On the Methodological Difficulty of Identifying Implicit Racial Beliefs and Stereotypes

Moa Bursell and Filip Olsson

Abstract
In “Status Characteristics, Implicit Bias, and the Production of Racial Inequality,” Melamed, Munn, Barry, Montgomery, and Okuwobi present an innovative and intriguing study on social influence, status beliefs, and implicit racial bias. To capture status-based expectancies, the authors measure implicit racial status beliefs using an Implicit Association Test (IAT) with words related to high and low status. We identify an important flaw in the study’s analytic approach that severely limits the conclusions that can be drawn based on the study. We argue that the authors neglected to separate the valence of the words included in the racial status IAT with the stereotype content of these words. It is therefore possible that the study’s racial IAT only captures implicit racial evaluations, and not status-based implicit racial beliefs.

Keywords
implicit bias, Implicit Association Test (IAT), attitudes, stereotypes, status characteristics theory (SCT)

In Melamed, Munn, Barry, Montgomery, and Okuwobi’s (2019) innovative and intriguing “Status Characteristics, Implicit Bias, and the Production of Racial Inequality,” they derive hypotheses from status characteristics theory (SCT) to explain how racial inequality is reproduced via status-based expectations in interracial interactions. Melamed and colleagues empirically evaluate how explicit and implicit racial status beliefs, that is, beliefs about racial groups having different status, mediate racial differences in individuals’ susceptibility to social influence from an other-race partner in a collaborative task. The introduction of a “new way to measure implicit racial status beliefs” (Melamed et al. 2019:1013) is highlighted as one of the study’s main contributions.

However, we have identified a serious flaw in the study’s analytic approach, which concerns how the authors identify implicit racial status beliefs by conducting a stereotype Implicit Association Test (IAT). Because the IAT is a crucial component in the study’s research design, we argue that it is only possible for Melamed and colleagues to evaluate the first of the article’s three hypotheses. We anticipate an increase in studies based on the IAT in sociology in the coming years, so we think it is important to address basic validity issues at an early stage of the incorporation of the IAT technique into sociology.

Institute for Futures Studies and Stockholm University
Stockholm University

Corresponding Author:
Moa Bursell, Institute for Futures Studies, 101 31 Stockholm; Department of Sociology, Stockholm University, Stockholm, 106 91
Email: moa.bursell@sociology.su.se
Drawing on SCT, Melamed and colleagues’ Hypothesis 1 states, “White participants will resist influence from black participants more than black participants will resist influence from white participants while working on a collectively-oriented task” (Melamed et al. 2019:1015–16). To identify the mechanism linking race and social influence, the authors stipulate Hypothesis 2b (2a is not tested, see Melamed et al. 2019:1031): “first-order status beliefs, as measured by the racial status IAT, will mediate the effect of race on social influence” (Melamed et al. 2019:1019). For unacquainted readers, SCT distinguishes between first-order (“I think that Whites have higher status...”) and third-order (“Most people think that...”) status beliefs. In their third hypothesis, Melamed and colleagues stipulate that if first- and third-order beliefs contradict, the latter will guide behavior: “Third-order status beliefs moderate the mediation effect of first-order status beliefs on behavioral inequalities” (Melamed et al. 2019:1019).

To evaluate their hypotheses, they conduct two studies. In Study 1, they construct and validate what they call a “racial status IAT” (we adopt this language throughout), a measure of first-order racial status beliefs, using participants from Amazon Mechanical Turk. The IAT relies on participant response times to measure one of two things: implicit attitudes or implicit stereotypes. Implicit attitudes are associations of social groups with positive or negative evaluations or valence, such as associating White, thin, heterosexual, young, or able people with positivity, or “good”; and stigmatized minorities, obese, homosexual, old-aged, or disabled people with negativity, or “bad.” Implicit stereotypes reflect associations of social groups with semantic attributes, such as associating women with liberal arts and the family domain, and men with science and the career domain. It is difficult, however, to accurately measure implicit racial stereotypes, as stereotypes tend to also be positively and negatively valenced (Kurdi et al. 2019; Schnabel, Asendorpf, and Greenwald 2008). When stereotypes are valenced, the IAT score might reflect valenced attitudes, rather than specific stereotypes or beliefs.

