Review of Philosophy and Psychology
https://doi.org/10.1007/s13164-022-00662-w

Testimonial Injustice: The Facts of the Matter

Migdalia Arcila-Valenzuela[1] · **Andrés Páez**[2]

Accepted: 11 October 2022

Abstract
To verify the occurrence of a singular instance of testimonial injustice three facts must be established. The first is whether the hearer in fact has an identity prejudice of which she may or may not be aware; the second is whether that prejudice was in fact the cause of the unjustified credibility deficit; and the third is whether there was in fact a credibility deficit in the testimonial exchange. These three elements constitute the facts of the matter of testimonial injustice. In this essay we argue that none of these facts can be established with any degree of confidence, and therefore that testimonial injustice is an undetectable phenomenon in singular instances. Our intention is not to undermine the idea of testimonial injustice, but rather to set limits to what can be justifiably asserted about it. According to our argument, although there are insufficient reasons to identify individual acts of testimonial injustice, it is possible to recognize recurrent patterns of epistemic responses to speakers who belong to specific social groups. *General* testimonial injustice can thus be characterized as a behavioral tendency of a prejudiced hearer.

1 Introduction

According to the standard definition, a speaker sustains a testimonial injustice if and only if she receives a credibility deficit owing to an identity prejudice in the hearer (Fricker 2007). This definition focuses on the epistemic effect of the hearer's

✉ Andrés Páez
apaez@uniandes.edu.co

Migdalia Arcila-Valenzuela
ma776@cornell.edu

[1] Sage School of Philosophy, Cornell University, 218 Goldwin Smith Hall, Ithaca, NY 14853, USA

[2] Department of Philosophy, Universidad de los Andes, Carrera 1 No. 18A-12 (G-533), Bogotá, DC 111711, Colombia

Published online: 02 November 2022

prejudice, but it does not distinguish between cases where the speaker's credibility is intentionally or unintentionally diminished. Each case requires a different type of analysis. Intentionally misrepresenting someone's beliefs as false or rationally unfounded does not lead to a misjudgment of the speaker's epistemic status; instead, the hearer intends to *manipulate* others to doubt the speaker's words (Fricker 2017), or to *undermine* the speaker's self-confidence by seeding doubt in her own beliefs, a phenomenon commonly known as "gaslighting" (Abramson, 2014). In contrast, the absence of deliberate, conscious manipulation leads to an unintended form of injustice that happens by way of a discriminatory but ingenuous *misjudgment* of the speaker's epistemic standing. This second form of testimonial injustice gives rise to an important epistemological problem that does not arise in the first case. If the misjudgment that generates this form of testimonial injustice is neither conscious nor intended, how can the hearer ever know that she has committed a testimonial injustice? Is there any evidence available to an external observer to establish that a testimonial injustice has occurred? Or is unintended testimonial injustice an opaque, undetectable phenomenon?

To verify the occurrence of an instance of testimonial injustice three facts must be established. The first is whether the hearer in fact has an identity prejudice of which she may or may not be aware; the second is whether that prejudice was in fact the cause of the credibility deficit; and the third is whether there was in fact a credibility deficit in the testimonial exchange. These three elements constitute the facts of the matter of testimonial injustice. In the case of an avowed racist or sexist, the first fact is easy to establish. But these cases are not the most interesting. Fricker, for example, rightly limits her most recent analysis of testimonial injustice to those cases that are "easy to miss" (2017, p. 54) because they do not arise from situations involving declared racist or sexist individuals. We will also limit the analysis to cases in which the hearer does not consciously accept that he or she has an identity prejudice.[1] The second fact, regarding the causal role of a person's prejudice, depends on how one parses individual attitudes and contextual influences in the determination of judgment and behavior. If implicit prejudice is construed along situationist lines, for example, the determination of causal influence in individual cases becomes a very difficult task. Finally, regarding the third fact, evidence of widespread cases of prejudiced credibility deficits is relatively easy to find. There are several statistical studies about racial and gender bias in hiring practices that can be best understood as cases of testimonial injustice (Bertrand and Mullainathan 2003; Norton et al. 2004; Quillian et al. 2017). More generally, the history of racism and sexism in many societies around the world is sufficient evidence to conclude that testimonial injustice is a common phenomenon. However, the question is not about the undeniable existence of testimonial injustice in general, but about the means of establishing the occurrence of an unjustified credibility deficit in *singular* cases. Since the hearer has made an ingenuous misjudgment, it must be left to others, or to the hearer at a different time,

[1] Although the first question is easy to answer in the case of avowed racists or sexists, this does not mean that it is obvious how a speaker can receive an *unintended* credibility deficit from such individuals. We leave the exploration of this question for future work.

to determine that the credibility owed to the speaker did not match the available evidence. How is this to be determined?

Our purpose in this paper is to show that in practice none of these three facts can be established, and therefore that testimonial injustice is an undetectable phenomenon in singular cases. In doing so, however, our intention is not to undermine the idea of testimonial injustice, but rather to set limits to what can be justifiably asserted about it. According to our argument, although there are insufficient reasons to identify individual acts of testimonial injustice, it is possible to recognize in an individual *recurrent patterns* of epistemic responses to speakers who belong to specific social groups. *General* testimonial injustice can thus be characterized as a behavioral tendency of a prejudiced hearer.[2] Being able to detect a negative behavioral tendency is sufficient to activate preventive and corrective strategies that do not depend on identifying singular cases of testimonial injustice.

The paper is structured in the following way. In the next section we show that claims about the existence of implicit identity prejudices as stable personal traits are based on evidence that has been recently discredited, and that no clear alternative has emerged for the empirical study of implicit prejudice. This section also discusses the uncertain causal role of prejudice, given that contextual elements and cognitive biases are also known to play a role in our perception of people's credibility. We show that establishing the contribution of each element is a task fraught with perils. In the third section we argue that recent interpretations of implicit measures offer empirical support to the idea of general testimonial injustice as a behavioral tendency not necessarily associated with a stable mental construct. In the fourth section we examine the assumption that the evidence available in a testimonial exchange determines the credibility owed to a speaker and that it is possible to establish whether a credibility deficit has occurred. In the fifth and final section we discuss a possible objection to our analysis, namely, that it scientizes testimonial injustice and sets an excessive burden of proof upon its victims. Finally, it should be noted that in this paper we only attempt the negative task of describing the epistemic opacity of singular testimonial injustice. The positive task of characterizing a general concept of testimonial injustice with adequate empirical support is left for future work.

2 Detecting Prejudice and Its Effects

The origin of testimonial injustice is the *existence* in the hearer of an identity prejudice that acts as the *cause* of her misjudgment of the speaker's credibility. In this section we will examine whether it is possible to establish either of these facts. As stated in the Introduction, we are only interested in cases in which the hearer is not aware that she has an identity prejudice. For that reason, the only way to overcome the asymmetry between her implicit and explicit attitudes will be to adopt an indirect method that does not rely on introspection-based self-reports.

