A Critical Analysis of Scientifically Pervasive Claims about Stereotypes, Prejudice, and Discrimination

Lee Jussim Sonia Yanovsky Akeela Careem
Rutgers

Nathan Honeycutt

Danica Finkelstein
Rutgers

To appear in T. Nelson (editor), *Handbook of Stereotyping, Prejudice and Discrimination*, Third Edition (expected publication 2025). New York: Routledge.

Abstract

In this chapter, we first describe some of the foundational principles underlying when conclusions claiming the mantle of "science" should and should not be granted credibility. We then apply those principles to a critical evaluation of claims made in four areas of intergroup relations: implicit bias, microaggressions, stereotype threat and job discrimination. We identify many common claims regarding implicit bias, microaggressions and stereotype threat that do not rise to the level of scientific facts. The evidence on racial discrimination is much sounder, but has produced some surprising results, including recent research finding that acts of discrimination occur at very low levels, even though disparate outcomes produced by discrimination can be substantial.

In this chapter, we first describe some of the foundational principles underlying when conclusions claiming the mantle of "science" should and should not be granted credibility. We then apply those principles to a critical evaluation of claims made in four areas of intergroup relations: implicit bias, microaggressions, stereotype threat and job discrimination.

Merton's Norms of Science

Merton (1942) argued that science *earned* a special place in the marketplace of ideas because it relied on norms that, when compared to other sources of ideas (such as religion or politics), dramatically elevated its validity and credibility. There is considerable debate about the extent to which scientists actually practice these norms (e.g., Anderson et al., 2010; Mitroff, 1974; Mulkay, 1976; Zimann, 2000). Nonetheless, as *prescriptive norms*, norms about how science *should* operate to deserve its special credibility status, our view is that he was right in the sense that, when scientists actually act according to them, their science is vastly more credible than when they do not. Given that social scientists often argue that their work deserves this heightened credibility on the (in our view, debatable) grounds that they follow one or more Mertonian norms (e.g., Fiske & Borgida, 2011; Jost, 2011; Van Bavel, Reinero, Harris, Roberts & Parnamets, 2020), we trust we are not alone in our belief that, at minimum, this is how social scientists should conduct themselves.

Organized skepticism. Merton (1942) used the term "skepticism" in its modern colloquial sense: To be dubious, critical, and even suspicious of any claim until the evidence in support of that claim was overwhelmingly clear and compelling. "Organized" meant that skepticism was not supposed to be restricted to individual scientists: it is intended to be a core feature of the scientific community. Scientific claims should withstand intense scrutiny, severe tests (Mayo, 2018) and attempts at falsification (Meehl, 1990; Popper, 1959/2005) before they become widely accepted as true.

Universalism. Universalism means that scientific knowledge does not hinge on the power, demographics, privilege, status, or fame of individual scientists. One could contest any idea on its merits – on the facts, evidence, and methods used – but one could reject no idea based on the identity of the scientist producing it. According to this norm, the same logical-empirical standards for supporting or refuting claims must apply to all scientists, no matter their status, demographics, identities, politics, ethnicity, or gender (Merton, 1942). The universalism norm is readily interpretable as rejecting positionality statements (in which authors discuss their identities in supposed attempts to articulate how this might bear on and bias their research) as part of scientific publishing (Savolainen, Casey, McBrayer & Schwerdtle, 2023). Self-reports of bias (as with positionality statements) are, in our view, almost completely useless, whereas systematic assessments of bias in specific research articles are extremely useful (for examples, see, e.g., Jussim & Honeycutt, 2023; Macnamara & Burgoyne, 2023).

Communitarianism. Merton (1942) used the term "communism," but to avoid confusion, most subsequent authors have called it "communalism" or "communitarianism." Data, methods, ideas, and theories, the core products of scientific knowledge, must be shared with both the wider community of scientists and the wider society, in part out of a moral principle that the fruits of science should be disseminated widely for maximum benefit, and partially out of the pragmatic concern that information sharing is necessary for organized skepticism to work. The pragmatic issue should be obvious: Skeptically vetting claims, data, theories or methods requires access to them; the wider the access, the greater the potential for the type of skeptical vetting necessary to ensure that only true ideas become canonized.

Disinterestedness. This norm exhorts scientists to prevent their personal or political values from influencing their scientific conclusions. Merton (1942) saw disinterestedness not as a property of the scientist, but as a property of the institution of science. The hope was that, because manifestly false claims would quickly be discovered by fellow scientists, few scientists would promote such claims, and, if they did, science would quickly self-correct.

The value of the four norms for scientific credibility can be contrasted with the consequences of failing to implement them. What are alternatives to skepticism? Credulity and confirmation bias. The history of science is littered with so many false claims once believed to be true (e.g., young Earth, spontaneous generation of life, implicit biases are "unconscious") that calls for science to be dominated by credulously accepting all claims at face value based on initial reports strikes us as absurd. A science dominated by confirmation bias would produce claims only consistent with what one already believes, meaning that science could never change anyone's belief. Thus, our view is that there is no serious scientific alternative to skepticism.

What is the alternative to universalism? Particularism (the idea that each person or researcher has their own personal "truths"), which rapidly gives rise to logical incoherence. If ideas must be evaluated on the basis of the identities of scientists, rather than their logic, data, and methods, what is to be done when two scientists who share an identity group produce opposing conclusions? They both cannot possibly be right, at least not if they are reaching conclusions about the same phenomena. We could bring up cases here where Black scholars disagreed about racism, White scholars disagreed about racism, Jewish scholars disagreed about antisemitism, and on and on for virtually any identity group pairing. Of course, this also means that White scholars may disagree with Black scholars about racism, and Jewish and nonJewish scholars may disagree with one another about antisemitism. These disagreements cannot be scientifically resolved by playing identity cards; they must be resolved by evidence and analysis.

The great modern manifestation of communitarianism in psychology emerged from the Replication Crisis. We refer specifically to efforts to change norms and practices so that researchers routinely post their data publicly whenever possible. The alternative here is for researchers to fail to do so; and this failure to permit other researchers to see one's data was one likely contributor to the Replication Crisis (e.g., Wicherts, Bakker, & Molenaar, 2011).

The alternative to disinterestedness is interestedness, which means being committed to a conclusion prior to conducting the research. Interestedness increases the risk of introducing biases into the research, raising the likelihood that one will likely end up with inaccurate or skewed research conclusions.

What Is And Is Not A Scientific Fact?

How does one know when a new scientific discovery has been established with sufficient credibility to warrant taking it seriously as a fact? In this section, we articulate our standards for believing some new discovery is sufficiently well-established to constitute a scientific fact.

Something is not a fact just because some scientist or some article says so. What, then, is a scientific fact and how would anyone know one when one sees one? We use the definition provided by the late evolutionary biologist, Stephen Jay Gould (1981): "In science, 'fact' can only mean 'confirmed to such a degree that it would be perverse to withhold provisional assent.'" This is worth unpacking in reverse order. "Provisional assent" refers to the idea that, in principle, all scientific ideas are subject to change if dramatic, overwhelming evidence is provided overturning some "established" conclusion. But doing so is not easy because when there is a mountain of evidence that some claim is true, withholding provisional assent is perverse. "Perverse" in this context means rigidly dogmatically wrong. It is, in the 21st century, perverse to believe that the Earth is flat, that women are, on average, taller than men, or that SAT scores are, on average, identical across American racial/ethnic groups (although the meaning and sources of those differences are hotly contested).

We rely on Gould's definition throughout this chapter. However, we also expand upon its implications by taking its converse (Jussim, Stevens, Honeycutt, Anglin & Fox, 2019, p. 281): "Anything *not* so well established that it would *not* be perverse to withhold provisional assent is *not* an established scientific fact." Given the Replication Crisis and the slew of known scientific biases reviewed next, our general perspective is that it is not perverse to withhold provisional assent with respect to many supposedly established claims in intergroup relations.

Different Types of Inaccurate Claims in Scholarship

A claim (including but not restricted to published claims) can be inaccurate in any of a variety of ways. It can be *false*, in that there is ample reason to know that the claim is untrue. Claiming that the Earth is flat is false.

Claims can also be inaccurate because they are *unjustified* by the evidence. This can occur when a published paper makes an empirical truth-claim without presenting *any evidence* or by citing a paper that *argues* for the claim but also presents no evidence for it. Claiming that Mars once had life is not *known to be false* because it is possible. But, at present, there is no conclusive evidence that Mars actually once had life on it, so claiming that it did is unjustified. Social psychology is replete with these sorts of unjustified claims.

Claims can also be inaccurate by virtue of being *misleading* when they present an incomplete and distorted view of the evidence. These claims are not *false* because, in this case, there *is some evidence* supporting the claim. And they are not *unjustified* in the sense of there being *no evidence supporting the claim.* For example, if one was to claim that New Jersey is colder than Alaska, this would be *misleading* even if one can point to some actual days in which the high temperature was colder somewhere in N.J. than somewhere in Alaska. This assertion overlooks the broader context of average temperatures, climate patterns, and seasonal variations, providing an incomplete and distorted view of the overall temperature conditions in both locations. This is obvious, but it applies to scholarship as well.

