 800-814.
- Fein, S., & Spencer, S. (1997). Prejudice as self-image maintenance: Affirming the self through derogating others. *Journal of Personality and Social Psychology*, *73*, 31-45.
- Frantz, C. M., Cuddy, A. J. C., Burnett, M., Ray, H., & Hart, A. (in press). A threat in the computer: The race Implicit Association Test as a stereotype threat experience. *Personality and Social Psychology Bulletin*.
- Karpinski, A., & Hilton, J. L. (2001). Attitudes and the Implicit Association Test. *Journal of Personality and Social Psychology*, *81*, 774-788.
- Karpinski, A., & Steinman, R. B. (2004). *Associate strength measures of attitudes: A*

comparison of the Single Category Association Test to the Implicit Association Test. Unpublished manuscript, Temple University.

Kinoshita, S., & Peek-O'Leary, M. (2004). *Implicit Association Test (IAT): Support for the figure-ground asymmetry account*. Unpublished manuscript, Macquarie University.

Lowery, B. S., Hardin, C. D., & Sinclair, S. (2001). Social influence effects on automatic racial prejudice. *Journal of Personality and Social Psychology*, *81*, 842-855.

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

Sekaquaptewa, D., & Espinoza, P. (2004). Biased processing of stereotype-incongruency is greater for low than high status groups. *Journal of Experimental Social Psychology*, *40*, 128-135.

Sekaquaptewa, D., Espinoza, P., Thompson, M., Vargas, P., & von Hippel, W. (2003). Stereotypic explanatory bias: Implicit stereotyping as a predictor of discrimination. *Journal of Experimental Social Psychology*, *39*, 75-82.

Spencer, S. J., Fein, S., Wolfe, C., Fong, C., & Dunn, M. (1998). Stereotype activation under cognitive load: The moderating role of self-image threat. *Personality and Social Psychology Bulletin*, *24*, 1139-1152.

Vargas, P. T., von Hippel, W., & Petty, R. E. (2004). Using "partially structured" attitude measures to enhance the attitude-behavior relationship. *Personality and Social Psychology Bulletin*, *30*, 197-211.

von Hippel, W., Sekaquaptewa, D., & Vargas, P. (1997). The Linguistic Intergroup Bias as an implicit indicator of prejudice. *Journal of Experimental Social Psychology*, *33*, 490-509.

von Hippel, W., Lakin, J. L., & Shakarchi, R. J. (2004). *Individual differences in motivated social cognition: The case of self-serving information processing.* Manuscript under review.