[2] The contrast established here between singular and general testimonial injustice resembles to some extent the epistemological issues involved in the distinction between singular and general causation (Davidson 1980; Hitchcock 1995; Danks 2017).

Implicit identity prejudices are often studied under the rubric of "implicit bias" (Wittenbrink et al. 1997; Brownstein 2018). Although implicit biases include both social and cognitive biases, it has become increasingly common to restrict its use to refer only to the former. We will therefore use the expressions "implicit prejudice" and "implicit bias" interchangeably. Now, the term "implicit bias" itself is used in the literature on implicit social cognition in a rather broad sense. Philosophers who are mostly interested in the social and political effects of implicit bias tend to define it functionally to refer to any mental content or process that affects or influences our actions, perceptions and decisions in undesirable or discriminatory ways (e.g., Saul 2013). This functional definition is unsatisfactory for our purposes because we need to understand the structure of the inner process itself if we want to find ways of identifying the mental origin of testimonial injustice.

In the psychological literature, implicit bias has been traditionally defined as a process in which stable associations stored in memory are unconsciously activated (Amodio and Mendoza 2010; Payne and Gawronski 2010; Nosek et al. 2012).[3] Implicit identity prejudices should then be understood as stable associations between, for example, gendered words or images and positive or negative attributes (Webb et al. 2010). Likewise, valenced attributes are often associated with words and images that reflect racial, religious, ethnic, and other identities. In recent years such associations have been detected using measures such as the implicit Association Test (IAT) (Greenwald et al. 1998), the Evaluative Priming Test (EPT) (Fazio et al. 1995), and the Affect Misattribution Procedure (AMP) (Payne et al. 2005). These analysis techniques are easy to interpret because they yield a single "bias" score for each participant, which seems to imply that there is a singular mental construct—an implicit bias—that is being measured. As we will see in what follows, this is a questionable assumption.

Because implicit measures only assess behavior, an implicit bias is best understood as a hypothesized construct that explains people's performance in these tests. The theoretical question about implicit bias then becomes one about the kind of underlying psychological structures and processes that best explain and predict the behavioral evidence. Brownstein et al. (2019) identify two debates related to implicit measures that are especially germane in the present context: (i) whether performance on these tests reflects temporally stable traits or occasion-specific states; and (ii) whether performance reflects characteristics of the person or of the situation in which he or she is taking the test.

The identity prejudice in the standard definition of testimonial injustice corresponds to a temporally stable personal trait, not to a spontaneous affective reaction

[3] To be sure, there are several other approaches to implicit bias in psychology (Byrd 2021) and there is a great deal of theoretical controversy. Some of these views flatly reject that implicit bias is underwritten by associations (Mandelbaum 2016), while other interactionist views allow that implicit bias can be predicated on associative and non-associative processes (Gawronski and Bodenhausen 2011). Distinctions can also be drawn depending on whether these processes are more or less reflective. It falls beyond the scope of this paper to explore these alternative views of implicit bias. We will restrict our analysis to the received view and the implicit measures closely associated with it. In particular, the issues related to the predictive validity and reliability of implicit measures are fairly independent from the theoretical viewpoint one adopts.

triggered by structural or accidental features of the situation.[4] There are several reasons that support this interpretation. First of all, negative identity prejudices are "epistemically culpable" because of their "resistance to counter-evidence owing to an ethically bad affective investment" (Fricker 2007, p. 35). Identity prejudices are thus entrenched epistemically and emotionally and preserve their identity through time. Furthermore, according to Fricker, in testimonial injustice the hearer exercises an "*agential* identity power" (p. 90, emphasis added) over the speaker. In contrast, the *structural* operation of identity power "is appropriate if one wishes to highlight the fact that all parties are to some extent under the control of a gender or racial ideology. But since my aim is to highlight the injustice that is occurring, and the sense in which the hearers are preventing the speakers from conveying knowledge, it is the agential description that is most relevant here" (pp. 90–91). Finally, if social prejudice were not a stable personal trait, it would be difficult to see why Fricker's primary solution to prevent testimonial injustice is based on personal virtues; otherwise, it would be more natural to advocate a more structural solution that eliminates contextual elements that favor the formation of negative affective reactions. To be sure, she does not discount the importance of structural solutions, but social change has to begin with individuals who display virtuous behavior.

The importance of the existence of stable personal identity prejudice comes out in one of Fricker's central examples in *Epistemic Injustice.* The case is designed precisely to illustrate why circumstantial epistemic bad luck in the absence of prejudice does not give rise to testimonial injustice: Suppose speaker S is a sincere but extraordinarily shy person, and S's shifty manner during an interview is judged by the hearer H to be a reliable sign that S is untrustworthy. S receives a credibility deficit because H is using an empirically reliable rule about credibility (p. 41).[5] H is non-culpable epistemically and ethically and therefore has not committed an epistemic injustice.

Interestingly, this is the only example in the book for which no social context is provided. Let's provide one. Suppose H is a white man, S is a shy African American woman and, as before, H judges S to be untrustworthy. Can we remain confident that H is still non-culpable and that this is still not a case of epistemic injustice? H can claim, as before, that he is using an empirically reliable rule, but there is a chance that the rule is being used more harshly because of the influence of racial or sexual prejudice, or even of an implicit prejudice against shy people. If so, it would be an instance of epistemic injustice. Thus, the issue hangs on whether H has an identity prejudice that is worsening the credibility deficit. The only way currently available to find out if he does is via an implicit measure, which brings us back to the fundamental distinctions regarding implicit bias: traits vs. states, and personal characteristics vs. reflections of the situation.

Regarding the first pair, the difference between a trait and a state has to do with how stable a given construct is over time and across situations. Attitudes, tastes, pref-

[4] Anyone can have a spontaneous negative affective reaction towards, for example, an African American man when the person is presented as the villain in a horror movie. The affective reaction can be exactly the opposite if the subject watches a documentary about Nelson Mandela (Brownstein et al. 2019). In both cases the context explains the valence of the affective transient state independently of the person's more stable racial associations.

[5] In Sect. 4 we will examine whether this is really an empirically reliable rule.

erences of various kinds can be more trait-like or more state-like depending on their stability in different contexts (Fazio 2007; Schwarz 2007). If implicit measures capture stable traits, their results should not fluctuate considerably over time, i.e., they should have a high test-retest reliability. However, multiple recent longitudinal studies have shown low correlations between a person's score on implicit measures across days, weeks and months (Cooley and Payne 2017; Gawronski, Morrison, et al.2017a, b). Also, stability varies in different content domains. Implicit measures of socially salient categories such as race and gender, which are central in testimonial injustice, are significantly less stable than implicit measures of political attitudes (Gawronski, Morrison, et al. 2017a, b). Rae and Olson (2018) report similar low reliability results for the race and gender IAT given to children and teens. Implicit measures in general are also less stable than explicit measures of the same attitudes (Gawronski, Brannon, et al. 2017). Implicit bias thus fails the reliability test for stable traits.