How then do researchers discern distinct implicit stereotypes about a social category, from an implicit attitude? Melamed and colleagues do not address this issue in their study. They simply design a racial status IAT by replacing “good” and “bad” with “high status” and “low status,” and they use words commonly associated with high status and low status to measure “first-order implicit racial status beliefs” (see their appendix for further details, Melamed et al. 2019:1032). They find positive scores for White, Asian, and Hispanic people and negative scores for Black people, which they interpret to indicate first-order implicit beliefs among White, Asian, and Hispanic individuals that White people are higher status than Black people.

The problem with this approach is that the status-related words in Melamed and colleagues’ IAT are systematically valenced: for example, “intelligent,” “worthy,” and “better” are positively valenced, whereas the low-status words (e.g., “unintelligent,” “unworthy,” and “worse”) have a negative valence. There is thus a significant possibility that their IAT results reflect implicit attitudes rather than implicit stereotypes. The IAT score might partially reflect both the semantic and evaluative properties of the words being used, but participants might also simply recode attributes such as “intelligent” and “unintelligent” into “positive” and “negative” (Schnabel et al. 2008).

In line with this, Kurdi and colleagues (2019) show how the variance of a stereotype race IAT was accounted for by the variance of an attitude race IAT (good/bad), but primarily if the stereotype IAT contained highly valenced words (e.g., smart/dumb). Even when measuring a seemingly incorrect or uncommon stereotype (Asian–dumb), the IAT produced a positive score (indicating negative stereotypes against Asians) for White participants, which was largely explained by the outcome of the attitude race IAT (White–good, Asian–bad).

The problem with Melamed and colleagues’ approach, then, is that researchers will likely observe higher scores for White
respondents on a racial (White–Black) IAT that uses valenced words because of the nature of how Black people are often evaluated by White people in the United States. To argue that one has identified implicit beliefs because of a positive score for White respondents, and a negative score for Black respondents, is thus insufficient. We may expect a positive score for White respondents, and a lower or negative score for Black respondents, even if none of the groups have implicit racial status beliefs, simply because of the valence of the words included in the racial status IAT (Kurdi et al. 2019). Because Melamed and colleagues are not taking valence into account, it still remains to be shown that their implicit racial status IAT is measuring something different than a standard racial attitude IAT, that is, implicit racial bias.

Melamed and colleagues do discuss the validity of their IAT, but they focus on the correspondence between their findings and results from previous research using the IAT. Because their findings replicate some patterns that are often observed in racial attitude IATs, such as ideological (higher scores for Republicans) and racial (higher scores for White, Asian, and Hispanic individuals, and low scores for Black individuals) differences, and a positive correlation between implicit and first-order explicit belief measures at conventional levels (\(r = .19\)), they conclude that their measure is valid (Melamed et al. 2019:1022).

But replicating well-known patterns of an established measure does not guarantee validity when constructing a new measure. On the contrary, similar patterns for the attitude race IAT and the racial status IAT may indicate they measure the same, and not distinct, constructs. Thus, we would expect that using an attitude IAT would produce scores highly correlated with those of the racial status IAT, and similarly correlated with the explicit measures of first-order beliefs.

After Melamed and colleagues conclude their measure of implicit racial status beliefs is valid, they move on to Study 2, where they conduct a laboratory study on race and social influence, in which participants are led to think they will work in interracial pairs to conduct an evaluative, collaborative task online. In accordance with the predictions of SCT, Black participants were more susceptible to social influence from a White partner than vice versa, that is, they found clear evidence confirming Hypothesis 1. Because this result does not depend on the validity of the racial status IAT, this conclusion is unaffected by the conflation of implicit attitudes with implicit stereotypes.

In Study 2, participants completed either the racial status IAT or a control “cats and dogs” IAT that compares racial differences in the implicit evaluations of cats and dogs using standard evaluative words such as “pleasure,” “beautiful,” “rotten,” and “tragic” (Melamed et al. 2019:1032). The results of the racial status IAT in Study 1 were reproduced: White participants associated White faces with high status, and Black participants associated Black faces with high status. Importantly, Melamed and colleagues found a negative effect of the racial status IAT on social influence (Table 5, Model 6, p. 1026). That is, controlling for race, individuals who scored high on the racial status IAT were less susceptible to being influenced by their “partner’s” evaluation. The authors interpret this as support for Hypothesis 2b, that first-order status beliefs, as measured by the racial status IAT, will mediate the effect of race on social influence (Melamed et al. 2019:1031). As we argued earlier, because Melamed and colleagues have not shown that their IAT is a measure of implicit racial status beliefs, this conclusion does not follow. We find it likely that if the authors used the evaluative words of the cats and dogs IAT in their racial status IAT, it would generate similar scores. What Melamed and colleagues have shown is that implicit racial bias, of undifferentiated nature, mediates the effect of race on social influence.