Known Threats To Validity Of Psych Science: In General

There are many known reasons to be skeptical of claims published in psychology journals, and sometimes even of whole literatures. Put differently, there are often many reasons not to treat published claims as having established new scientific facts. We review some of those reasons here: the Replication Crisis, allegiance bias, use of "persuasive communication devices," propaganda scholarship, and social pressure to make or not make some claims. There can be overlap between some of them. One may sometimes be deployed in the service of another (e.g., persuasive communication devices may be deployed in the service of propaganda scholarship or political biases; propaganda often involves politicized claims but it can refer to other types of claims as well). Regardless, there are enough differences among them to warrant treating each separately.

The Replication Crisis. The Replication Crisis refers to a series of events starting around 2009 that revealed a slew of suboptimal scientific practices common throughout psychological science and one of their main manifestations, research that other scientists could not replicate (see, e.g., Nelson, Simmons, & Simonsohn, 2018, for a review). In this context, it was then not surprising that the long term success rate for replications in psychology has hovered right around 50% (e.g., Boyce, Mathur, & Frank, 2023; Open Science Collaboration, 2015; Scheel et al., 2021). This means that published psychology studies should not be treated as establishing new "scientific facts" just because they have been published in a peer reviewed journal. Instead, fact status should generally be reserved for findings that have been replicated by independent teams of researchers, preferably using some version of the registered report format (in which the research is accepted in principle by a journal based on the proposal, so that the results do not influence publication; e.g., Nosek & Lakens, 2014). Absent this type of replication, it is not perverse to withhold provisional assent that something published has established a new scientific fact.

Allegiance bias. Allegiance bias refers to the discovery that, sometimes, research produced by advocates of some phenomenon or intervention report larger effects than that produced by more disinterested (in the Mertonian sense) scientists. Allegiance biases were first identified when those studying the effectiveness of different types of psychotherapy (Smith, Glass, & Miller, 1980; Luborsky et al., 1999) discovered that proponents sometimes yielded stronger effects for the type of therapy they advocated than for alternative therapeutic approaches. For example, Smith et al. (1980) found that proponents obtained effect sizes of d=.29 larger than those obtained by others. Similarly, Luborsky et al. (1999) found that effect sizes (d) ranged from .25 to 1.00 *higher* when the researchers were advocates for that type of therapy.

Although allegiance biases do not *always* occur, the effect has been found to be robust (e.g., Munder et al., 2013) and is not restricted to clinical therapies. A recent meta-analysis found that growth mindset studies produced a small but statistically significant overall effect of d=.05, which was reduced to d=.02 and statistical nonsignificance when only studies conducted by authors without a financial incentive (such as being registered with a speakers bureau to give paid talks or having lucrative book sales promoting growth mindset) were included (Macnamara & Burgoyne, 2023).

The point is not that all research produced by researchers with high allegiance to some intervention or phenomenon is bad or biased. It is, instead, that allegiance bias raises credibility questions about effects produced by advocates. Our view is that a prudent approach to interpreting such work is to withhold belief in findings by researchers with allegiances pending confirmation by more disinterested research teams.

Persuasive communication devices. "Persuasive communication devices" refers to the rhetorical techniques used by writers of academic articles to: 1. Render their work clear and engaging, which is completely legitimate *if* such writing accurately reflects the underlying empirical research, including its weaknesses and limitations; or 2. Render their work to appear more persuasive, more important, and less flawed than it really is, which, we argue, is not at all legitimate (Corneille et al., 2023). Corneille et al. (2023) identified 21 such devices, although they also pointed out that their list was far from comprehensive and invited other scholars to identify additional ones. Although we do not review all 21 devices here, we briefly describe three in order to give a sense of how researchers can produce impressive-seeming but ultimately unjustified scientific claims:

- Ignoring previous work that is inconsistent with one's preferred conclusion or narrative.
- Selective reporting: not reporting or downplaying results that would weaken the paper.
- Selective appeal to rigor: holding claims one opposes to a higher scientific, evidentiary or methodological standard than one applies to claims one supports.

Propaganda scholarship. We define propaganda as attempts to persuade others of something that is either outright false, unjustified or misleading (as described above). As might be expected, the *persuasive communication devices* described previously can and often are used in propaganda scholarship. *Propaganda scholarship,* then, refers to scholarship that seeks to persuade others that certain things are facts when they are, at the time, known to be either false, unjustified or misleading.

Of course, science sometimes gets things honestly wrong, but this can be distinguished from propaganda scholarship. The key distinction is whether the wrong claim is knowably false, unjustified or misleading at the time. Although there may be gray areas, this is often obvious, such as when a "scholarly" article makes a truth-claim without citing any evidence. This is unjustified propaganda (see Jussim, 2012, for many examples of claims of stereotypes being "inaccurate" without referencing studies of the accuracy of stereotypes). Similarly, sometimes, peer reviewed articles make manifestly unjustified claims, such as when Ellemers (2018) claimed that gender stereotypes were broadly inaccurate without *even mentioning, let alone reviewing or critiquing,* any of the 11 articles reporting 16 separate studies assessing gender stereotype accuracy and, typically, finding it substantial (see Jussim, 2018 for a list of those 11 articles).

Gambrill (2010, 2012; Gambrill & Reimann, 2011) presented a perspective on propaganda scholarship. Gambrill did not primarily focus on *political propaganda* and, instead, focused on propaganda promoting false, unjustified, and misleading *social scientific claims,* mostly but not exclusively in what she referred to as "the helping professions" (social work, psychotherapy, etc.).

Gambrill and Reimann (2011) presented a checklist identifying features of propaganda that could be used by peer reviewers of scholarly articles, and we believe this can be used by any professional when *critically evaluating* articles, so that its value goes beyond *pre-publication peer review* and is just as useful for formal and informal *post-publication peer review* (such as shall be conducted herein). Gambrill (2010) and Gambrill and Reimann's (2011) included some propaganda-indicators that overlap with both the *persuasive communication devices* and Bodi's (1995) indicators discussed previously (such as selectively ignoring contrary findings). They presented 32 such indicators, of which we present only a few here that do not overlap with others already presented:

- Possible harms of the view presented are not described
- Vague terms are used
- Evidence is not described in quantitative terms
- Use of personal attacks and innuendo

- Oversimplification
- Hypocrisy (extolling some virtue or behavior but not implementing it one's self).
- Using social pressure or censorship to attempt to suppress, rather than debate or refute, alternative views.

Social pressure to present or not present some finding or conclusion. Social pressure to make or not make a particular claim can also undermine the validity of entire literatures. Although Bodi (1995) and Gambrill (2010) highlighted attempts to suppress certain ideas as a marker of propaganda, it was Joshi (2022) who first fully articulated how this works. Consider first social pressure to *not* make a claim, whether informal or official (e.g., through outright censorship – see Clark et al., 2023, for a review), say that "Conclusion X is Wrong." To the extent that such social pressure is successful at suppressing some or all of the falsification of X, the "scientific" literature may have many published papers appearing to find that X is True, whereas the full scope of the evidence (i.e., where they may actually be some, or even far more, evidence showing that X is False) does not get represented in the published literature. Thus, the literature may overwhelmingly find that X is True, but this is only because much of the evidence showing that X is False has been suppressed.

Demonstrating that a literature is *wrong* because some evidence has been suppressed would seem to be an impossible task because, one might assume, all that one can actually show is what *is* in the literature, not what is not in the literature. However, there are several ways to show that social pressure to make or not make certain claims likely does exist:

1. The presence of academic "book burning." Jussim, Honeycutt, et al. (in press) used this term to refer to a certain obvious type of academic censorship. These acts of "book burning" included several examples of retractions of papers, usually at the hands of academic outrage mobs (typically via social media), despite no evidence that those papers violated the retraction guidelines of the Committee on Publication Ethics (n.d.). Instead, those retracted papers made claims that violated the social justice sensibilities of the mob (such as defending colonialism or criticizing diversity, equity and inclusion initiatives). This is exactly the type of censorship highlighted by Gambrill (2010) as indicative of propaganda.
2. Stark differences between the conventional published literature and registered reports (the format whereby journals make an accept/reject decision based on a research proposal and allow the paper to be published, regardless of the subsequent results). When registered reports produce starkly different findings than routinely appears in a published literature, it at least raises red flags about social pressure distorting the published literature. This could occur, for example, because once a finding appears in the literature, it is easier to publish similar papers than contrary papers.
3. Vast citation differences to papers finding X than to papers of comparable or higher quality finding Not X or the opposite of X (see Honeycutt & Jussim, 2020 for an example). Researchers respond to incentives (e.g., Jussim et al., 2019), and high citation counts have many benefits. Sometimes, they are used in hiring and promotion decisions to evaluate "impact" (indeed, some versions of citation counts are called "impact factors"). Furthermore, if academics would prefer to have more influence and visibility rather than less, knowing that a certain type of claim produces higher citations may create social pressure to make those claims.

