American Psychologist

Adversarial Collaboration: An Undervalued Approach in Behavioral Science

Stephen J. Ceci, Cory J. Clark, Lee Jussim, and Wendy M. Williams

Online First Publication, August 15, 2024. https://dx.doi.org/10.1037/amp0001391

American Psychologist

ISSN: 0003-066X
https://doi.org/10.1037/amp0001391

Adversarial Collaboration: An Undervalued Approach in Behavioral Science

Stephen J. Ceci[1], Cory J. Clark[2], Lee Jussim[3], and Wendy M. Williams[1]
[1] Department of Psychology, Cornell University
[2] The Wharton School, School and the College of Arts and Sciences, University of Pennsylvania
[3] Department of Psychology, Rutgers University

Open Science initiatives such as preregistration, publicly available procedures and data, and power analyses have rightly been lauded for increasing the *reliability* of findings. However, a potentially equally important initiative—aimed at increasing the *validity* of science—has largely been ignored. Adversarial collaborations (ACs) refer to team science in which members are chosen to represent diverse (and even contradictory) perspectives and hypotheses, with or without a neutral team member to referee disputes. Here, we provide background about ACs and argue that they are effective, essential, and underutilized. We explain how and why ACs can enhance both the reliability and validity of science and why their benefit extends beyond the realm of team science to include venues such as fact-checking, wisdom of crowds, journal reviewing, and sequential editing. Improving scientific validity would increase the efficacy of policy and interventions stemming from behavioral science research, and over time, it could help salvage the reputation of our discipline because its products would be perceived as resulting from a serious, open-minded consideration of diverse views.

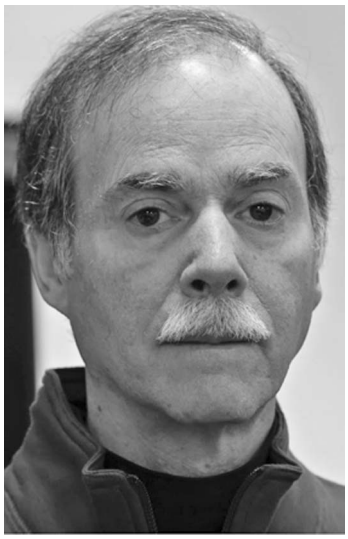
Public Significance Statement
It has been alleged by scholars that science proceeds "one funeral at a time": Disagreements are usually not resolved through the give-and-take of scientific studies but rather when one party to a scientific dispute retires or dies and the other side emerges as dominant. Adversarial collaborations are intended to avoid this troubling depiction by pairing opposing scholars to work together on a project designed to resolve their disagreements. The collaboration requires agreement on the framing of hypotheses, choice of procedures and dependent measures, and most importantly, a priori agreement about what evidence would lead each side to update their empirical beliefs. Although it can be difficult to persuade opponents to collaborate, adversarial collaborations hold the promise of advancing not only the replicability of science but also its validity.

Keywords: adversarial collaboration, validity, reliability, scientific bias, metascience

On September 12, 2022, we asked a nonscientific sample of 36 social scientists the following question:

> In your opinion, what is the most important advance made in the social sciences in the past several decades? We are not talking about specific research findings or movements (e.g., positive psychology or neuroscience), but rather innovations across the social sciences that have been broadly embraced by researchers in the field (e.g., preregistration, replication, public data repositories, ethnic/gender diversification of research samples, power analysis, adversarial collaboration, data-sharing, meta-analysis, IRB oversight, push for ecological validity). What would be your top picks?

Rick Hoyle served as action editor.
Stephen J. Ceci https://orcid.org/0000-0002-2835-9707
Stephen J. Ceci played a lead role in conceptualization and writing–original draft and an equal role in writing–review and editing. Cory J. Clark played an equal role in conceptualization, writing–original draft, and writing–review and editing. Lee Jussim played an equal role in visualization and writing–original draft. Wendy M. Williams played an equal role in conceptualization, writing–original draft, and writing–review and editing.
Correspondence concerning this article should be addressed to Stephen J. Ceci, Department of Psychology, Cornell University, Ithaca, NY 14853, United States. Email: sjc9@cornell.edu



Stephen J. Ceci

These 36 social scientists nominated 15 unique advances, which we then sent to a sample of 150 psychologists who spanned all areas and ranks of psychology, selected from a national canvass developed for a previous study. We asked them to indicate which of these 15 nominations they thought were the most important scientific advances or to add their own advance(s) if not on the list. Sixty-one of the 150 psychologists responded with usable data.