To address Hypothesis 3, a subset (42 participants) of the group that participated in the implicit racial status IAT also filled out a questionnaire including the explicit first-order status beliefs measure identified in Study 1, and an additional measure of explicit
third-order status beliefs. Analyzing these outcomes, Melamed and colleagues find support for Hypothesis 3, that third-order status beliefs mediate the effect that implicit first-order status beliefs have on social influence (see their Model 8 in Table 5, the discussion on 1028–29, and Figure 5). Again, we argue that this is not what they have shown. They have shown that third-order status beliefs mediate the effect that implicit racial bias has on social influence.

Linking implicit racial bias with explicit attitudes and racial differences in the outcomes of the collaboration task is an intriguing result, and there are still relatively few studies that link implicit racial bias and social influence. Unfortunately, however, this does not identify the “underlying cause” with the precision that Melamed and colleagues set out to do (p. 1016). We still know little about the nature of the implicit bias captured by the IAT.

Additionally, Melamed and colleagues’ interpretation of the IAT as a measure of first-order beliefs is a bit bold. There is no consensus within the IAT research field that the IAT measures “I” associations rather than “most people” associations or even “situations” (Arkes and Tetlock 2004; Payne, Vuletich, and Lundberg 2017). Melamed and colleagues’ results would be difficult to make sense of in terms of implicit third-order beliefs, because this would result in contradictory patterns for Black participants (who expressed in-group preference on the racial status IAT, and explicit third-order beliefs that favored White people), but their results may at least, to some extent, be mediated by unobserved situational factors.

A proper validation of Melamed and colleagues’ IAT might show that it does measure implicit racial status beliefs. But a significant amount of work remains before it can be argued with any confidence that this is the case. In the remainder of this comment, we will focus on the possibility of identifying implicit racial status beliefs among White individuals in the United States. One approach is to expand on Melamed and colleagues’ research design by comparing a group of participants’ results on a standard attitude race IAT (White–Black, good–bad) with the racial status IAT, similar to the procedure in Rudman and Ashmore (2007) and Amodio and Devine (2006).

If the two tests are weakly correlated, and if the tests differ in their ability to predict behavioral outcomes, it might indicate that the attitude race IAT and the racial status IAT measure two distinct constructs (provided that other factors, such as valence strength and word length/complexity, are kept constant, and that the order in which the tests are taken is randomized). Such a procedure, however, would be limited in its ability to establish whether or not we are measuring implicit status beliefs, as it might instead reflect other unobserved differences between the tests.

The next step would be to identify a racial minority group, Minority X, that (1) generates a similar outcome in an attitude race IAT (White–Minority X, good–bad), but (2) is not associated with the same negative status beliefs as Black persons. We would then compare how a new set of participants (who conduct the IATs in a randomized order) score on the original racial status IAT (White–Black) with how they score on a White–Minority X racial status IAT. Table 1 illustrates a hypothetical outcome that would confirm, or at the very least significantly strengthen, the validity of Melamed and colleagues’ racial status IAT. If the same participants score significantly lower in the White–Minority X racial status IAT than in the White–Black racial status IAT, then it would provide a significantly stronger indicator that the racial status IAT taps White respondents’ implicit status beliefs about Black people. The practical challenge of this analytic strategy will be to identify a group that meets the criteria of Minority X.

Previous research, primarily based on the Stereotype Content Model (Cuddy, Fiske, and Glick 2008), has successfully detailed how several social groups do differ in terms of perceived warmth and competence. Asian people, for example, are traditionally associated with high competence and low warmth in the U.S. context, whereas immigrants from the Middle East are associated with low competence and low warmth. Because previous
research shows that the “competence” dimension correlates strongly with conceptions of “status” (e.g., Fiske 2015), SCM may provide a good starting point for identifying a group that matches the requirements of Minority X.