A Heuristic Guide to When Scholarly Claims Should be Treated as Facts

The Replication Crisis means that no single study or paper should be treated as "fact" until replicated by independent researchers. Allegiance biases mean that no line of research should be treated as "fact" until confirmed by independent researchers without such allegiances. Persuasive communication devices, propaganda scholarship, and political biases mean that claims cannot automatically be taken at face value *simply* because they appear in a peer reviewed social science outlet. In all such cases, it is *not* perverse to believe otherwise, although it is always possible that, in the fullness of the time, the results of the single paper or of the advocates will prove to be robust and replicable, and therefore may, someday in the future, rise to the level of scientific fact.

When, then, *can* social science be treated as a scientific fact? Because a detailed discourse on social science validity and epistemology has often required an entire book (e.g., Popper, 1959/2005), we provide some useful heuristics here rather than some comprehensive guide to necessary and sufficient conditions. These heuristics include:

- Absence of the types of threats to validity described above.
- Large samples and/or high statistical power.
- Large effect sizes.
- Few p-values between .01 and .05 for critical findings.
- Adversarial collaborations for theoretically or politically controversial topics.
- Use of registered reports (or at least pre-registration *if* the research sticks to the pre-registration or deviations are clearly and transparently reported).
- Many successful pre-registered replications
- Includes open data and materials.

For reviews supporting these as useful criteria when evaluating the strength of psychological research see, e.g., Clark et al., 2022; Fraley et al., 2022; Jussim et al., 2019; Nelson et al., 2018; Scheel et al., 2021). Even this list, however, is not complete and other criteria may be invoked, depending on the methods, statistics and claims. Furthermore, depending upon the claim, these criteria may be irrelevant. Consider the claim “there are only white swans.” One needs none of these criteria for a credible report (say, multiple scientists as eye-witnesses and clear undoctored photographic evidence) if one finds a single black swan. Similarly, one needs no carefully controlled experiments, low p-values or pre-registration to conclude that it is a fact that a person who jumps out of a plane flying at 20,000 feet without a parachute or flying equipment will die.

This also raises the question of how to evaluate older literatures that predate the new norms around large samples, pre-registration, open materials and adversarial collaborations. In general, our view is that more skepticism is justified regarding such older literatures, but this does not mean *nothing should be believed.* First, some of it does meet some of the standards listed above (absence of the red flags identified herein, large samples, large effects, etc.). Second, many *results* may be replicable, even if their common interpretations are dubious. For example, conservatives clearly score higher on “subtle” measures of racism, such as symbolic and modern racism, even if the *interpretation* of those as pure measures of racism per se is dubious (Cramer, 2020). Third, some of the older literature *has* been confirmed in one or more replications decades after the original, often with large samples. For example, the original Asch (1956) conformity in judgments of line length study has recently been replicated with very similar results (Franzen & Mader, 2023). Fourth, some topics have received extensive attention over the decades and consistently produce moderate to large effects or relationships regardless of researcher, precise operationalization, context, and when they have been studied. Some examples include the heritability of intelligence and its manifestations (such as academic achievement; e.g., Harden, 2021; Neisser et al., 1996); most of the heuristics and biases first uncovered by Kahneman & Tversky (e.g., Tversky & Kahneman, 1974; Brewer et al., 2007); and the strong relationship of party identification to presidential voting (Dalton, 2021; Declercq et al., 1975).

With these issues in mind, we now proceed to four areas in intergroup relations that have, at one time or another, been wildly popular and about which strong claims have been made. These cases have been purposely selected because, according to the standards articulated herein, most (but not all) fail to rise to the level of “scientific fact” in the sense of our corollary to Gould’s definition.

We do not present a thorough review of these topics. Instead, in each case, we focus on some of the most extreme or common claims made by advocates. Thus, although in each case, we present reviews by advocates for completeness, our focus here is not on a thorough review of the relevant literatures. It is, instead, on whether certain central claims are credible.

Consider, for example, the intermittent appearance in the psychological literature of claims to have discovered paranormal phenomena (Bem, 2011; Cardeña, 2018). We dismiss them without review on several grounds:

1. They violate known laws of physics (e.g., Reber & Alcock, 2020)

2. Psi researchers often try to render their phenomena "consistent" with physics but:
    a. This still does not make them true (Wagenmakers et al., 2011).
    b. The phenomena can be consistent with some laws and still inconsistent with others, and if they are inconsistent with any, dismissing the claims out of hand is warranted.
    c. We challenge any reader to find a single article in a peer reviewed physics journal that reports confirmation of paranormal phenomena.
3. The processes involved in mindreading, future-telling, "seeing" at a distance and the like have never been identified.

Unitil these problems are overcome, we see no need to review the vast evidence produced by psi researchers and critically evaluate its strengths and weaknesses. These weaknesses are plausibly viewed as sufficient to dismiss the whole area for failing to meet conventional scientific standards, including logical coherence, so that it does not matter how many studies have been conducted or how well-controlled they are. But even if one does not go quite that far, these flaws are sufficiently severe that no one would be "perverse" if they withheld "provisional assent" that psi is an established scientific fact.

Social science on intergroup relations is not usually as scientifically unsound as research on psi, but that is not the point. Instead, the point is that when an area has certain, or, in some cases, too many, weaknesses, flaws, limitations, and misinterpretations, we believe no scientist would be perverse to withhold belief. Therefore, our review emphasizes areas of intergroup relations that are or have been highly popular and treated as fact, but which, in our view, do not rise to the level of scientific fact, and we explain why in each instance. When the flaws are sufficiently severe, a full "even-handed" review of all the pros and cons regarding the evidence is not necessary to withhold belief.

Implicit Bias, Microaggressions and Stereotype Threat: Strong Claims, Few Scientific Facts

In this section, we review research on implicit bias, microaggressions, and stereotype threat, because they are central concepts in modern psychological work on intergroup relations and because they share the following characteristics:

1. Advocates have presented their conclusions as if they are scientific facts.
2. They are not scientific facts in the Gouldian sense described herein. Put differently, in some cases there are hundreds, maybe thousands, of academic papers treating these theories, hypotheses or conclusions as if they are true, even though, as we shall show, it is not perverse to believe otherwise.

As such, this review focuses on identifying the critical weaknesses in these areas sufficiently severe to undercut claims or assumptions that they represent "facts." It does not aspire to review all evidence that bears on any of these key claims.

Implicit Racial Bias

Many reviews by advocates that presume or argue that implicit biases are real, socially important, and relevant to the real world can be readily found (e.g., Banaji & Greenwald, 2016; Greenwald & Lai, 2020; Kang & Banaji, 2006). Interested readers should consult them, but we do not delve into them here because we find neither their arguments nor evidence persuasive regarding the validity of once-common claims about implicit bias.

The definition problem: Falsification, logical incoherence, inconsistency and vacuousness. It is useful, however, to consider what those common claims about implicit bias once were (and still sometimes are). In a 2017 conference on the implicit bias controversies sponsored by The National Science Foundation (which the first author attended, and which produced an edited book – Krosnick et al., in press), Greenwald (2017, one of the originators of the IAT and one of its strongest advocates) presented this as the working definition of implicit bias from 1998-2017:"Introspectively unidentified (or inaccurately identified) effects of past experience that mediate discriminatory behavior."

Almost every aspect of this definition is false or incoherent:

- IAT scores are not "introspectively unidentified" (e.g., Hahn et al., 2014).

- The IAT is a reaction time measure; i.e., it does not measure behavioral discrimination.
- The IAT is a single measure; as such, by itself, it constitutes mediation of nothing because no single measure can ever constitute evidence of mediation, which requires at least three variables.

If B mediates the effects of A on C, then A, B, and C must be three different variables. Therefore, given that the IAT is by far the most common measure of "implicit bias" (see any of the reviews cited above), if Greenwald's definition aptly captures what researchers thought they were measuring with the IAT, they were wrong and the research does not show what they believed.

Of course, many papers offer other definitions of implicit bias. The problem is that those definitions are often widely different from one another. Some emphasize culture, some don't; some emphasize behavior, some don't; some emphasize automaticity or unconsciousness, some don't; and, worse still, many vacuously offer no definition at all and seem to implicitly assume that "implicit bias is whatever is measured by the IAT" (see Jussim, Careem et al., in press, for a review). Thus, the field has no consensus definition as to what implicit bias even is.

Unidentified flying egalitarianism. To this day, it is widely assumed that scores of 0 on the IAT correspond to egalitarian responding, when, in fact, there is: 1. No evidence that this is true; 2. Considerable evidence that scores well above 0 correspond to egalitarian judgments and behavior; and 3. The point of egalitarian-ness moves around from study to study and, when it has been reported, has been found to be well above zero (e.g., Blanton, Jaccard, Strauts et al., 2015). Furthermore, the ubiquitous transformation of IAT reaction times to *D-scores* (Greenwald et al., 2003) is capable of producing a massive disproportion of scores above 0 (supposedly indicating "bias"), even when only 10% of respondents are actually biased (Blanton, Jaccard, & Burrows, 2015).