The most popular choices concerned Open Science (OS) initiatives to maximize data integrity, databases being accessible to assess replicability, and preregistered hypotheses and analyses to minimize *p*-hacking (various analytical and reporting tactics, such as file-drawering of nonconfirmatory results, that make supporting evidence look stronger than it is) and HARKing (hypothesizing after results are known; Simmons et al., 2011). Many expressed the view that OS, with the attendant access to online repositories of data, has been a game changer, resulting in better psychological science and greater reproducibility (although this assertion has yet to be demonstrated empirically; but see Clark, Connor, & Isch, 2023).

As empirical behavioral scientists, we agree that OS initiatives are indeed highly valuable for minimizing certain questionable research practices that have contributed to low replication rates in psychology and related disciplines (Camerer et al., 2018). Improving replication rates increases *reliability*: We can be more confident that specific methods will reproduce a specific set of results. However, OS practices are less useful for an equally (or perhaps even more) serious problem: a contradictory and systematically biased research corpus. This problem impacts not only the *reliability* of psychological science but also its *validity*. Plenty of scholarship in psychological science *is* replicable, but the conclusions directly contradict other replicable scholarship because different scholars frame their research questions differently, utilize different operationalizations of independent and dependent variables (DVs), rely on different procedures and selectively "file-drawer," and ignore different sets of results. To address these problems, we need adversarial collaborations (ACs). ACs require disagreeing scholars to (a) *work together* in pursuit of truth, (b) articulate their precise and empirically testable disagreements, (c) mutually agree upon rigorous and unbiased tests of their competing hypotheses, (d) commit to conditions of falsifiability or at least belief updating, and (e) pursue publication of the results regardless of their outcomes (Clark & Tetlock, 2023).

Advantages of AC

Only two of the 61 respondents in our 2022 survey nominated ACs as an important advance—this ACs despite being an option in the instructions and also being in the list of alternatives they were given. Perhaps the reason was because many respondents had never heard of ACs or why they are important; in contrast, OS initiatives such as preregistration have become well understood. Indeed, whereas OS practices rapidly rose to popularity, ACs are still unpopular (Clark et al., 2022a) even decades after prominent scholars began advocating for them (Bateman et al., 2005; Kahneman, 2003; Mellers et al., 2001). Our goal here is to familiarize readers with ACs and argue for their wider use to tackle one of psychology's most important challenges: A research corpus plagued with contradictory claims. (We also note in passing the potential value of ACs to other venues such as sequential editing, fact-checking, reviewing, editorializing, and calibrating the wisdom of crowds.)

Many areas of the social sciences can be characterized as "he-said-she-said" in the sense that opposing findings can coexist and receive repeated support over very long periods and sometimes over the course of entire careers. One team will publish a finding only to be countered by another team's contrary finding, and this back-and-forth can go on for decades. Or, worse, the two contrary lines of scholarship occur in parallel, but the advocates for each side rarely grapple with or even acknowledge the issues raised by the other side. Each group of scholars publishes article after article with only weak tests of contrary findings or acting as if the other side does not exist.

ACs are intended to avoid this result. They are teams of scientists with divergent positions who work collaboratively to design, conduct, and analyze studies that attempt to resolve their differences using mutually agreed-upon research procedures. Some ACs comprised only the sparring sides but others also contain a neutral referee agreed upon by the adversaries or appointed by an outside body such as a journal's editorial board.

As an illustration of the need for ACs, consider a few of the multitude of issues that ACs could examine, presented

Cory J. Clark

in Table 1.[1] Opposing teams have vied to support contradictory claims about such topics—and hundreds (if not thousands) of others—for decades. Occasionally, a resolution of contradictory claims emerges about a given topic, but usually, it is due to the retirement or death of a key proponent on one side rather than being the result of a deep resolution that scientists on both sides agree with. A team of economists who have studied this issue quantitatively by analyzing patterns of citations concluded that "scientific progress advances one funeral at a time" (Azoulay et al., 2019, p. 2889). A similar notion was advanced long ago by Nobel Laureate Max Planck: "A new scientific truth does not triumph by convincing its opponents and making them see the light, but rather because its opponents eventually die, and a new generation grows up that is familiar with it" (Planck, 1950, p. 33). Although there are notable exceptions to this pattern (see, e.g., Hull et al., 1978), the existence of countless long-running competing camps of scholars suggests that scholarly standoffs are commonplace. This, we think, is a rather embarrassing situation for researchers; we must die for science to relinquish our cherished theories. Surely, we can do better. ACs could be an important part of the solution.