If implicit measures capture stable personal traits, they should also have high predictive validity given the causal role attributed to them in social behavior. As noted by Rae and Olson, "predicting behavior is a key motivation behind the use of implicit measures" (2018, p. 309, quoted by Machery 2021). Several studies have reported correlations between implicit measures and behavior. The evidence is used to argue for the causal importance of automatically retrieved associations (e.g., Greenwald et al. 2009; Devine et al. 2012). However, the correlations between implicit measures and behavior tend to be smallest for topics in which automatic and deliberate processes are least likely to be aligned, such as race relations (Greenwald et al. 2009; Schimmack 2021). More recent evidence shows that changes in implicit measures do not, in general, result in changes in behavior. The influential meta-analysis of Oswald et al. (2013) examined the predictive validity of the race and ethnicity IATs for a wide range of criterion measures of discrimination. They found that IATs were poor predictors of every criterion category, and that the IATs performed no better than simple explicit measures. More recently, Forscher et al. (2019) presented a meta-analysis of 492 studies (87,418 participants) to investigate the effectiveness of procedures to change implicit measures, and whether implicit measure change translates into change in actual or intended behavior. They concluded:

> To get closer to questions of causality, we looked at whether changes in implicit measures correspond with and mediate changes in behavior in our sample of randomized experiments. We found that the effect of procedures on behavior were trivial by conventional standards, with the exception of threat which had a small-to-moderate effect on behavior. We found no evidence that changes in implicit measures mediate changes in behavior (p. 543).

Finally, the existence of a causal relationship between the construct detected by implicit measures and behavior should allow the design of efficient implicit bias training programs that focus on manipulating the construct. Despite being used by

many police departments around the world, such programs have proved to be largely inefficient (Carter et al. 2020).[6]

The implication of the low predictive reliability of implicit measures for testimonial injustice is that even if we were to trust an implicit measure that indicates that the hearer H has an implicit prejudice,[7] there is no evidence of a causal connection between the implicit measure and any of H's judgments and behaviors. In the example of the shy interviewee, it is empirically impossible to determine whether, on this occasion, a biased hearer used an empirically reliable rule about credibility or committed an epistemic injustice. One might be tempted to say that it is not important to find the exact cause as long as we detect the epistemic wrong. But we should resist this temptation because H will remain epistemically and ethically non-culpable until proven guilty of a prejudiced misjudgment.

3 Traits as General Behavioral Tendencies

The difficulties detected in the previous section indicate that the prospects of detecting individual cases of testimonial injustice are slim. The current discussion of the weaknesses of implicit measures seems to be converging towards the view that they are not good measures of people's individual biases (Greenwald et al. 2015),[8] and new explanatory alternatives of the outcomes have emerged. In this section we will present some recent developments that lend credence to the idea that testimonial injustice can only be understood as a general behavioral tendency.

Mitchell and Tetlock (2006) present various studies that show that implicit measures such as the IAT measure a host of alternative processes that do not involve implicit *negative* bias toward social groups. In the race IAT, for example, there is an apparent compatibility effect between the "pleasant" attribute and the "white" category. Instead of interpreting this result as a reflection of a permanent affective valence in the subject, an alternative interpretation is that greater familiarity with the white category makes it more salient (Kinoshita and Peek-O'Leary 2005). Uhlmann et al. (2006) argue that White Americans' negative automatic associations with African Americans may partly result from associating members of low status groups with unfair circumstances and not with negative attributes. In a similar vein, Andreychik and Gill (2012) argue that measures of implicit evaluation fail to detect the difference

[6] A recent strategy to improve the validity and reliability of the IAT is to develop models that separate processes related to cognitive control, stimulus encoding, associations between concepts and categories, and processes unrelated to the choice itself. Although it is a promising route, it is still in its infancy. Some of the best-known models, the Quadruple Process (Conrey et al. 2005) and the ReAL model (Meissner and Rothermund 2013), require simplifications and modifications of the original IAT, thus impeding the possibility of re-analyzing older data. Kvam et al. (2022) offer a computational model that uses the original IAT but it has not been peer reviewed as of this writing.

[7] There is, of course, the additional problem of establishing a nonarbitrary score that indicates the existence of said prejudice (Blanton and Jaccard 2006; Mitchell and Tetlock 2017).

[8] Besides their low test-retest reliability and low predictive validity, implicit measures have other weaknesses: scores vary according to situational factors, but the influence of the latter is transient; correlations between indirect measures as well as between indirect and direct measures vary substantially; and indirect measures do not correlate with one another (Machery 2016, 2021).

between empathy-based and prejudice-based associations. Thus, implicit measures of prejudice can tap negative, yet egalitarian associations. Finally, Arkes and Tetlock (2004) argue that the reaction times in the IAT may reflect shared cultural stereotypes rather than personal animus. This is just a small sample of experimental results that explain the results of implicit measures without appealing to the existence of a stable negative identity prejudice in the subject.

Payne et al. (2017) take a different approach in response to the crisis of implicit measures. In their view, implicit bias can still be attributed to a subject on the basis of implicit measures if we adopt "a situationist view of implicit bias." This brings us back to the second debate identified by Brownstein et al. (2019): implicit measures as reflections of personal characteristics vs. reflections of the situation in which the person is taking the test. Payne et al. propose that indirect measures do not gauge a stable construct but rather the transient, situational variations in the strength of the connections in the conceptual network that represents social categories in an individual. These variations explain both the low test-retest reliability and the low predictive validity of implicit measures. They also explain why young children show levels of implicit bias similar to adults (Dunham et al. 2008). The authors call their model "the bias of crowds." Their thesis is based on the fact that implicit measures can be easily affected by features of situations. For example, implicit racial bias scores have been shown to be affected by the interaction with a Black experimenter, listening to rap music, or looking at photos of Black celebrities (see Lai et al. 2013, for a review).

There are, of course, researchers who still defend implicit bias as an individual construct. Machery (2017) argues that the solution to the low test-retest reliability of implicit measures is not to average at the group level but to do so at the individual level:

> One would obtain a stable individual measurement of this individual's bias by aggregating across her time slices. (…) And exactly as group-level measurement is predictive of group-level discriminatory behavior, an aggregate individual-level measure of bias would be predictive, not of individual discriminatory behavior but of aggregate discriminatory behavior of a single individual (p. 289).

In a sense, an implicit bias would be a trait, "a disposition to perceive, attend, cognize, and behave in a particular way in a range of social and nonsocial circumstances" (p. 289).