These findings have clear implications for other "implicit bias" claims. If the IAT has no stable score reflecting egalitarianism, one cannot be in the business of declaring what proportion of people show racial preferences and whether any particular IAT score constitutes a "weak" or a "strong" implicit racial preference (as was once common, see also the reviews by the advocates cited above). And if egalitarianism usually corresponds to scores well above zero, but conclusions of "preference" are based on proportion of scores above 0 or even some low cutoff (such as D = .1), these will be overestimates, perhaps extreme ones. Claims about how many people show "implicit preferences" for one group or another – which typically refer to how many claims are above some arbitrary cutoff – need not be taken seriously and are almost certainly overestimates.

A related major failure of this line of research is that it is not normative for researchers to report the benchmarked (against their other variables) point of egalitarianism for their IAT scores. This is not hard to do. The simple bivariate regression prediction equation is:

$$Y = C + b(X).$$

Y is the outcome, C is the regression constant, X is the predictor, and b is the coefficient relating X to Y. When Y is some outcome (let's say, discrimination), and X is the IAT, the equation becomes:

$$\text{Discrimination} = C + b(\text{IAT}).$$

Anyone can now solve for the IAT score that corresponds to zero discrimination:

$$0 = C + b(\text{IAT}), \text{ so}$$

$$0 - C = b(\text{IAT}), \text{ so}$$

$$- C = b(\text{IAT}), \text{ so}$$

$$-C/b = \text{IAT}$$

-C/b is the IAT score that corresponds to egalitarian behavior (i.e., zero discrimination). Anyone with an elementary understanding of regression can do this. At best, it is not common to do so, though it should be. At worst, and as far as we know, no one conducting original research using the IAT has ever reported this. If they did, the field could get much clearer information regarding what IAT scores correspond to egalitarian judgments and behavior.

Problems of temporal instability, lack of discriminant validity, failure to demonstrate causal effects, and measurement artifacts. Some of the additional limitations and weaknesses of the IAT and implicit bias construct include:

- The temporal stability of IAT scores is low, meaning that scores fluctuate considerably from administration to administration for the same people, and the reliability of other measures of implicit bias are also low (LeBel & Paunonen, 2011; Machery, 2022a).
- Despite once common conclusions to the contrary (see reviews by the advocates referenced earlier), claims that the IAT measured something different than explicit measures are dubious and contested at best (Machery, 2022a; Schimmack, 2021).
- The evidence that implicit biases (as measured by the IAT or any other method) actually cause inequality is nonexistent (Cesario, 2022a; Jussim, Careem, et al., in press; Machery, 2022a).
- Consistent with the "no causal effect" hypothesis, the one published study that examined whether changing IAT scores changed behavior found that it did not (Forscher et al., 2019).
- The IAT is not even a pure measure of strength of category association, let alone racism, and is instead tainted by artifacts and cognitive processes irrelevant to strength of association (Bluemke & Fiedler, 2009; Conrey et al., 2005).

An emerging scientific consensus around deep skepticism? Once one gets beyond the early advocates (e.g., those listed at the beginning of this section), there seems to be an emerging consensus of deep skepticism about most claims regarding implicit bias. Returning to the three ways scientific articles can promote inaccurate claims, few claims about the IAT or implicit bias are *outright false* but nearly all are either *unjustified or misleading.* Because of these and other issues, many social scientists have expressed deep skepticism about the implicit bias construct. It has been called "delusive" (Corneille & Hutter, 2020) and a line of research suffering "degeneration" (Cyrus-Lai et al., 2022). Unconscious racism has been called a "construct in search of a measure" (Blanton & Jaccard, 2008) and the IAT has been called "a method in search of a construct" (Schimmack, 2021). This is why Cesario (2022b, p. 164) concluded that, "The history of implicit bias research provides many cautionary lessons, from poorly defined concepts, to over-extension of the explanatory realm, to political activism far before a reasonable time."

Although we have mostly focused on the IAT here, other implicit measures perform no better and, often, even worse (e.g., Gawronski, Ledgerwood, & Eastwick, 2022a; LeBel & Paunonen, 2011). Or, as Gawronski, Ledgerwood, & Eastwick (2022b, p. 229) put it, "Theoretically, it seems very plausible that unconscious effects of social category cues contribute to social disparities in a significant manner. Yet, based on the currently available evidence, any such claims are premature." Put differently, few of the once-central claims about "implicit bias" rise to the level of scientific fact. Advocates can believe them and try to persuade others. But it is not perverse to provisionally reject claims that the most common measures of implicit bias detect unconscious racism, capture anything much different than explicit measures or cause real world disparities.

How did the field spend 25 years chasing a will-o-the-wisp? In a very interesting exchange (target article plus commentary):

- Gawronski, Brownstein, & Madva (2022) argued,"Finally, we note that Machery's [2022a] repeated invocation of how much remains unsettled "30 years after the implicit revolution" (p. 5) takes it as self-evident that 30 years is too long for a scientific research program to settle on the kinds of debates and frameworks we discuss here. This is not at all self-evident. Basic theorizing about emotion, perception, attention, and consciousness is multipronged, conflicted, and ongoing—appropriately so—after much more than 30 years."

- To which Machery (2022b) replied: "The problem is that 30 years after the development of contemporary indirect measures, foundational questions are up for grabs; worse some are just starting to be seriously addressed."

It is possible that both Gawronski et al. (2022) and Machery (2022b) are right. Perhaps it is appropriate to wait for 30 years of research to be sure the scientific community has had ample opportunity to skeptically vet the methods and conclusions in a new area of research; this is consistent with Merton's (1942) norm of "organized skepticism." If so, then it is not surprising that Machery (2022b) would also be right – if it takes 30 years, then the conclusions based on the first five to ten years of research should be taken as tentative and preliminary, rather than as a conclusive demonstration of new "scientific facts" (in the Gouldian sense).

IAT as propaganda scholarship. The early claims about the IAT and implicit bias by advocates has many of the earmarks of propaganda scholarship. It involved leaping to conclusions based on thin or nonexistent evidence. There never was evidence directly testing whether IAT scores were unconscious until the Hahn et al. (2014) study came along; no one bothered to benchmark IAT scores to behavioral measures of discrimination to find out whether 0 corresponded to egalitarianness until Blanton et al. (2015b). Second, the rush to attempt to change policy and law (Greenwald & Krieger, 2006 and Kang & Banaji, 2006 were both published in law journals) speaks of a rush *to advocacy*. If it is appropriate to expect 30 years or more to thoroughly vet claims of important new social scientific discoveries (as argued by Gawronski et al., 2022), then psychologists should be in no rush to change law and policy after several papers, or maybe even several dozen papers, published in *Journal of Personality and Social Psychology* or related outlets.

This is not to say that research on implicit biases or using the IAT is *inherently* propaganda. One reason we now know that so many of the early claims about implicit bias were false or dubious is because of the work that was not propaganda, some of which was conducted by those who might be considered advocates. Nonetheless, some critics of the work argue that the term implicit bias should be abandoned altogether (Corneille & Hutter, 2020) or that the IAT adds nothing to simpler self-reports (Schimmack, 2021). Still, it is possible that advocates of this work will provide sufficient modification to their methods and claims so that future research will produce credible scientific facts (in the Gouldian sense) about implicit bias. As Cyrus-Lai et al. (2022, p. 204) put it:

> To maintain itself, auxiliary assumptions such as multiple moderators in conjunction lead to respectable predictive validity correlations (Kurdi et al., 2019), social desirability bias on laboratory behavioral measures (Tierney et al., 2020), the cumulative consequences of minute discriminatory biases (Greenwald et al., 2015; Hardy et al., 2022), mismatched and suboptimal behavioral outcomes in studies examining causality (Gawronski et al., this issue), and aggregate-level crowd biases (Payne et al., 2017) must be invoked. Some or even all these defenses may hold empirically. And yet this heavily modified theoretical structure would still represent a major retreat from earlier models in which pervasive individual level implicit prejudices and stereotypes constitute major causal contributors to societal inequities.

Whether the failures of research on implicit bias justify abandoning work in the area is, therefore, a matter of scientific and professional judgment. Such judgments probably hinge, in part, on factors outside of scientific validity (such as whether scientists believe conducting such research will advance their careers, e.g., Jussim et al., 2019). Furthermore, the failures of work on implicit bias is not a general indictment of the scientific study of prejudice and discrimination. Indeed, the best solution to bad research or propaganda scholarship is better scholarship devoid of propaganda.

Social psychology has a long and troubled history of producing great outbursts of research purporting to demonstrate some amazing, dramatic phenomena, including biases produced by supposedly unconscious processes, only to find out, years later, that they had failed to establish credible new scientific facts (Jussim & Honeycutt, 2024; Jussim et al, 2016; Meehl, 1978, 1990). As Meehl (1990, p. 196) put it:

> …theories in the "soft areas" of psychology have a tendency to go through periods of initial enthusiasm leading to large amounts of empirical investigation with ambiguous over-all results. This period of infatuation is followed by various kinds of amendment and the proliferation of ad hoc hypotheses. Finally, in the long run, experimenters lose interest rather than deliberately discard a theory as clearly falsified.

Whether implicit bias research will fit Meehl's (1990) analysis or produce something more enduring remains to be seen. We do, however, recommend a moratorium on advocacy and applications (such as implicit bias trainings and judges' instructions to juries), except as experimental projects for research purposes. Such interventions could be justified after there is a long track record of producing scientific facts that constitute a credible foundation for them, clear evidence of the effectiveness of those interventions, and evidence that their beneficial effects are worth more than their costs and unintended negative effects.