The core problem is that researcher autonomy permits scientists to create bubbles of like-minded collaborators who can collaborate to "prove" a conclusion is true, employing procedures curated to support their preferred hypothesis. If a proponent has published evidence that has garnered praise, speaking invitations, contracts, appointments, and so forth, then this may inspire students to seek this person out as an advisor, and then other researchers who see the payoffs for studying that topic and reaching similar conclusions may follow, and consequently, these researchers will likely have increased opportunities to work with like-minded scientists. Their subsequent projects on this topic are then more likely to comprise a team of like-minded colleagues and students who will agree on the framing of hypotheses, the choice of paradigms and DVs, the definition of constructs, and (most importantly) what evidence they will accept as sufficient to falsify their past positions—often with little, if any, attempt to provide the type of severe tests that might falsify their hypotheses. Indeed, scholars often work to avoid falsifying their hypotheses. Because of the biases built into the way questions are framed and the methods used to address them (i.e., selection of the procedures that "work" the best), such work may be highly replicable, yet the conclusions based on it may be distorted or wrong altogether (see Jussim et al., 2016, for numerous examples of findings that are reliable but not valid). Research is seldom framed around what could falsify the theory. And because researchers have the freedom to frame research questions and focus on variables and methods amenable to their preferred outcome, their work may overstate the power of the phenomenon. This is one reason Ioannidis concluded that science is not necessarily self-correcting and that even when true associations are discovered, their strength is often inflated (Ioannidis, 2012; Jussim et al., 2023). ACs can protect against such biases.

Divergent teams, we speculate, are far more likely to use *strong inference* (Mayo, 2018) and *severe testing* (Lakens, 2024; Mayo, 2018; Platt, 1964). Strong inference refers to setting up a study to test mutually exclusive predictions from competing theories, rather than seeking only confirmatory evidence and dismissing null results as studies that simply "did not work." The lower the Type I and Type II error rates, the more severe is the test (Lakens, 2024). Although no single study in psychology can conclusively refute any particular theory, by comparing competing theories and making different predictions, psychological science can determine which theory has the better track record. Severe testing is the related idea that a claim is only accepted after it passes tests specifically designed to find flaws in that idea if they exist. Opponents in ACs will be highly motivated to severely test *the other side's* predictions, so we can have more confidence in whichever predictions are ultimately confirmed.

The journal *Psychological Science in the Public Interest* (*PSPI*) was created by the Association for Psychological Science to address the problem of long-term controversies by commissioning AC teams of scholars with histories of publishing opposing claims. After vetting by the *PSPI* editorial board, these scholars are asked to collaborate in an attempt to resolve their past and continuing disagreements. For example, over a 30-year period, there have been repeated contradictory claims about the relationship between eyewitness confidence and accuracy, with numerous findings published on both sides of the relationship. *PSPI* commissioned John Wixted and Gary

[1] Note that because the authors of the present article are from the United States, this list likely overrepresents controversies in the United States and similar Western cultures.

Lee Jussim

Wells—two of the chief proponents of the contradictory positions—to form an AC. They spent a year hashing out their contradictory claims and produced a synthesis that, while it did not reconcile every aspect of their prior disagreement, was a major improvement over decades of partisan reports that were written by opposing homogeneous teams (Wixted & Wells, 2017). The same can be said of other ACs (e.g., Löckenhoff et al., 2014; Mellers et al., 2001). As yet another example, Halpern and her divergent team of colleagues spent nearly 2 years debating what the data did and did not prove about sex differences in mathematics and science (Halpern et al., 2007). Although very valuable, these ACs strove for resolutions of scientific debates by argumentation and reanalyses rather than through the implementation of new methods, testable and falsifiable hypotheses, and the collection of new data. Other ACs led to the design and conduct of critical experiments (e.g., Kekecs et al., 2023). *PSPI* is a rarity among journals in commissioning adversarial scholars to conduct joint reviews, but it only publishes three articles per year. This is a drop in a vast bucket.