Recent evidence calls into question the feasibility of Machery's proposal. Hannay and Payne (2022) show that aggregating multiple tests per person "might provide researchers with slightly greater validity due to reduced person-level error variance [i.e., noise]. However, the absolute size of the test-retest correlations and validity correlations remained small by conventional standards." (p. 5). In fact the authors explicitly interpret these results as a refutation of Machery's idea that increasing the number of person-level measurements will reveal large correlations. It is also intended as a refutation of the idea that the IAT's low stability reflects a large amount of random measurement error. The results show that aggregating multiple measurement does not eliminate any additional noise after two or three tests. Hannay and

Payne also point out that their paper, together with Forscher et al. (2017), are the only studies that have collected more than three repeated implicit measurements, so there is no further evidence based on aggregation that can presently be considered.

Notice that even if Machery is right and implicit measures are predictive of a person's aggregate discriminatory behavior, this would not solve the epistemological problem of *singular* testimonial injustice because a person's aggregate individual-level measure of bias would not be proof of discriminatory behavior in individual instances, which is what is required to establish that a case of testimonial injustice has occurred. Intergroup biases, in particular, still have low predictive validity in Machery's approach. At best, an aggregate implicit measure would allow us to say that a person is on average more racist and therefore more prone to commit testimonial injustice. Machery's approach could thus be the starting point for an account of *general* testimonial injustice.

Our purpose in the previous two sections has not been to settle which of the alternative explanations of the results of implicit measures is correct, but rather to point to a lack of consensus about the very *existence* of the stable personal traits that play an essential role in the definition of testimonial injustice,[9] and a lack of evidence about their *causal* role. Without such evidence, singular testimonial injustice remains epistemically opaque.

4 The Assessment of a Speaker's Credibility

The existence of a credibility deficit is the third fact that must be established to assert the occurrence of a singular instance of testimonial injustice.[10] How should we understand the idea of a deficit in this context? A credibility deficit implies that there is a minimum degree of credibility that the speaker should have been given by the hearer in light of the available evidence. Let us call this the *minimum credibility thesis*. Alternatively, if we do not want to commit ourselves to the idea of degrees of credibility, a credibility deficit implies that there is at most one propositional attitude (belief, disbelief, or suspension of judgment) that the hearer ought to have adopted towards the speaker's words. This is a version of the well-known "uniqueness thesis" (Feldman 2007, p. 205). In this section we examine the plausibility of these theses and whether it is possible at all for a hearer to purposefully fulfill her purported epistemic duty to the speaker from either of these two perspectives.

Let us begin with the minimum credibility thesis.[11] It is doubtful that there is a minimum (or an exact) degree of credibility owed to a speaker in a given context. As a theoretical construct in the mind of the hearer, the credibility attributed to a speaker

[9] In a recent survey of implicit measures, Machery states that "a basic issue in implicit attitude research—what do indirect measures measure?—is still unanswered" (2021, p. 6).

[10] Lackey (2020) argues that under certain circumstances an unwarranted credibility *excess* can be understood as a case of testimonial injustice. In this paper we will only discuss credibility deficits, but the main argument applies to both cases.

[11] The minimum credibility thesis can be strengthened and transformed into a version of the uniqueness thesis if we interpret "attitude" as credence (Cohen 2013, p. 101). The arguments presented here against the minimum credibility thesis apply *pari passu* to the Bayesian version of the uniqueness thesis.

is always underdetermined by the available evidence. It is possible for noncognitive values to enter into evidential reasoning, and they often do, especially in everyday contexts that are not ruled by strict methodological principles. Naturally, if the noncognitive values in question are prejudices or social biases, we have a case of testimonial injustice. But even in their absence, there are other noncognitive values that fill the logical gap between evidence and hypothesis. Among them are economic values such as risk aversion, socially determined preferences, and culturally and institutionally filtered evaluations.[12] To put it another way, any testimonial exchange is personally, socially and culturally situated, and there is no neutral context in which a minimum degree of credibility owed to the speaker can be established. Without such a normative minimum measure, it becomes impossible to say that there was in fact a credibility deficit in a singular testimonial exchange.

It could be argued that if personal, social, and cultural values are factored into the context of a testimonial exchange, the idea of a normative minimum of credibility can be restored. For any socially and culturally situated hearer, the required minimum would be the credibility appraisal of the speaker that he or she would reach in the absence of implicit prejudice. But this strategy would only dissolve the idea of an epistemic standard for credibility. If any credibility assessment of a speaker is as good as any other—as long as there is no prejudice involved—then there is no general epistemic norm that is being satisfied by any particular individual. Without a normative standard, all talk of a credibility deficit is rendered meaningless. Furthermore, it could lead to absurdity. For example, the degree of credibility attributed to a speaker by a prejudiced individual might end up being higher than that of an unprejudiced but very skeptical individual with an extremely high epistemic risk aversion. In brief, if the standard is understood in terms of degrees of credibility, there is no way to set up a minimum standard that gives content to the idea of credibility deficit. And if the standard is defined in negative terms, as the absence of prejudice, it loses any normative force.

The uniqueness thesis, which is framed in terms of rough-grained propositional attitudes, seems more plausible than the minimum credibility thesis. An initial drawback of the uniqueness thesis is that it impoverishes the concept of testimonial injustice. Although many examples used in the literature on testimonial injustice focus on cases in which the speaker is disbelieved as the result of the hearer's prejudices, not all cases involve a change in propositional attitude. Consider the case of an employer who decides to hire a highly qualified female, but due to his sexist implicit prejudice gives her less responsibilities than he would have given a male employee. Or an investor who is advised by her very competent African American stockbroker to buy $1000 in shares of company X but ends up buying only $600 because of her implicit racial prejudice. In both cases the hearers trust the speakers to a lesser extent than they should have, but they trust them nonetheless. Perhaps the intuitive appeal of testimonial injustice comes from cases in which there is a change of attitude towards the speaker, but the idea of a credibility deficit should also include cases like these.

[12] There is an ongoing lively debate about how to tell beneficial from noxious noncognitive values in theory choice in science (Hicks 2014; Psillos 2015; Goldenberg 2015). It seems unlikely that such a debate can be fruitful outside of the regimented context of science.

Ignoring them would run counter Fricker's stated purpose of bringing to light cases of testimonial injustice that are "easy to miss."

Suppose we settle for this restricted sense of credibility deficit, thereby limiting the scope of testimonial injustice. Isn't this thesis vulnerable to the same objection based on noncognitive values discussed above? So-called "permissivists" have used arguments along these lines to attack the uniqueness thesis.[13] Defenders of the uniqueness thesis have replied that the influence of noncognitive values is less definitive here. They might switch a hearer's propositional attitude in boundary cases, but not in general. Even in the absence of methodological rules in everyday life, the argument goes, there are implicit and explicit prudential principles and sufficient inductive evidence that people follow when assessing a speaker's credibility. Lackey, for example, argues that hearers in a testimonial exchange will have "a substantial amount of inductive evidence for believing that … reports made with sustained eye contact are typically sincere ones, or that reports made ably and confidently are typically confident ones" (2006, p. 173). The question is whether this inductive evidence is sufficiently strong to support accurate individual credibility assessments in all circumstances. The psychological literature on trust and deception seems to indicate that it is not.