Microaggressions

The definition problem. A minimum standard for taking an operationalization of some phenomena seriously is logical consistency between the definition and the operationalization. For example, defining flight speed as the rate of travel in the air means that it could not be measured by radar guns pointing at cars. Those radar guns may be technically superb, capturing the speed of cars with almost no error. They may be highly effective at catching speeders on highways. There is nothing wrong with such radar guns, *per se*, but operationalizing flight speed as the speed measured on radar guns pointed at cars would still be useless as a measure of flight speed, because cars do not fly.

Thus, for example, defining something as "unconscious" requires some operationalization empirically demonstrated to capture something of which people are unaware. If such evidence is not provided, the claim of "unconscious" is unjustified. Thus, a minimal condition to take any operationalization seriously is that it is logically consistent with the definition. If that definition includes an empirical claim, there needs to be strong evidence supporting that claim. If these conditions are not met, claims involving that operationalization do not rise to the level of "scientific fact. " In the same way that it is not perverse to disbelieve claims about *airplane* flight speed based on radar detectors pointing at *cars*, it is generally not perverse to disbelieve claims based on operationalizations that do not meet the standards required by the definition of the constructs being operationalized.

Another example, and one which we have already seen with implicit bias, is that, if a definition includes a causal relationship between two variables, say, "X is what happens when A causes B," any operationalization must meet three conditions:

1. It must have a logically coherent operationalization of A
2. It must have a logically coherent operationalization of B
3. It must provide evidence that A causes B.

This is not necessarily a problem, despite the subheader of this section. If all of these conditions are met, the X advocate may have strong justification for concluding that "X has occurred," though even this will depend on many other assumptions we put aside here (such as validation of A and B and the quality of the methods used to infer a causal relationship between them).

It can be considered an *a priori* problem for any researcher invoking such definitions, because the conditions required by the definition – in this case, that A caused B – must be established *before* one can treat X as an example of the construct. This is a very high and difficult standard to meet, if one wishes to conclude that one has found X. If any one of these conditions are *not* met for some operationalization, then it is *not* justified to refer to that operationalization or any findings involving it, as X. If X advocates still refer to that operationalization as "X," they are wrong, and, at minimum, no one would be "perverse" to disbelieve any claim they make about X. It is a critical point because as we shall demonstrate, work on microaggressions fails in exactly this manner.

Definitions of microaggressions. Racial microaggressions are usually defined as some sort of subtle racist insults and indignities. Sue et al. (2007, p. 273) defined microaggressions as:

> Microaggressions are brief, everyday exchanges that send denigrating messages to people of color because they belong to a racial minority group. In the world of business, the term 'microinequities' is used to describe the pattern of being overlooked, under-respected, and devalued because of one's race or gender. Microaggressions are often unconsciously delivered in the form of subtle snubs or dismissive looks, gestures, and tones. These exchanges are so pervasive and automatic in daily conversations and interactions that they are often dismissed and glossed over as being innocent and innocuous.

Williams (2020, p. 3) relied on Pierce's (1974) definition: "black-white racial interactions [that] are characterized by white put-downs, done in an automatic, preconscious, or unconscious fashion." She continues, "Microaggressions are by definition caused by socially conditioned racial biases and prejudices" (Williams, 2020, p. 6).

Disconnects between the definition and the operationalizations. Let's unpack the key claims here:

1. Microaggressions are acts committed by transgressors. This implies that the assessment of microaggressions requires measuring something about the people committing microaggressions.
2. Microaggressions are typically "automatic, pre-conscious, or unconscious," which would require assessing evidence that they are committed outside of conscious awareness.
3. They are caused by racism.

Should empirical research produce evidence of subtle insults perpetrated by transgressors, committed unconsciously (or automatically or pre-consciously), that are caused by racism, and that they are pervasive, there might *begin* to be justification for considering such evidence as at least documenting the existence of microaggressions, given the definition (depending on conventional standards for the quality of that evidence, including, e.g., pre-registration, well-validated measures, large samples, p-values below .01 etc.). But if empirical research produces *none of this evidence*, it does not deserve to be taken seriously as research on "microaggressions" by the advocates' own definition of the construct. Most research on microaggressions fits this latter category. Although we do not provide a thorough review of all research on microaggressions, we do review some of the work touted as "important" by microaggression advocates and show that even this work fails these standards.

Common measures of microaggressions assess targets' perceptions, not perpetrators' commission, of microaggressions. Questionnaires assessing "microaggressions" never assess microaggressions. Instead, at best (and more on this later), they assess targets' claims to perceive or experience microaggressions (e.g., Anderson et al., 2022; Nadal, 2011; Torres-Harding et al., 2012). These questionnaires provide no information about the behavior of transgressors at all. Consequently, they provide no direct information about microaggressions. Whether they provide any indirect evidence regarding microaggressions would require an assessment of the relationship of microaggressive behaviors of actual transgressors with scores on these scales. This has never been done.

No evidence that microaggressions are unconscious. To take Williams' (2020) definition as serious science, any operationalization of microaggressions would require scientific evidence that they are committed in an "automatic, preconscious, or unconscious fashion." This would be extraordinarily difficult to demonstrate, as shown in the prior section about the failures of the IAT. Nonetheless, it has not even been attempted. Instead, ostensible measures of microaggressions focus exclusively on the self-reported experiences of supposed victims (see Wong et al., 2014 for a review). Putting aside what such measures do capture, such work can say nothing at all about the "automatic, preconscious, or unconscious" nature of their supposed transgressions. Further, we are not aware of a single review of microaggressions that has even attempted to evaluate evidence regarding how "automatic, unconscious, or preconscious" microaggressions are (e.g., Sue et al., 2007; Williams, 2020, 2021; Wong et al., 2014). That is because there is no such evidence.

No evidence that microaggressions are caused by racism. One could imagine experimental studies that first identify some condition that increases racism, and then creates and manipulates that condition to evaluate whether White participants produce more microaggressions. Experimental attempts such as this appear nowhere in common reviews of microaggression research (e.g., Sue et al., 2007; Williams, 2020, 2021) except to point out that the area needs such work (Wong et al., 2014), a call that has not yet been answered.

One could also test for racism causing microaggressions in naturalistic studies. Although proving causality is much more difficult in such studies, it is not impossible (Rohrer, 2018). One could start with a well-validated measure of racism and if someone could produce a well-validated measure of microaggressions, one could administer them longitudinally, along with a slew of control variables to reduce third variable explanations. If racism at time 1 predicted microaggressions at time 2 (independently of other control variables, and also the measure of microaggressions at time 1) one would have at least preliminary evidence consistent with the idea that racism causes microaggressions. This has never even been attempted. Claims that microaggressions are caused by racism may or may not be true, but they are scientifically unjustified by the complete evidence of relevant data..

Further methodological and conceptual failures. Identification of what constitutes a microaggression has occurred mainly by researcher fiat. For example, Williams (2020, p. 5) argued that "microaggressions are, by nature, offensive in the sense that they are a form of racism." By nature according to whom? Presumably, by the anointed (Sowell, 1995). Earlier in the same paper, Williams (2020, p. 4), argued that figuring out whether something is a microaggression "is not based on the conscious intent of the offender or the perception of the target." This is a stunning statement. If it is based on neither intent of the offender, nor the perception of the target, determining if something is actually a microaggression would appear to be scientifically impossible. It can, however, be "determined" by fiat by the anointed.

As Cantu & Jussim (2021, p. 226) put it:

> "...researchers argue that listed microaggressive items have intrinsically embedded racist meanings, notwithstanding the non-racist intent of the speaker or the lack of malign interpretation by the recipient. As such, the legitimacy of lists of microaggressions depends on researchers being able to divine objectively racist meaning in facially innocuous acts that others cannot detect. And the propagative success of the CMC [current microaggression construct] has relied on the public believing that researchers are able to do just this.

Table 2 in Williams (2020) lists many statements as "microaggressions," including: "What is your nationality" and "how did you get so good at science." We agree that *it is possible* that situations exist wherein these statements would be insults, but that is a far cry from producing evidence that "socially conditioned racial biases and prejudices" (Williams, 2020, p. 6) causes them, either in general, or in any particular instance. It is *also possible* that in many, most, or even all instances in the real world, *no* such statements would be caused by racial biases (absent evidence, anything is possible) and they are, instead, caused by curiosity about someone's national origin or a belief that hard work usually pays off (respectively). No method for separating these alternatives doing so has yet been devised, let alone deployed, to produce scientific evidence that racial bias produces these types of statements.

In a startling confession, Williams (2020, p. 12) wrote:

> Lilienfeld (2017b) argued that there is no evidence that the commission of microaggressions is related to racial prejudice. Admittedly, those of us who study microaggressions have not felt a need to prove this because the connection between racism and microaggressions appears evident through our research and lived experiences.

Sowell's (1995) anointed may claim anything they like, but no one should pretend that this constitutes any type of scientific evidence.