Table 1 lists 60 controversies, many of which regard dozens or hundreds of smaller debates (e.g., "Do men and women have different psychological characteristics? If so, where do they come from?" "How much variance do genes explain in [fill in outcome here]?" "Does [fill in effect name here] fail to replicate? If so, why? And does psychology have a replication crisis?"), many of which involve many thousands of scholars (e.g., a Google Scholar search for "replication crisis" returned 640,000 hits). Although we cannot know for certain, we would estimate that the number of scholars who have published an article that contradicts another article is well into the millions. For example, Google Scholar searches for the words "overestimated" and "underestimated," which return many articles claiming that other claims have been overestimated or underestimated, return over 2 million results combined. Yet, ACs remain rare. As of writing this sentence, only 2,250 academic articles have used the word "adversarial collaboration"—and many of these articles, like the present one, are *about* them, not one of them.

Bias Among Team Members May Be Unconscious

The bias that results from the typical homogeneity among collaborators need not be conscious, as Kahneman (2012) has argued, because the critical issues are seldom explicitly considered due to the like-mindedness of members of the team. Commenting on three ACs in which he participated, Kahneman (2012) argued:

> I believe that we theorists are not fully aware of the extent to which the experiments we plan and carry out are biased to favor our own theoretical point of view. I'm not alluding to a file drawer problem, to people hiding research that they don't like. The bias enters at the design stage. When you consider possible experiments, you apply your intuition to select those that are likely to support your view.

Often, based on prior findings and familiarity with competing studies, a principal investigator of a traditional, homogeneous collaboration may pose specific hypotheses, DVs, and analytic methods and operationalize core definitions that unwittingly tilt the odds toward a preferred outcome. As noted, such leaders may have a professional stake in the outcome because they built their reputations on past arguments they published, and they were hired, remunerated, given accolades and speaking invitations, and awarded publishing contracts (Clark et al., 2022a). When all team members expect or even desire a similar outcome, it can lead to designs and DVs that increase the likelihood that the outcome will confirm their hypothesis, and as Kahneman (2012) argued, this may be an implicit process.

Importantly, bias resulting from team viewpoint bubbles can occur regardless of whether the most common and well-known OS practices are employed—even if hypotheses, methods, power analyses, and statistical procedures are preregistered, and the raw data are made publicly accessible. This is because such initiatives place no constraints on the selection of methodological procedures: In preregistering a study, there is no requirement that a team includes someone to serve as a devil's advocate, to strive to falsify rather than confirm; no one is chosen to propose alternative hypotheses, DVs, operationalizations of core constructs, and methodologies (e.g., whether multilevel modeling should be employed). No one is necessarily tasked with implementing severe tests. Team members are not selected to disagree about what findings ought to be seen as disconfirming evidence for their theory. And null results can still be file-drawered.

One organized attempt to include viewpoint diversity in a research project is Lakens' (2020) concept of "red team" members who act as "a designated 'devil's advocate' charged

Wendy M. Williams

to find holes and errors in ongoing work and to challenge dominant assumptions, … teams of scientists should engage with red teams at each phase of a research project and incorporate their criticism." Lakens proposed this concept to thwart the dissemination of flawed findings related to the pandemic, but he points out that the concept has currency in other domains, such as cybersecurity where external members are tasked with finding flaws in security code. (Lakens did not intend for red teams to be limited to the pandemic; the word pandemic was added to the title by an editor at *Nature* because of its timeliness.) Indeed, Lakens applied the red team concept in his own recent meta-analysis examining audit studies of hiring discrimination against women including researchers who have staked out a wide variety of positions on gender discrimination and found that there has been none for many years (Schaerer et al., 2023).