People deploy two monitoring strategies to evaluate the credibility of speakers. The first is to detect positive evidence that one's interlocutor is trustworthy or competent; the second, to identify traits that reveal that the speaker is deceptive.[14] According to Shieber, these strategies face two problems: "there may well be no uniform, stable set—or sets—of traits signaling trustworthiness or deceptiveness [and] even if there are traits signaling trustworthiness or deceptiveness, subjects aren't reliably sensitive to those traits" (2012, p. 6).

The first problem was diagnosed long ago in the social psychological literature. According to interpersonal deception theory (Buller and Burgoon 1996), speakers have different goals, motivations, emotions, strategies, and cognitive abilities, and interact with hearers with whom they have different degrees of familiarity on matters of different importance in contexts that vary widely. The complexity of interpersonal communication makes it very unlikely that there will be one profile of honest or deceptive behavior. For example, the criteria mentioned by Lackey have proved to be completely useless: liars in fact maintain more sustained levels of eye contact than truth tellers (Sitton and Griffin 1981) and produce no more nervous smiles than sincere interlocutors (Hartwig and Bond 2011). Microexpressions, which were once heralded as a useful technique for catching liars (Ekman 2001), have been largely discredited (DePaulo et al. 2003; Mercier 2020). Furthermore, traits that people do not consciously associate with trustworthiness, such as a speakers' physical attractiveness (Chaiken 1979), the fact that they are wearing uniforms (Bickman 1974) or using jargon (Cooper et al. 1996) end up having a large positive effect on judgments of credibility. In sum, the traits we tend to believe are reliable, are not; and the traits that we do not tend to count as reasons for our credibility judgments affect us sub-

[13] See Jackson and Turnbull (2023) for an overview of the literature.

[14] There is empirical evidence that these strategies employ two distinct cognitive mechanisms (Ekman et al. 1999, p. 265).

consciously. The absence of a stable inductive basis for the assessment of a speaker's credibility makes all such singular judgments unfounded. Even if people generally agree on their credibility judgments and on the reasons they offer for those judgments, the non-existence of consistent credibility-signaling traits removes all force from an epistemic standard based on those criteria.

The second problem described by Shieber is equally detrimental to the idea of a credibility standard for individual testimonial exchanges. In many experiments with people who are instructed to lie or to be truthful, observers have failed systematically to detect deception. In a well-known study, Ekman and O'Sullivan (1991) used a videotape that showed 10 people who were either lying or telling the truth in describing their feelings. The authors evaluated 509 people including law enforcement personnel, such as members of the US Secret Service, Central Intelligence Agency, Federal Bureau of Investigation, National Security Agency, Drug Enforcement Agency, California police and judges, as well as psychiatrists, college students, and working adults. Only members of the Secret Service performed better than chance.

Although demeanor cues are completely unreliable, there are other behavioral and contextual cues that are helpful in credibility judgments. Mercier (2020) suggests that hearers can search for signs of a speaker's diligence to provide valuable information and try to determine whether the hearer and the speaker's incentives are alligned. "We can trust speakers to be diligent when their incentives align with ours" (p. 92). But since diligence and incentives are often difficult to detect, and since people are often negligent and their incentives do not align with ours, humans have developed an effective method to control for truthfulness: reputation. "Being a diligent communicator is a crucial trait of a good cooperation partner. Receivers should be able to keep track of who is diligent and who isn't, and adjust their future behavior on that basis" (pp. 88–89). Plausible as this sounds from an evolutionary perspective, it is not very helpful in the dialogical conditions in which testimonial injustice often occurs. It generally involves complete strangers who interact for the first time, with limited information of the speaker's incentives or her track record for truthfulness or diligence. Furthermore, Mercier's approach reveals that gauging credibility takes time and more than a few testimonial exchanges. People who sustain a testimonial injustice in a job interview or in a court hearing are not afforded the time or the opportunity to reveal much about their own competence and honesty. To be sure, a speaker can be subject to testimonial injustice for an extended period of time covering many testimonial interactions with a prejudiced hearer, but it is quite likely that the credibility deficit occurred from the very beginning.

These results indicate that there is no "correct" way of attributing credibility to others in individual testimonial exchanges because there is no general inductive basis to do so, and therefore no standard against which to establish a credibility deficit. Furthermore, even if an inductive basis were to be established, people are incapable of detecting the tell-tale signs of liars.[15] And yet, we have accepted that there is evidence of widespread cases of testimonial injustice. How can those two positions be reconciled?

[15] Due to space restrictions, we refrain from discussing the implications of these findings for the dispute between reductionists and anti-reductionists in the philosophy of testimony.

Our position is that it is possible to obtain statistical measures of credibility *inequality* in a population. These measures provide the evidence for widespread testimonial injustice. Moss-Rascusin et al. (2012), Honeycutt et al. (2020), and older studies like Bertrand and Mullainathan (2003) have detected credibility inequalities in different populations. The first two studies focused on STEM faculty and gender bias, while the third one detected significant racial discrimination among potential employers in Chicago and Boston.[16] It should also be possible to design a within-subject experiment comparing several instances of a person's credibility judgments. The experiment could detect an individual pattern of credibility inequality, but it would not help us set a standard for credibility. Suppose the task is to rate equivalent CVs from men and women, and the latter consistently receive a lower score. Is the score given to men's CVs the standard against which a credibility deficit should be established? Not at all. It might well be that men receive excessive credibility from the experimental subject. The only thing we can establish for sure is that the credibility judgments are unequal and biased, but no credibility benchmark can be inferred from the data, which is the point we wanted to establish in this section.

5 Act-Based vs. Victim-Centered Approaches to Testimonial Injustice

In the previous sections we have used evidence from social and cognitive psychology to call into question our epistemic access to individual instances of testimonial injustice. In this section we want to briefly defend this naturalistic methodological approach and respond to a possible objection. We have been accused, in discussion, of "scientizing" testimonial injustice, of requiring scientific evidence for a phenomenon that is quotidian and easily detectable by its victims, perhaps by means of a simple inference to the best explanation. In brief, to some, our approach ignores the victim's perspective and places an impossible probative burden on her.