Williams (2020), did, however, invoke one of her prior studies as providing the evidence that racism causes microaggressions. Kanter et al. (2017) correlated measures of racism with measures of microaggressions. Williams (2020, p. 12) described the study as providing "important empirical support for something that diversity researchers knew all along—microaggressive acts are rooted in racist beliefs" because correlations between measures of racism and some microaggressions were substantial (*r* = .4 or higher).

This sounds like strong evidence until one examines Kanter et al. (2017). First, it was a small-scale study (33 black and 118 white students) from college students in Kentucky. These numbers are so small and so unrepresentative of any population that the entire study should be viewed as little more than preliminary. Indeed, its title begins with "A Preliminary Report…" This is fine, but no one should be declaring anything definitive (such as that the study provides important empirical support for the assumption that microaggressions are rooted in racism) on the basis of a preliminary report. Along those lines, Kanter et al. (2017) has not been subjected to attempted replication by independent scientists. But even taking the study at face value (as having found what correlations between racism and some microaggression items), it fails to provide the evidence Williams (2020) claimed.

First, as every first year graduate student knows, one cannot infer a particular causal direction from a correlation, so the observed correlations provide no direct evidence about whether or not racism causes microaggressions. Second, only 14 of the 30 microaggressions correlated with racism at $p < .05$, meaning that 16 were statistically indistinguishable from zero. Five of them 14 had p-values below .05 but above .01, which are notorious for being unlikely to replicate (Benjamin et al., 2018), and, therefore, should not be treated as fact pending replication. Compounding these statistical issues, no method was used to account for the large number of statistical tests (conventional methods would involve requiring a lower p-value threshold, rendering even fewer of the items "significantly" correlated with racism).

This means that for over two-thirds of the supposed microaggressions, the study effectively showed no meaningful relationship, or, at best a suspect one, between racial prejudice and a white participant's likelihood of expressing it. Importantly, among the 21 items that failed to meaningfully correlate with racism (or did so at $p > .01$), most were the ambiguous items for which claims of overreach are most trenchant (items such as "You seem more intelligent than I thought"). Thus, this study provides little, if any, evidence that these behaviors constitute manifestations of racism.

Furthermore, a handful of items seem like blatantly racist statements on their face and were perceived by large majorities of the Black respondents (72% or more) as racist (such as "You are smart for your race"). Unsurprisingly, these items often had substantial correlations with racism (.3 or higher). When summed together, the set of items produced the substantial overall correlation with racism touted by Williams (2020) as a smoking gun of racism-infused microaggressions. But this is no smoking gun at all. The small number of blatantly racist statements that actually do correlate with measures of racism drives much of the scale's overall correlation with racism. The items least likely to be racist ride on the correlative coattails of the more blatant racist items, creating a false veneer of legitimacy to the entire set of questions.

If researchers in the area merely claimed that blatantly racist statements correlate with racism, we suspect that few would see this as anything but a banal validation of the racism measure. Instead, this method of smuggling in the weak and ambiguous items under the protective umbrella of the blatant items can be used to rhetorically "justify" the unjustified: that microaggressions are more pervasive and varied than they really are.

Or consider a scale described by Williams (2020, p. 13) as "Another important measure of microaggression frequency." A series of studies were conducted using some sophisticated statistical techniques (such as confirmatory factor analysis) to develop and validate a questionnaire assessing microaggressions (Nadal, 2011). Even if one accepts the questionnaire as completely valid, its main findings were a stunning rebuke of the claims by microaggression advocates as pervasive.

POC respondents were provided with supposed examples of subtle racism, such as "someone assumed I would have a lower education because of my race." They were then asked how frequently they had experienced this supposed microaggression in the prior six months. For most items, most respondents reported that they either had experienced the supposed microaggression in the past six months between zero and three times. This is hardly "pervasive."

Again, it goes downhill from there, because there are no reasons to take Nadal's (2011) studies as capturing actual microaggressions. First, it is entirely from the target's standpoint so we have no idea what the supposed perpetrators actually did or said. Second, retrospective memory is known to be imperfect (Schacter et al., 2023), so even the low estimates of microaggressions should not be presumed valid absent validation evidence.

Third, many of the items require mindreading, and, if we exclude the work on paranormal phenomena discussed previously, this is a power people do not have. For example, items ask things like, "Someone assumed that I would have a lower education because of my race" and "Someone avoided walking near me on the street because of my race." Questions like these require respondents to read others' minds or, at best, attribute beliefs and motives to others. Attributional processes are well-known to also be fraught with errors and biases (e.g, Gawronski, 2004).

In response to these sorts of criticisms (see also Cantu & Jussim, 2021; Lilienfeld, 2017), Williams et al. (2021) highlighted recent research purporting to find microaggressions to be experienced far more frequently than found by Nadal (2011). That is indeed what the authors of the study Williams et al. (2021) touted. Anderson et al. (2022, p. 303) claimed: "Our first major finding was that medical students frequently experience microaggressions."

Unfortunately, the authors' claims notwithstanding, they did not assess "microaggressions" as defined by the advocates. They assessed variations on "How often do you think someone has been mean to you?" Here are just two items: "People trivialize my ideas in classroom discussions" and "I am made to feel unwelcome in a group." There is nothing about race or racism here (or in their other questions). There is no assessment of unconsciousness. These types of experiences have probably happened to everyone.

In sum, then, much of the work on microaggressions has so far *failed to establish* that:

1. supposed microaggressions are caused by racism;
2. they are unconscious, even sometimes;
3. they are pervasive
4. except for the most blatantly racist items few are even perceived as racist
5. measures of the actions advocates label as microaggressions are perceived accurately when assessed exclusively among the potential targets of microaggressions.

Put differently, although one might be able to pull out some empirically justified claims from this work, it provides nothing whatsoever about microaggressions as commonly defined (Nadal, 2011; Sue et al., 2007; Williams, 2020, 2021) – as pervasively experienced subtle, generally unconscious, slights due to racism. The evidence may be highly persuasive to microaggression advocates – but it is far from perverse to believe that it has not established its main claims as scientific facts about microaggressions in the Gouldian sense.

Microaggressions as propaganda scholarship. Microaggression research shows many of the same earmarks of propaganda scholarship as does work on implicit bias, so we only briefly review them here. It involves repeated leaping to and subterranean importation of conclusions. It involves truth claims with no evidence. It has been used to justify interventions (such as college microaggression trainings and bias response systems) on a very thin scientific basis. And, like work on implicit bias, it has likely created a new generation of Sowell's (1995) anointed who believe they can "see" subtle manifestations of racism that the benighted plebians are too ignorant or bigoted themselves to notice.

Stereotype Threat and African American Standardized Test Performance

Definition and key ideas. Stereotype threat refers to performance debilitating effects produced by fear of confirming negative societal stereotypes about one's group (Steele & Aronson, 1995). In three studies, Steele & Aronson (1995) found that making some aspect of stereotypes of African American intellectual inferiority salient was enough to reduce African American students' standardized test scores.

This initial work inspired an immense body of empirical research, and readers are directed to reviews by its advocates if so interested (e.g., Schmader, 2010; Spencer et al., 2016; Walton et al., 2015). In this section, however, we do not consider the myriad of studies involving stereotype threat. We consider the fundamental question: "Are stereotype threat effects a scientific fact?"

Widespread academic misinformation about stereotype threat. The work was widely misinterpreted and misrepresented when it first came out as showing "but for stereotype threat, Black and White standardized test scores would be the same" (see reviews by Jussim et al., 2016; Sackett et al., 2004). Such a claim is false; no study has ever found this. Instead, the common use of analysis of covariance to control for initial Black/White differences in standardized test scores means that, post-stereotype threat induction, those differences were *not reduced* in the conditions in which the *covariate-adjusted* means were equal (as in Steele & Aronson, 1995). Prior initial differences were simply maintained.

To illustrate how this sort of covariate adjustment works, Jussim et al. (2016) performed an analysis of covariance testing whether the average daily high temperature in Nome, Alaska was any different that in Tampa Florida, controlling for temperatures on the prior day. Just as Steele & Aronson (1995) showed that "there were no Black/White differences in test scores, controlling for prior test scores in the no threat condition, Jussim et al. (2016) showed that there were no differences in the daily high temperatures of Tampa and Nome, controlling for prior temperatures..

Returning to the Steele & Aronson (1995) studies, they did find an effect of stereotype threat. The difference in *covariate-adjusted* mean obtained under threat occurred because those differences *increased* in the threat conditions (at least if one puts aside the many reasons to believe that analysis of covariance is completely inappropriate for experimental studies of stereotype threat, e.g., Wichert, 2005). Nonetheless, the false interpretation of Steele & Aronson (1995) as showing "but for stereotype threat, Black and White test scores would be the same" lives on in peer review and in textbooks. From 2005 to 2019 (i.e., after the Sackett et al., 2004 paper exposed that the "but for stereotype threat, Black and White test scores would be equal" was false), over 60% of peer reviewed journal articles and over 40% of introductory psychology texts examined addressing between group effects of stereotype threat *still* made inaccurate claims about it (Tomeh & Sackett, 2022). Thus, by virtue of promoting demonstrably false claims, this is *a lot* of propaganda scholarship.