A related concept is that of appointing a contrarian member of a research team. In his advice for future center directors, Gigerenzer singles out the important role that a contrarian played on his team at the Max Planck Society:

> Every research group can benefit from (at least) one contrarian, that is, a person who dares to question the group's and the director's wisdom, plays devil's advocate, insists on evidence, and questions what others take for granted. Such a person is sometimes frustrating but actually provides a great service by protecting the group from falling prey to groupthink. For that reason, when selecting new group members, we preferred those who found some fault or disagreement with our research findings … rather than those who politely praised our research. (Gigerenzer, 2022)

A nascent effort, the Adversarial Collaboration Project at the University of Pennsylvania, aims precisely to encourage and support ACs. The project members are currently carrying out multiple ACs and have successfully collaborated with a few academic journals, including *Advances in Methods and Practices in Psychological Science*, *Journal of Experimental Social Psychology*, and *Theory and Society* to encourage AC research. However, these efforts are small and new, and although they have successfully persuaded many scholars to participate in the projects, they have pursued and coordinated, they have had to work unusually hard to regrant funds for other scholars to launch their own ACs. And invitations for ACs are often declined (e.g., Costello et al., 2022; Tetlock & Mitchell, 2009).

Red team members, contrarians, devil's advocates, and their equivalents are a rarity in psychological science, although occasionally they are employed. Of course, reviewers and editors can play the role of devil's advocate, but for the reasons mentioned below, they may not be the best individuals to provide critical perspectives, and their role is at best reactive rather than proactive. Reviewers and editors are not involved in the design stage when biases are most likely to manifest and may not be sufficiently motivated to oppose the team's scientific or sociopolitical orientation, which can be particularly relevant when the research is sociopolitical in nature, a point elaborated below.

Research on Topics With Sociopolitical Implications

In addition to the advantage of resolving long-standing theoretical or empirical disputes, ACs can result in a second outcome that can enhance both the validity and credibility of scientific conclusions. Here, we refer to the representation of evidence that has sociopolitical consequences, rather than the purely theoretical disputes of the type Kahneman (2012) referred to above. Of course, work with sociopolitical consequences *is also theoretical and substantive*. Theoretical, substantive scholarship *with sociopolitical implications* is merely a subset of theoretical, substantive scholarship more generally. It is, however, useful to focus on this subset of work because it often has a direct influence on wider society and because theoretical and substantive research imbued with sociopolitical issues may be among the most highly contentious—and, as such, are a prime area for improvement via ACs. As Tetlock and Mitchell (2009) noted, ACs are most needed (and least feasible) in domains in which "the scientific community … is fractured into opposing camps that engage in ad hominem posturing and have intimate ties to political actors who see any concession as weakness" (p. 31). Nonetheless, many of our recommendations here apply as well to nonpolitical scholarship as to research with sociopolitical aspects.

Take, for example, the topic of whether there is evidence of gender bias in tenure-track hiring. The literature is rife with demonstrations of gender bias by ideologically similar teams, often published in the premier journals, and by prestigious professional organizations. Consider a few such claims regarding gender bias in tenure-track hiring presented in Table 2.