To respond to this worry we think it is relevant to differentiate between a victim-centered and an act-based approach to testimonial injustice. This distinction has been recently used in the case of microaggressions (Freeman and Stewart 2021). An act-based approach privileges questions about intentionality, causation, and responsibility for *inflicted* harms, while a victim-centered approach focuses on the consequences of *experienced* harms. According to Freeman and Stewart (2021), the main limitation of the former approach is that it displaces the victim from the central theoretical and practical role she should have. There is an epistemic dimension of testimonial injustice that can only be grasped by its victims, as is widely recognized by standpoint theory (e.g., Collins 2002; Wylie 2003; Freeman and Stewart 2020). Moreover, the

[16] In one of the very few studies focused specifically on testimonial injustice, Díaz and Almagro (2021) found no evidence that women are given less credibility then men, at least among the participants in their study. But even if the experimental evidence for widespread testimonial injustice turns out to be mixed, this does not imply that there are no individual patterns of credibility inequality. The authors allow for this possibility: "…should we conclude that testimonial injustice is not real? The answer to this question is a clear no. The fact that, on average, the participants in our studies did not attribute less credibility to women than men does not mean that there are no cases in which women are given less credibility than their male peers" (p. 19).

experience of the oppressed is crucial to remediate the harms caused by testimonial injustice. Only the receiving party can give voice to the complexity of the epistemic, emotional, and material harms imposed upon her.

Imposing a burden of proof on the victim also seems unjustified when there is strong evidence of widespread testimonial injustice in a restricted domain. Within the healthcare system, for example, there is evidence that women's pain is disregarded and misdiagnosed much more often than men's pain (Kent et al. 2012; Samulowitz et al. 2018).[17] Discounting women's pain reports can have devastating consequences for them, such as overlooking lethal coronary conditions (Chiaramonte et al., 2006). In this context, the available evidence should be sufficient to move healthcare providers to implement preventive measures without having to establish that whether the credibility inequality is caused by the prejudices of the individual medical personnel.

However, focusing *exclusively* on the victim's perspective is to miss the opportunity of finding measures that can prevent, or at least mitigate, the appearance of discriminatory behavior in prejudiced individuals. As we have argued in previous sections, implicit bias is a complex phenomenon in which individual, structural and situational factors play a role. An act-based approach to testimonial injustice focuses on empirically-validated methods that try to establish how these personal and contextual causes operate. Contrary to what Freeman and Stewart (2021) assert, advancing an act-based approach to testimonial injustice does not mean displacing or disregarding the victim's experience. It is precisely to avoid the exacerbation of the injustices suffered by marginalized groups that it is important to give equal attention to a victim-centered and an act-based approach.

Finally, we want to insist that we are not denying the existence of testimonial injustice, but advocating for a reconceptualization of the phenomenon. As we argue in the third section, even though we cannot identify single instances of testimonial injustice, and attribute responsibilities thereof, we do have statistical evidence that indicates the existence of widespread credibility inequality. We can also obtain evidence about an individual's pattern of unequal credibility judgments. In the second case, we can detect a behavioral tendency that we have characterized as *general* testimonial injustice.

6 Conclusion

Our purpose in this paper has been to point out the lack of evidential support for a singularist account of testimonial injustice. A general account seems to fare much better in evidential terms and does not depend on interpreting implicit measures in personalist terms. A generalist perspective will make it easier to devise effective strategies to counter the negative effects of testimonial injustice. It is undeniable that there are historically marginalized groups and that marginalization can take hidden and complex forms. Even if it is unlikely that we will ever be able to identify individual cases of testimonial injustice with any degree of confidence, it is our duty to detect contextual factors that increase the statistical risk that marginalized individuals will

[17] We are grateful to an anonymous reviewer for calling our attention to this specific context.

be degraded as knowers. When liberal and progressive institutions do not make an effort to eliminate these factors using solid scientific evidence, they are paying lip service to their professed goals.

Abandoning the use of testimonial injustice as a useful concept in individual instances has other theoretical consequences. The concept has been applied in contexts as varied as psychiatry (Kurs and Grinshpoon 2017), medicine (Carel and Kidd 2014), law (Fyfe 2018), and education (Kotzee 2013). We do not deny the usefulness of the concept in these areas, as long as it is used as a steppingstone towards changing the toxic circumstances in which many testimonial exchanges arise. For example, the credibility deficit suffered by minority students can be lessened using changes in admission policies that increase diversity in the student population. This change of focus from individual testimonial exchanges to favorable structural changes will contribute to lessen the effect of a phenomenon that is, in principle, undetectable.

Acknowledgements The authors would like to thank Manuela Fernández-Pinto, Ignacio Ávila, Santiago Amaya, Manuel Vargas, Nick Byrd, and an anonymous reviewer for this journal for helpful comments on an earlier draft.

Funding Open Access funding provided by Colombia Consortium

Conflict of interest None.

References

Amodio, D. M., and S. A. Mendoza. 2010. Implicit intergroup bias: Cognitive, affective, and motivational underpinnings. In *Handbook of implicit social cognition. Measurement, theory, and applications*, eds. B. Gawronski, and B. K. Payne, 353–374. New York: The Guilford Press.

Andreychik, M. R., and M. J. Gill. 2012. Do negative implicit associations indicate negative attitudes? Social explanations moderate whether ostensible "negative" associations are prejudice-based or empathy-based. *Journal of Experimental Social Psychology* 48: 1082–1093.

Arkes, H. R., and P. E. Tetlock. 2004. Attributions of implicit prejudice, or "Would Jesse Jackson 'fail' the Implicit Association Test?". *Psychological Inquiry* 15 (4): 257–278.

Bertrand, M., and S. Mullainathan. 2003. Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *American Economic Review* 94 (4): 991–1013.

Bickman, L. 1974. The social power of a uniform. *Journal of Applied Social Psychology* 4 (1): 47–61.

Blanton, H., and J. Jaccard. 2006. Arbitrary metrics in psychology. *American Psychologist* 61: 27–41.

Brownstein, M. 2018. *The implicit mind: Cognitive architecture, the self, and ethics*. New York: Oxford University Press.

Brownstein, M., A. Madva, and B. Garownski. 2019. What do implicit measures measure? *Wiley Interdisciplinary Reviews Cognitive Science* 10 (5): e1501.

Buller, D. B., and J. K. Burgoon. 1996. Interpersonal deception theory. *Communication Theory* 6 (3): 203–242.

Byrd, N. 2021. What we can (and can't) infer about implicit bias from debiasing experiments. *Synthese* 198 (2): 1427–1455.

Carel, H., and I. J. Kidd. 2014. Epistemic injustice in healthcare: A philosophical analysis. *Medicine Health Care and Philosophy* 17 (4): 529–540.

Carter, E. R., I. N. Onyeador, and N. A. Lewis Jr. 2020. Developing & delivering effective anti-bias training: Challenges & recommendations. *Behavioral Science & Policy* 6 (1): 57–70.

Chaiken, S. 1979. Communicator physical attractiveness and persuasion. *Journal of Personality and Social Psychology* 37 (8): 1387–1397.

Chiaramonte, G. R., and R. Friend. 2006. Medical students' and residents' gender bias in the diagnosis, treatment, and interpretation of coronary heart disease symptoms. *Health Psychology* 25 (3): 255–266.