Why even the more modest stereotype threat claim is not an established scientific fact. Even if the original claims were overblown, it would be possible that the more modest claim – that stereotype threat undermines the achievement test performance of African American test takers – rises to the level of scientific fact. Next, therefore, we explain why it does not.

There is a general pattern in published psychology wherein (Scheel et al., 2021):

1. The conventional literature produced results that supported the tested hypothesis 96% of the time.
2. Registered reports produced results supporting the tested hypothesis 44% of the time.

The key difference between a conventional publication and a registered report is that, with registered reports, a journal commits to publishing the study on the basis of the proposed research, regardless of how the results of the studies turn out. In contrast, the conventional literature has no such constraint and the use of "statistical significance" (typically, p < .05) constitutes a gatekeeper for entry into the published literature (e.g., Simmons et al., 2011). This raises the possibility that any literature on a particular psychological topic may constitute a distorted picture of reality (e.g., Ioannidis, 2005). Put differently, in the absence of registered reports replicating some effect, the study by Scheel et al. (2021) means that any particular researcher has justification for believing that even a literature with 96% published confirmation has less than a 50% chance of actually being true (in the sense of being replicable).

We know of no registered replication reports regarding stereotype threat and African American test taking. Therefore, it is reasonable to believe that registered reports would have less than a 50% chance of replicating the effect.

Further evidence in support of this sort of skepticism involves work on stereotype threat in the related area of women and performance in math. The core idea is the same – fear of confirming negative stereotypes about women's performance in math undermined women's actual performance in math (e.g., Spencer et al., 1999). This also inspired a vast amount of research, and interested readers should consult reviews by advocates (e.g., Schmader, 2010; Spencer et al., 2016).

There have been two registered reports testing stereotype threat effects among women. Despite large samples (590 and 2064, respectively), both failed to produce a statistically significant effect of stereotype threat (Finnigan & Corker, 2016; Flore et al., 2018). These findings are consistent with Scheel et al. (2021) showing that even a vast published literature teeming with statistically significant effects may not be readily replicable. We are not claiming there is no stereotype threat effect on women in math. Our claim is more modest: that these findings mean it is not perverse to withhold provisional consent with respect to the stereotype threat hypothesis, pending successful registered replications.

How does this apply to stereotype threat and African American test performance? The logic is inexorable:

1. There are no published registered replication reports of stereotype threat and African American test performance.
2. Scheel et al.'s (2021) study suggests that such reports would have less than a 50% chance of succeeding.
3. We know for a fact that the only two registered reports testing the stereotype threat effect for women and math failed to detect the effect, so that
4. It is not perverse to withhold provisional assent with respect to claims about stereotype threat and African American test performance.

Interestingly, there is a large scale "many labs" type registered replication project in progress right now, as we write this chapter (Forscher et al., 2020). When the project is completed and published we may have clearer information as to how replicable – i.e., how scientifically well-established – is the main stereotype threat phenomenon. Pending demonstrable replicability in the registered report format, advocates can point to hundreds of studies for support, but it is not perverse to withhold provisional assent. That is, there is no reason for anyone besides the advocates to consider stereotype threat an established scientific fact.

The Discrimination Paradox

Racial discrimination is a vast topic and this is not a comprehensive review of that literature. Instead, our focus is more limited. In the last few years, several empirical studies of racial discrimination have been published, *seeming* to produce wildly different results. The entire set of studies focus on *individual acts of discrimination* rather than "systemic racism," which we do not address in this chapter. The focus of our review is four papers including dozens of studies: 1. A meta-analysis of job discrimination (Quillian et al., 2017); and 2. Three large sample papers assessing discrimination in either the real world (Campbell & Brauer, 2021; Nødtvedt et al., 2021) or experimental laboratory (Peyton & Huber, 2021). The meta-analysis produced evidence of substantial levels of discrimination; the others of very low levels of discrimination. In contrast to the work reviewed so far on implicit bias, microaggressions, and stereotype threat, as we describe below, we consider all of this work sound and credible.

How, then, can it be that strong studies produce both evidence of substantial discrimination and evidence of minimal discrimination? We refer to this empirical pattern as *the discrimination paradox.* And in this section, we resolve the paradox by showing that the two seemingly very different findings regarding levels of discrimination are, in fact, completely compatible with one another.

Audit studies find substantial racial discrimination. Audit studies refer to a class of experimental studies, conducted in the real world, wherein targets who are otherwise identical (say, they have identical or equivalent resumes) differ on some demographic characteristic and apply for something (such as a job). It can be almost any demographic such characteristic, but herein we focus on those manipulating whether the job applicants are Black or White. The main outcome is whether Black or White applicants are treated similarly or differently (say, via call backs or interviews).

A review and meta-analysis found 21 audit studies of racial discrimination in hiring since 1989 and three additional ones going back to 1972 (Quillian et al., 2017). The studies included over 55,000 applications submitted for over 26,000 jobs. There were two headline findings:

1. On average, White applicants received 36% more callbacks than did Black applicants.
2. This difference did not decline between either 1972 or 1989 and 2015. Indeed, there was weak evidence that it had actually increased over that time.

Even audit studies are not beyond criticism. Most were conducted before registered reports or even pre-registration became common. Quillian et al. (2017) did include procedures to test for publication bias (they found some small evidence of it), but also used a variety of methods to mitigate it. The types of jobs studied are typically entry level (Reilly, 2021) so provide little information about whether, how much, or which direction discrimination occurs for mid-level or professional jobs. Similarly, Reilly (2021) reported that none of the audit studies conducted to that point had examined discrimination at the over 18% of minority-owned companies in America.

Whether any of these issues rise to the level of *dismissing or disbelieving* the results of the Quillian et al. (2017) meta-analysis is a matter of judgment. For our part, although we think Reilly's (2021) points about types of jobs are well-taken, we also think Quillian et al.'s (2017) results are credible, at least for White-owned businesses and entry-level jobs. We treat them as such throughout this section.

Recent studies showing very low levels of discrimination. One study found anti-Black discrimination 1.3 percent of the time, which is the same as saying they found no anti-Black discrimination the other 98.7 percent of the time (Peyton & Huber, 2021). In the study, they had over 700 people play the ultimatum game with either Black or White partners. This is a game often used in experimental studies. The first player proposes to the second how to divide some money. For example, the first player may be given a dollar to divide, and offers 30 cents to the second. If the second player accepts, then the first gets 70 cents and the second gets 30 cents. If the second rejects this division, neither gets anything.

Participants played the ultimatum game 25 times with either Black or White partners, so the total number of offers accepted or refused was over 18,000. Peyton and Huber (2021, p. 1830) described what constituted racial discrimination in their study: "Racial discrimination occurs when a white individual rejects an offer from a Black individual that would be accepted if offered by a white individual." This happened 1.3 percent of the time.

The participants in this study were Mechanical Turk workers, which is important because they are not a representative sample of Americans. Whether the 1.3 percent figure would generalize to "Americans" is unknowable from this study. Also, whereas the 98.7 percent nondiscrimination is very high, it was not a real-world context. Although this renders its implications for real-world discrimination unclear, the next two papers addressed discrimination in the real world.

Campbell and Brauer (2021) conducted five surveys, eight experiments and a meta-analysis examining discrimination at the University of Wisconsin-Madison. Although we focus exclusively on the seven experiments addressing racial discrimination (including discrimination against Muslims), results were similar for the study addressing discrimination against homosexuals. All studies examined naturally-occurring interactions, such as door-holding, asking directions, sitting next to a target on a bus, as students went about their business on campus.

The discrimination studies begin with Study 5, which found that 5% of students held a door for a White person but not for a Black person, a difference that was not statistically significant. Study 6a found that a White actor requesting directions received them 9% more often than an Asian actor and 6% more often than a Muslim actor, differences that were, again, not statistically significant. In Study 7a, a White actor received help 18% more often than did a Muslim actor (a result that did not quite reach statistical significance), but 20% less often than did an Asian actor (a result that was statistically significant). Study 8 found a Muslim actor was treated with more social distance on a bus 6% of the time, a result that did not reach statistical significance despite being tested in multiple ways. Studies 9a and 9b were job application studies. Study 9a found that a White applicant received 7% more responses than an Arab applicant; Study 9b found that a White applicant received 8% more responses than did a Black applicant (neither of these differences were statistically significant).

The meta-analysis that included only the Black and Muslim targets was statistically significant, indicating a small overall tendency to favor White targets. Simply averaging the differences for these groups produces an overall discrimination rate of about 8%. Of course, these studies were only conducted among college students at a single university, so their generalizability is unknown.

Nødtvedt et al. (2021) examined discrimination in the selection of Airbnb listings among a nationally representative sample of 801 Norwegians. The host was either identified as ethnically Norwegian or ethnically Somali. Overall, there was a 9.3 percent preference for the listing by the Norwegian ethnic. Of course, this study was conducted in Norway and whether its results generalize to anywhere else is an open empirical question.

Although these three publications (including 9 separate studies) each have unique limitations, they are also broadly consistent with a recent meta-analysis finding very low levels of racial disparities in "adjudication" (imprisonment, sentence length, etc) since 2005 (Ferguson & Smith, 2023) and a surprising lack of racial discrimination in studies assessing the predictive validity of the IAT (Jussim et al., in press).