Table 1
Examples of Debates That Could Benefit From Adversarial Collaboration

Example
1. Is there gender bias in STEM fields? And if so, in favor of men or women?
2. How reliable is children's testimony?
3. Can grit be cultivated?
4. What are the short- and long-term consequences of COVID-19 vaccines?
5. What are the consequences of mass immigration of undocumented people?
6. Is perceived harm a fundamental component of all moral judgment?
7. Do violent video games increase violence?
8. When, if ever, can implicit racial attitudes explain discrimination?
9. Is human reasoning designed to pursue truth first and foremost?
10. Are stereotypes a self-fulfilling prophecy and/or a reflection of empirical reality?
11. Does racial bias among police explain the disproportionate shooting of minority civilians?
12. Is the political right more prone to motivated and inflexible thinking than the political left?
13. What do intelligence tests measure?
14. Do everyday people think scientific determinism is compatible with free will?
15. Does religion promote prosocial behavior?
16. Are the social sciences politically biased?
17. Do men and women have different psychological characteristics? If so, where do they come from?
18. Does ovulation influence female mating behavior?
19. What is the relationship between biological sex and gender?
20. Are there average psychological differences between ethnic groups? If so, what causes them?
21. Is the mind modular?
22. Are police shootings racially biased?
23. In which contexts, if any, does stereotype threat undermine performance?
24. What are the pros and cons of adoptees' rights?
25. Is smartphone use by children emotionally damaging?
26. Are there sex differences in intelligence?
27. What are the costs and benefits of gender-affirming surgery and care?
28. Does a Universal Basic Income disincentivize work?
29. What causes fake news acceptance?
30. How much variance do genes explain in (fill in the outcome here)?
31. Are political rightists more authoritarian than leftists?
32. How easily implanted are false memories?
33. What best explains inequality? And which proposed causes do not?
34. Is the Dunning–Kruger effect real?
35. Is the IAT reliable and valid?
36. What are the positive and negative consequences of electric cars?
37. What are the causes and consequences of microaggressions?
38. Which psychotherapies are validated and for which problems?
39. Is exercise a better treatment for depression than medication?
40. How much can personality change? And what can change it?
41. Do some psychotherapeutic interventions cause more harm than good?
42. Is social media harmful to children?
43. What is the size and cause of the gender pay gap?
44. Why do people attribute more intentionality to harmful side effects than helpful ones?
45. What, if anything, does mindfulness improve?
46. Does attachment to parents influence attachment to romantic partners?
47. Does contact with the criminal justice system increase or decrease recidivism?
48. When, if ever, do mindsets matter?
49. How accurate is eyewitness identification?
50. Is transgender identity sometimes the product of social influence?
51. How much does parenting matter?
52. Which human group differences exist? And which are the product of evolution? And why?
53. What best explains altruism?
54. When, if ever, does intellectual humility have negative consequences?
55. Which environmental interventions improve cognitive performance?
56. What are the costs and benefits of various diets?
57. What are the costs and benefits of polyamory?
58. Which factors do and do not explain declining fertility rates?
59. Why do men dominate leadership positions?
60. Does (fill in effect name here) fail to replicate? If so, why? And does psychology have a replication crisis?

Note. STEM = science, technology, engineering and mathematics; IAT = implicit association test.

Table 2
Claims Regarding Gender Bias in Tenure-Track Hiring

Claims in the elite science media regarding bias in hiring
"Researchers in recent years have found that women are less likely than men to be hired and promoted, and face greater barriers to getting their work published." (Casselman, 2021, *New York Times*, Section B, p. 1)
"Women … are penalized in hiring decisions when compared with equally qualified men." (Fortunato et al., 2018, p. eaao0185, *Science*)
"When fictitious or real people are presented as women in randomized experiments, they receive lower ratings of competence from scientists." (Witteman et al., 2019, *The Lancet*, p. 531)
"Implicit bias is pervasive. Men are preferred to women even if they have the same accomplishments. Psychologists have shown this by testing scientists' responses to fictitious curriculum vitaes that are identical other than coming from 'John' or 'Jennifer.'" (Witze, 2020, *Nature*, p. 583)
"Considerable research has shown. … Evaluation criteria contain arbitrary and subjective components that disadvantage women faculty." (National Academy of Sciences, 2007, pp. 4–5)
"Research has pointed to bias in peer review and hiring … a female applicant had to … publish at least three more papers in a prestigious science journal or an additional 20 papers in lesser-known specialty journals to be judged as productive as a male applicant." (Hill et al., 2010, p. 24)
"Even after earning STEM degrees, women are less likely to be hired into STEM jobs compared with equally qualified men." (Cech & Blair-Loy, 2019, *PNAS*)
"Bias, discrimination, and harassment are major drivers of the underrepresentation of women [in STEM]." (National Academies of Sciences, Engineering, and Medicine, 2020, p. 5)
"Every major criterion on which scientists are evaluated … has been shown to be biased in favour of (white) men." (Urry, 2015, *Nature*, p. 472)
"Studies have shown that people rate men's competence more highly than women's when assessing identical job applications." (*Nature* Editorial, March 7, 2024)

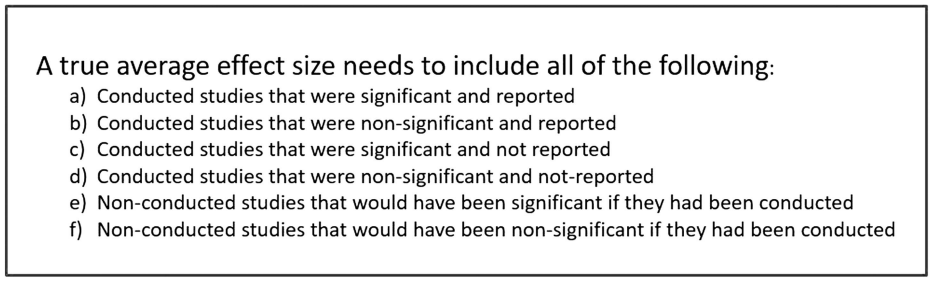
Note. STEM = science, technology, engineering and mathematics; PNAS = Proceedings of the National Academy of Sciences.