Cohen, S. 2013. A defense of the (almost) equal weight view. In *The epistemology of disagreement: New essays*, eds. D. Christensen, and J. Lackey, 98–120. Oxford: Oxford University Press.

Collins, P. H. 2002. *Black feminist thought: Knowledge, consciousness, and the politics of empowerment*. Routledge.

Conrey, F. R., J. W. Sherman, B. Gawronski, K. Hugenberg, and C. J. Groom. 2005. Separating multiple processes in implicit social cognition: the quad model of implicit task performance. *Journal of Personality and Social Psychology* 89 (4): 469–487.

Cooley, E., and B. K. Payne. 2017. Using groups to measure intergroup prejudice. *Personality and Social Psychology Bulletin* 43: 46–59.

Cooper, J., E. Bennett, and H. Sukel. 1996. Complex scientific testimony: How do juries make decisions? *Law and Human Behavior* 20 (4): 379–394.

Danks, D. 2017. Singular causation. In *The Oxford handbook of causal reasoning*, ed. R. Waldmann, 201–215. New York: Oxford University Press.

Davidson, D. 1980. Causal relations. In *Essays on actions and events*, 149–162. Oxford: Clarendon Press.

DePaulo, B. M., J. J. Lindsay, B. E. Malone, L. Muhlenbruck, K. Charlton, and H. Cooper. 2003. Cues to deception. *Psychological Bulletin* 129 (1): 74–118.

Devine, P. G., P. S. Forscher, A. J. Austin, and W. T. Cox. 2012. Long-term reduction in implicit race bias: A prejudice habit-breaking intervention. *Journal of Experimental Social Psychology* 48: 1267–1278.

Díaz, R., and M. Almagro. 2021. You are just being emotional! Testimonial injustice and folk-psychological attributions. *Synthese* 198 (6): 5709–5730.

Dunham, Y., A. S. Baron, and M. R. Banaji. 2008. The development of implicit intergroup cognition. *Trends in Cognitive Sciences* 12 (7): 248–253.

Ekman, P. 2001. *Telling lies: Clues to deceit in the marketplace, politics, and marriage*. New York: Norton.

Ekman, P., and M. O'Sullivan. 1991. Who can catch a liar? *American Psychologist* 46 (9): 913.

Ekman, P., M. O'Sullivan, and M. G. Frank. 1999. A few can catch a liar. *Psychological Science* 10 (3): 263–266.

Fazio, R. H. 2007. Attitudes as object-evaluation associations of varying strength. *Social Cognition* 25: 603–637.

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

Feldman, R. 2007. Reasonable religious disagreements. In *Philosophers without gods*, ed. L. M. Antony, 194–214. Oxford: Oxford University Press.

Forscher, P. S., C. Mitamura, E. L. Dix, W. T. Cox, and P. G. Devine. 2017. Breaking the prejudice habit: Mechanisms, timecourse, and longevity. *Journal of Experimental Social Psychology* 72: 133–146.

Forscher, P. S., C. Lai, J. Axt, C. Ebersole, M. Herman, P. Devine, and B. Nosek. 2019. A meta-analysis of change in implicit bias. *Journal of Personality and Social Psychology* 117 (3): 522–559.

Freeman, L., and H. Stewart. 2020. Sticks and stones can break your bones and words can really hurt you: A standpoint epistemological reply to critics of the microaggression research program. In *Microaggressions and philosophy*, eds. L. Freeman, and J. Weekes Schroer, 36–66. Routledge.

Freeman, L., and H. Stewart. 2021. Toward a harm-based account of microaggressions. *Perspectives on Psychological Science* 16 (5): 1008–1023.

Fricker, M. 2007. *Epistemic injustice: Power and the ethics of knowing*. Oxford: Oxford University Press.

Fricker, M. 2017. Evolving concepts of epistemic injustice. In *Routledge handbook of epistemic injustice*, eds. I. J. Kidd, and J. Medina, & G., Polhaus Jr., New York: Routledge.

Fyfe, S. 2018. Testimonial injustice in international criminal law. *Symposion* 5: 155–171.
Gawronski, B., and G. V. Bodenhausen. 2011. The associative–propositional evaluation model: Theory, evidence, and open questions. *Advances in Experimental Social Psychology* 44: 59–127.
Gawronski, B., S. Brannon, and G. Bodenhausen. 2017. The associative-propositional duality in the representation, formation, and expression of attitudes. In *Reflective and impulsive determinants of human behavior*, eds. R. Deutsch, B. Gawronski, and W. Hofmann, New York: Psychology Press.
Gawronski, B., M. Morrison, C. Phills, and S. Galdi. 2017. Temporal stability of implicit and explicit measures: A longitudinal analysis. *Personality and Social Psychology Bulletin* 43: 300–312.
Goldenberg, M. J. 2015. Whose social values? Evaluating Canada's "Death of evidence" controversy. *Canadian Journal of Philosophy* 45: 404–424.
Greenwald, A. G., M. R. Banaji, and B. A. Nosek. 2015. Statistically small effects of the Implicit Association Test can have societally large effects. *Journal of Personality and Social Psychology* 108: 553–561.
Greenwald, A. G., D. McGhee, and J. Schwartz. 1998. Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology* 74: 1464–1480.
Greenwald, A. G., T. A. Poehlman, E. L. Uhlmann, and M. R. Banaji. 2009. Understanding and using the Implicit Association Test: III. Meta-analysis of predictive validity. *Journal of Personality and Social Psychology* 97: 17–41.
Hannay, J. W., and B. K. Payne. 2022. Effects of aggregation on implicit bias measurement. *Journal of Experimental Social Psychology* 101: 104331.
Hartwig, M., and C. H. Bond. 2011. Why do lie-catchers fail? A lens model meta-analysis of human lie judgments. *Psychological Bulletin* 137 (4): 643–659.
Hicks, D. J. 2014. A new direction for science and values. *Synthese* 191: 3271–3295.
Hitchcock, C. 1995. The mishap at Reichenbach Fall: Singular vs. general causation. *Philosophical Studies* 78: 257–291.
Honeycutt, N., L. Jussim, A. Careem, and J. Neil Lewis. 2020. Are STEM faculty biased against female applicants? A robust replication and extension of Moss-Racusin and colleagues (2012). PsyArXiv. https://psyarxiv.com/ezp6d/.
Jackson, E., and M. G. Turnbull. 2023. Permissivism, underdetermination, and evidence. In *The Routledge handbook of the philosophy of evidence*, eds. M. Lasonen-Aarnio, and C. Littlejohn, London: Routledge.
Kent, J. A., V. Patel, and N. A. Varela. 2012. Gender disparities in health care. *Mount Sinai Journal of Medicine: A Journal of Translational and Personalized Medicine* 79 (5): 555–559.
Kinoshita, S., and M. Peek-O'Leary. 2005. Does the compatibility effect in the race Implicit Association Test reflect familiarity or affect? *Psychonomic Bulletin & Review* 12 (3): 442–452.
Kotzee, B. 2013. Educational justice, epistemic justice, and leveling down. *Educational Theory* 63 (4): 331–350.
Kurs, R., and A. Grinshpoon. 2017. Vulnerability of individuals with mental disorders to epistemic injustice in both clinical and social domains. *Ethics & Behavior* 28 (4): 336–346.
Kvam, P. D., C. Smith, L. H. Irving, and K. Sokratous. 2022. Improving the reliability and validity of the IAT with a dynamic model driven by associations. https://psyarxiv.com/ke7cp/.
Lackey, J. 2006. It takes two to tango. In *The epistemology of testimony*, eds. J. Lackey, and E. Sosa, 160–189. Oxford: Oxford University Press.
Lackey, J. 2020. False confessions and testimonial injustice. *Journal of Criminal Law & Criminology* 110 (1): 43–68.
Lai, C. K., K. M. Hoffman, and B. A. Nosek. 2013. Reducing implicit prejudice. *Social and Personality Psychology Compass* 7 (5): 315–330.
Machery, E. 2016. De-Freuding implicit attitudes. In *Implicit bias and philosophy, Metaphysics and epistemology, Vol. 1*, eds. M. Brownstein, and J. Saul, Oxford: Oxford University Press.
Machery, E. 2017. Do indirect measures of biases measure traits or situations? *Psychological Inquiry* 28 (4): 288–291.
Machery, E. 2021. Anomalies in implicit attitudes research. Wiley Interdisciplinary Reviews: Cognitive Science, e1569. https://doi.org/10.1002/wcs.1569.
Mandelbaum, E. 2016. Attitude, inference, association: On the propositional structure of implicit bias. *Noûs* 50 (3): 629–658.
Meissner, F., and K. Rothermund. 2013. Estimating the contributions of associations and recoding in the implicit association test: the ReAL model for the IAT. *Journal of Personality and Social Psychology* 104 (1): 45.