Another course of action would be to construct a racial status IAT with words that are either valence neutral or valence balanced (Schnabel et al. 2008). Such a test would consist of words that are synonymous with the words used in the racial status IAT but have matching or neutral valence. Rudman, Greenwald, and McGhee (2001) used this approach in an early study of implicit gender stereotypes, where words such as “weak” were exchanged for words like “delicate.” Similarly, Amodio and Devine (2006) measured racial stereotypes by comparing two positively valenced categories—athleticism and intelligence (for potential problems with this approach, see Kurdi et al. 2019).

Summing up, Melamed and colleagues’ study provides an interesting contribution concerning the link between implicit racial bias, social influence, and unequal outcomes. But a significant amount of work remains before we can identify a more precise mechanism, such as status beliefs, as the driver of these implicit biases.

The takeaway for future research on implicit racial bias or beliefs/stereotypes is that developing new IAT measures requires meticulous validation, and we should be careful in our interpretation of IATs. We are pleased to see that an increasing number of sociologists are incorporating the IAT methodology into their research to improve our understanding of some of sociology’s core questions. It is crucial, however, that we recognize the possibilities as well as the limitations of the IAT methodology.

To the extent that we are merely interested in the role of implicit bias, we can adopt standard IATs and draw on established recommendations for interpretation and validity checks (see, e.g., Greenwald, Nosek, and Banaji 2003). However, more ambitious and innovative approaches— Attempts to understand the meaning of these biases—will have to control for a range of confounders (see, e.g., Kurdi et al. 2019). Sociologists interested in implicit beliefs may also find it necessary to take a position in the ongoing debate on whether implicit bias is associative or propositional (Bursell and Olsson 2020). We hope our comment on Melamed and colleagues’ ambitious and inspiring work will contribute to an open and constructive discussion on how best to incorporate the IAT into sociological studies on social inequality and other related fields.

### Funding

This work has been funded by The Swedish Research Council for Health, Working Life and Welfare (DNR 2016-00173).

### ORCID iDs
Moa Bursell [https://orcid.org/0000-0002-8289-4401](https://orcid.org/0000-0002-8289-4401)
Filip Olsson [https://orcid.org/0000-0001-9888-071X](https://orcid.org/0000-0001-9888-071X)

### Notes

1. An additional problem is that only two of the word-pairs used to capture status in the explicit third-order belief measure overlap with those used in the racial status IAT (see Melamed et al. 2019:1024, 1032).

2. There is ongoing discussion in psychology about what type of inner state the IAT measures. Greenwald and colleagues, the creators of the IAT, contend that both implicit attitudes and stereotypes are implicit associations, and thus not assigned a truth value (Greenwald and Banaji 1995). Another position, which we have been influenced by, argues that the

### Table 1. Hypothetical Outcome (IAT Score) That Would Confirm the Validity of Melamed and Colleagues’ Racial Status IAT

| Race IAT (Good–Bad) | Racial Status IAT (High–Low Status) |
|---------------------|-------------------------------------|
| White–Black         | .4                                  |
| White–Minority X    | .4                                  |
|                     | .6                                  |
|                     | .4                                  |
IAT may capture implicit propositional beliefs (see Hughes, Barnes-Holmes, and De Houwer 2011). According to the propositional account, implicit bias does not only consist of associations (e.g., associating “White people” with “intelligent”), but it can also contain information on how two concepts are related (e.g., “White people are intelligent”). It is clear, though, that the link between outcomes of the IAT and the inner state it corresponds to has not been convincingly established (for an insightful discussion, see Fazio and Olsen 2003).

3. The racial status IAT includes the word-pairs worthy–unworthy and better–worse. Previous work on racial and moral boundaries suggests Black Americans contest the relationship between high socioeconomic status and moral worthiness (Lamont et al. 2016:57). “Worthiness” is also a characteristic that sorts under “warmth” in the stereotype content model’s warmth–competence dimensions (see Fiske 2015), a literature that shares many similarities with SCT. Using words that are ambiguous in relation to these dimensions, such as worthy–unworthy, better–worse, in a status IAT may cause systematic bias in the data, because socioeconomically subordinate groups have been found to assign warmth–characteristics higher importance in the evaluation of individuals and groups (Fiske 2015). For simplicity, we therefore focus our continued discussion on White individuals’ implicit racial status beliefs. This concern could, however, be easily addressed in future studies by conducting a pre-study in which participants score how they associate a broad set of potentially status-relevant words.