Taking these findings altogether, we have the paradox. Quillian et al. (2017) in a high quality meta-analysis finds job discrimination at 36%; the recent studies reviewed in detail here (Campbell & Brauer, 2021; Nøtvedt et al., 2021; Peyton & Huber, 2021), along with many other studies (e.g., Ferguson & Smith, 2023; Jussim et al., in press) find discrimination at very low levels, typically single digits.

It might seem that something is wrong somewhere. Perhaps there are deep flaws in the studies finding little discrimination but not in the ones finding substantial discrimination (or vice versa) and we simply failed to uncover them. Perhaps the situations are too different to justify any comparison (even though two of the Campbell & Brauer, 2021 studies were audit studies, as in the Quillian et al., 2017, meta-analysis).

These are all possible. However, it is also possible that they are all strong and credible studies – but if that were true, we would have an apparent paradox of strong studies producing seemingly strikingly contrasting findings. In the next section, we show how the seemingly conflicting findings of the different highlighted studies are completely compatible; there is an apparent conflict, but no actual conflict.

Resolving the discrimination paradox. There is no single number for the amount of discrimination a group experiences. Discrimination varies in type (hate crimes, harassment, exclusion, etc.), and there are many different methods for assessing discrimination, which often yield different estimates. Still, to illustrate the discrimination paradox, we need to use actual numbers. We use the 36%, based on the Quillian et al. (2017) meta-analysis of racial discrimination in hiring as our starting point.

The key to resolving the discrimination paradox is understanding that discrimination can be assessed at two different levels of analysis. The 36% figure obtained by Qullian et al. (2017) is based on the differences in callbacks received by Black and White applicants. That is, it is a difference between the experiences of Black and White applicants. It is not the difference between the responses of companies to Black and White applicants. In contrast, the three papers finding single digit discrimination (Campbell & Brauer, 2021; Nøtvedt et al., 2021; Peyton & Huber, 2021) addressed acts by potential perpetrators of discrimination.

The importance of this difference can be readily seen with an example that starts with Quillian et al.'s (2017) figure of White applicants receiving 36% more callbacks than did Black applicants. First consider a simple hypothetical:

> There are 1000 applicants for a type of job. There are 500 Black and 500 White applicants with equivalent records. They receive a total of 236 callbacks, combined.
>
> 1. If there were no discrimination, Black and White applicants would receive identical numbers of callbacks, 118 in each case.
> 2. In this hypothetical where there are Quillian et al. (2017) levels of discrimination, White applicants receive 136 callbacks; Black applicants 100, thereby reflecting the 36% figure of Quillian et al. (2017).

3. Discriminatory acts have occurred 18/1000 times, or 1.8% of the time (118, which is egalitarian, plus 18 discriminatory callbacks, equals the 136 White callbacks).

This resolves the discrimination paradox because it shows how 36% discrimination from the target's standpoint results from acts of discrimination occurring only 1.8% of the time. There is no substantial conflict between the results of Quillian et al's (2017) meta-analysis, and those of the recent studies finding single digit levels of discrimination (Campbell & Brauer, 2021; Nøtvedt et al., 2021; Peyton & Huber, 2021).

Table 1 displays several alternative scenarios, all showing the same type of thing, whereby 36% of discrimination from the target's standpoint (or slightly more) results from acts of discrimination occurring in single digits. The first three rows purposely used equal numbers of Black and White applicants to make it easy to see the math underlying our resolution to the discrimination paradox. The last three rows use numbers that more closely approximate the Black and White population proportions in the U.S.

Table 1: Resolving the Discrimination Paradox

Black Applicants	White Applicants	Callbacks to Black Applicants	Callbacks to White Applicants	Quillian et al. (2017) Discrimination Rate	Acts of Discrimination Rate
500	500	200	272	36.00%	3.6%
500	500	100	136	36.00%	1.8%
500	500	50	68	36.00%	0.9%
200	800	75	408	36.00%	1.1%
200	800	38	207	36.18%	0.6%
200	800	8	44	37.50%	0.1%

The Quillian column always equals at least 36%. Only whole people can receive callbacks. In the last two hypothetical examples, the nearest round numbers for callbacks produced discrimination of slightly above 36%. These examples assume the Black and White applicants are equally qualified, as can be assured experimentally in audit studies.

One of the things notable from Table 1 is that, as the number of callbacks goes down, so does the proportion of acts of discrimination necessary to approximate their 36% rate. This can be seen in a simple example. Consider an example in which 500 applicants were Black and 500 were White, and there were only four callbacks. With no discrimination, 2 callbacks would go to Black applicants and 2 would go to White applicants. A single act of discrimination would mean that only one Black applicant would receive a callback whereas three White applicants would receive a callback. In this case, White applicants are 200% more likely to receive a callback (three versus one), even though there are only 1/1000 or 0.01% acts of discrimination.

Implications. One implication is that minimal levels of acts of discrimination can have a substantial impact on the targets of discrimination. There are long running debates about whether even small biases are important. Our resolution to the discrimination paradox strongly suggests that some small biases do indeed likely produce larger disparities than one might assume, e.g., if one merely knew that acts of discrimination only occur in the single digits. One important question, then, is what to do about this?

Although this is not a policy paper, the effectiveness of various types of diversity initiatives has been the subject of empirical investigation. Diversity training (Devine & Ash, 2022) and implicit bias training (Paluck et al., 2021) are generally ineffective. The current review may help explain why – because of a floor effect. If the research reviewed herein applies broadly, then acts of discrimination are rare. It is likely to be exceedingly difficult to significantly (both statistically and practically) reduce an already-rare event. The recent studies described here found acts of discrimination most typically in the single digits (Campbell & Brauer, 2021; Nødtvedt et al., 2021; Peyton & Huber, 2021). It also seems like a waste of time and effort if nearly all of those subjected to such trainings are already not engaging in discrimination.

To be sure, however, this only potentially explains why the evidence for the effectiveness of a variety of diversity trainings is so weak. It does not mean that the effort to reduce or eliminate racial discrimination is inherently misguided. It does suggest, however, that efforts probably should be directed elsewhere. It is possible that there is no one-size-fits-all effective intervention (such as diversity or bias trainings). Perhaps interventions need to be tailored to specific organizations and settings, by first identifying how much discrimination is occurring (if any) and in what contexts. We speculate that such targeted interventions are far more likely to be effective than the ubiquitous yet ineffective trainings.

Preferential selection can ensure any level of representation desired, but most people, including most members of the minority groups that it is supposed to benefit, opposed it (Graf, 2019; Horowitz, 2019). This may help explain why California, a state with only a minority of White voters, has twice voted to support referendums banning this type of affirmative action. Furthermore, at least for college admissions, the U.S. Supreme Court recently ruled that preferential selection based on race/ethnicity is illegal. Work on the "mismatch effect" has found that preferential selection was actually associated with fewer Black students passing the bar (Sander, 2004). This occurred because, when applicants are selected exclusively on merit, however imperfect indicators of merit may be, they are more likely to fit in well with other students at the school and with faculty expectations. In contrast, when students with records far below that of their classmates are admitted, failure and dropout is more likely.

In addition, there is abundant evidence that, when perceivers have and attend to a great deal of relevant individuating information, they overwhelmingly use that information, rather than make judgments based on demographic categories (Jussim et al., 2009). Thus, another strong contender for eliminating discrimination on the job or in admissions is to adopt practices that emphasize focusing on and evaluating merit (Abbot et al., 2023).

But how to do this? Another intervention for which there is good evidence for effectiveness is adding diversity managers to an organization (Dobbin & Kalev, 2016). These are administrators tasked with overseeing the embrace of diversity within an organization. Because they provide expertise, information, and accountability, the evidence is that having such managers generally succeeds in increasing organizational diversity.

Conclusion

In this chapter, we first pointed out that skepticism, rather than being an activity of science deniers, is justifiably elevated to a "norm of science" (Merton, 1942), because, to deserve a special place of credibility, scientific claims should be intensely vetted for limitations, failures, and alternative explanations. Only the claims that emerge largely unscathed by this process deserve to be taken seriously. In tandem with this, we presented a philosophical and conceptual review of when some claim should be considered a "scientific fact" and concluded that the answer was "only when it would be perverse to withhold provisional assent.

In this context, we then presented a skeptical review of research on implicit bias, microaggressions, stereotype threat, and discrimination. We concluded that it is most definitely not perverse to withhold provisional assent to most of the common claims about implicit bias, microaggressions, and stereotype threat.

In contrast, the work we reviewed on discrimination emerged largely unscathed. As we showed when we resolved the discrimination paradox, even seemingly contradictory findings are, in fact, completely consistent with one another. Thus, whereas this review indicated that deep skepticism remains justified regarding many of the most common psychological *explanations* for discrimination, it is not similarly skeptical of the ongoing *existence* of discrimination. Reducing discrimination is, however, far more likely when its sources are scientifically well-justified than when faulty understandings are based on wishful thinking, leaping to conclusions, and fiat-by-researcher.