Mercier, H. 2020. *Not born yesterday*. Princeton: Princeton University Press.
Mitchell, G., and P. Tetlock. 2006. Antidiscrimination law and the perils of mindreading. *Ohio State Law* 6: 1023–1121.
Mitchell, G., and P. E. Tetlock. 2017. Popularity as a poor proxy for utility: The case of implicit prejudice. In *Psychological science under scrutiny: Recent challenges and proposed solutions*, eds. S. Lilienfeld, and I. Waldman, New York: Wiley-Blackwell.
Moss-Racusin, C. A., J. F. Dovidio, V. L. Brescoll, M. J. Graham, and J. Handelsman. 2012. Science faculty's subtle gender biases favor male students. *Proceedings of the National Academy of Sciences*, *109*(41), 16474–16479.
Norton, M. I., J. A. Vandello, and J. M. Darley. 2004. Casuistry and social category bias. *Journal of Personality and Social Psychology* 87 (6): 817–831.
Nosek, B. A., C. B. Hawkins, and R. S. Frazier. 2012. Implicit social cognition. In *Handbook of Social Cognition*, eds. S. Fiske, and C. N. Macrae, 31–53. New York: Sage.
Oswald, F. L., G. Mitchell, H. Blanton, J. Jaccard, and P. E. Tetlock. 2013. Predicting ethnic and racial discrimination: A meta-analysis of IAT criterion studies. *Journal of Personality and Social Psychology* 105: 171–192.
Payne, B. K., C. M. Cheng, O. Govorun, and B. D. Stewart. 2005. An inkblot for attitudes: Affect misattribution as implicit measurement. *Journal of Personality and Social Psychology* 89: 277–293.
Payne, B. K., and B. Gawronski. 2010. A history of implicit social cognition. Where is it coming from? Where is it now? Where is it going? In *Handbook of implicit social cognition. Measurement, theory, and applications*, eds. B. Gawronski, and B. K. Payne, 1–15. New York: The Guilford Press.
Payne, B. K., H. A. Vuletich, and K. B. Lundberg. 2017. The bias of crowds: How implicit bias bridges personal and systemic prejudice. *Psychological Inquiry* 28: 233–248.
Psillos, S. 2015. Evidence: Wanted, alive or dead. *Canadian Journal of Philosophy* 45: 357–381.
Quillian, L., D. Pager, O. Hexel, and A. H. Midtbøen. 2017. Meta-analysis of field experiments shows no change in racial discrimination in hiring over time. *Proceedings of the National Academy of Sciences*, *114*(41), 10870–10875.
Rae, J. R., and K. R. Olson. 2018. Test-retest reliability and predictive validity of the Implicit Association Test in children. *Developmental Psychology* 54: 308–330.
Samulowitz, A., I. Gremyr, E. Eriksson, and G. Hensing. 2018. "Brave men" and "emotional women": A theory-guided literature review on gender bias in health care and gendered norms towards patients with chronic pain. *Pain Research and Management*, *2018*.
Saul, J. 2013. Implicit bias, stereotype threat and women in philosophy. In *Women in philosophy: What needs to change?* eds. F. Jenkins, and K. Hutchison, 39–60. Oxford: Oxford University Press.
Schimmack, U. 2021. The Implicit Association Test: A method in search of a construct. *Perspectives on Psychological Science* 16 (2): 396–414.
Schwarz, N. 2007. Attitude construction: evaluation in context. *Social Cognition* 25: 638–656.
Shieber, J. 2012. Against credibility. *Australasian Journal of Philosophy* 90: 1–18.
Sitton, S. C., and S. T. Griffin. 1981. Detection of deception from clients' eye contact patterns. *Journal of Counseling Psychology* 28 (3): 269–271.
Uhlmann, E. L., V. L. Brescoll, and E. L. Paluck. 2006. Are members of low status groups perceived as bad, or badly off? Egalitarian negative associations and automatic prejudice. *Journal of Experimental Social Psychology* 42 (4): 491–499.
Webb, T. L., P. Sheeran, and J. Pepper. 2010. Gaining control over responses to implicit attitude tests: Implementation intentions engender fast responses on attitude-incongruent trials. *British Journal of Social Psychology* 51 (1): 13–32.
Wittenbrink, B., C. M. Judd, and B. Park. 1997. Evidence for racial prejudice at the implicit level and its relationship with questionnaire measures. *Journal of Personality and Social Psychology* 72 (2): 262–274.
Wylie, A. 2003. Why standpoint matters. In *Science and other cultures: Issues in philosophies of science and technologies*, eds. R. Figueroa, and S. Harding, 26–48. Routledge.
Abramson, K. 2014. Turning up the lights on gaslighting. *Philosophical Perspectives* 28(1): 1-30.