References

Amodio, David M., and Patricia G. Devine. 2006. “Stereotyping and Evaluation in Implicit Race Bias: Evidence for Independent Constructs and Unique Effects on Behavior.” Journal of Personality and Social Psychology 91(4):652–61.

Arkes, Hal R., and Philip E. Tetlock. 2004. “Attributions of Implicit Prejudice, or ‘Would Jesse Jackson ‘Fail’ the Implicit Association Test?’” Psychological Inquiry 15(4):257–78.

Bursell, Moa, and Filip Olsson. 2020. “Do We Need Dual-Process Theory to Understand Implicit Bias? A Study of the Nature of Implicit Bias against Muslims.” Working paper, Stockholm University.

Cuddy, Amy J. C., Susan T. Fiske, and Peter Glick. 2008. “Warmth and Competence as Universal Dimensions of Social Perception: The Stereotype Content Model and the BIAS Map.” Advances in Experimental Social Psychology 40:61–149.

Fazio, Russell, and Michael Olson. 2003. “Implicit Measures in Social Cognition Research: Their Meaning and Use.” Annual Review of Psychology 54:297–327.

Fiske, Susan T. 2015. “Intergroup Biases: A Focus on Stereotype Content.” Current Opinion in Behavioral Sciences 3:45–50.

Greenwald, Anthony G., and Mahzarin R. Banaji. 1995. “Implicit Social Cognition: Attitudes, Self-Esteem, and Stereotypes.” Psychological Review 102(1):4–27.

Greenwald, Anthony G., Brian A. Nosek, and Mahzarin R. Banaji. 2003. “Understanding and Using the Implicit Association Test: I. An Improved Scoring Algorithm.” Journal of Personality and Social Psychology 85(2):197–216.

Hughes, Sean, Dermot Barnes-Holmes, and Jan De Houwer. 2011. “The Dominance of Associative Theorizing in Implicit Attitude Research: Propositional and Behavioral Alternatives.” The Psychological Record 61:465–96.

Kurdi, Benedek, Thomas Mann, Tessa Charlesworth, and Mahzarin R. Banaji. 2019. “The Relationship between Implicit Intergroup Attitudes and Beliefs.” PNAS 116(13):5862–71.

Lamont, Michèlle, Graziella Moraes Silva, Jessica Wellburn, Joshua Guetskow, Nissim Mizrahi, Hannah Herz hog, and Elisa Reis. 2016. Getting Respect: Responding to Stigma in the United States, Brazil and Israel. Princeton, NJ: Princeton University Press.

Melamed, David, Christopher W. Munn, Leanne Barry, Bradley Montgomery, and Oneya F. Okuwobi. 2019. “Status Characteristics, Implicit Bias, and the Production of Racial Inequality.” American Sociological Review 84(6):1013–36.

Payne, Keith, Heidi Vuletich, and Kristjen Lundberg. 2017. “The Bias of Crowds: How Implicit Bias Bridges Personal and Systemic Prejudice.” Psychological Inquiry 28(4):233–48.

Rudman, Laurie A., and Richard Ashmore. 2007. “Discrimination and the Implicit Association Test.” Group Processes and Intergroup Relations 10(3):359–72.

Rudman, Laurie A., Anthony G. Greenwald, and Debbie E. McGhee. 2001. “Implicit Self-Concept and Evaluative Implicit Gender Stereotypes: Self and Ingroup Share Desirable Traits.” Personality and Social Psychology Bulletin 27(9):1164–78.

Schnabel, Konrad, Jens Asendorpf, and Anthony G. Greenwald. 2008. “Understanding and Using the Implicit Association Test: V. Measuring Semantic Aspects of Trait Self-Concepts.” European Journal of Personality 22(8):695–706.

Moa Bursell holds a PhD in Sociology from Stockholm University. She is currently a research leader at the Institute for Futures Studies in Stockholm, Sweden. Her research focuses on ethnicity, implicit ethnic bias, ethnic boundaries, intersectionality, discrimination, and experimental methods.

Filip Olsson is a PhD student at the Department of Sociology, Stockholm University. His current research centers around implicit cognitions, discrimination, and nationalism.