One might conclude from the myriad assertions of gender bias in tenure-track hiring that it is settled science. After all, highly respected scientific teams commissioned by organizations like the National Academy of Sciences and disseminated in top journals like *Nature* and *Science*, as well as respected legacy media such as the *New York Times*, have repeatedly ratified this assertion, and occasionally, a diverse team has attempted to produce a consensus report (Gruber et al., 2021). A possible approach to testing this claim is to conduct a meta-analysis of tenure-track hiring studies to determine whether the view of gender bias in tenure-track hiring is supported. However, meta-analysis can only produce effects that are based on what has been reported; it cannot fathom findings from studies that were never designed, conducted, or published.

As originally Hedges (1984) and subsequently many others (e.g., Nelson, 2018) have argued, estimating the mean meta-analytic effect requires information that is unavailable, particularly (e) and (f) in Figure 1 below, and for reasons mentioned below, these lacunae may not be randomly distributed. Because a true effect size is an effect in a defined population, (a), (b), (c), and (d) below are all part of the true effect size, whereas there is no defined population of "nonconducted studies" that can be formalized, and therefore, there is no "true effect size" for (e) and (f). This is an example of the difference between reliability and validity: Even if nonconducted studies might be more valid, they are not part of the "true average effect size," and they cannot be used to compute a "true" average across values from a nondefined reference class. Thus, the performed studies might not be valid, and more valid studies may not be performed, which emphasizes the need to distinguish this from biased effect sizes due to publication bias.

Beyond these missing components, there is good reason to believe that systematic publication biases impact all components. For example, it is likely that hypotheses that

Figure 1
Required Information to Estimate a True Average Effect Size



challenge the perspective that there is no gender bias in tenure-track hiring are less likely to be posed, tested, and published due to the extreme asymmetry in sociopolitical leanings among researchers in the academy (Duarte et al., 2015). A reviewer's sociopolitical orientation predicts their willingness to accept articles for publication, selection of participants for symposia, ratings of grant proposals, whom they would prefer to hire on tenure track (Honeycutt & Freberg, 2017; Inbar & Lammers, 2012), and which research conclusions they would discourage peers from pursuing (Clark, Fjeldmark, et al., 2024). Thus, the literature that gets meta-analyzed can be lopsided and disproportionately framed by those who are sociopolitically progressive.

The above list of quotes alleging gender bias in tenure-track is a concrete manifestation of this point because recently an AC team comprised authors who had previously published contrary findings has shown that the above claims are often incorrect or lacking needed qualifications (Ceci et al., 2023). Yet, it took decades to correct the erroneous record through the efforts of an AC comprised members on opposite sides of the issue who concluded that claims of pervasive gender bias in the tenure-track academy were not in fact consistent with the totality of evidence; while they found evidence of bias against women in salary and student ratings, they found no gender gaps in grant success, journal acceptances, and letter of recommendation. And they found bias in favor of women in tenure-track hiring. Laypeople may have some general understanding of these processes, which may help explain why they ascribe lower credibility to scholarship, the more politically asymmetric they believe are the constituents of academia (Buss & von Hippel, 2018). ACs can serve as a tonic against such skepticism because partisans on each side of the debate may be assuaged by the presence of someone from their side being author of the AC report.

ACs can help address the biases introduced into the scientific literature because of a lack of political diversity. Academics on the political right might be in a distinct minority, but academics who are willing to challenge particular prevailing dogmas are comparably plentiful. One primary way scholars earn status in academia is by forwarding seemingly novel findings that challenge prevailing theories and perspectives. It is thus easier to find a scholar who challenges one left-leaning finding than a scholar who would challenge all left-leaning ideas or findings.

ACs are one way to avoid the limited progress that results from an absence of viewpoint diversity among team members. For example, the entire Winter 2024 issue of *Daedalus*, the journal of the *American Academy of Arts and Sciences*, was devoted to the IAT, but none of its 19 contributions were written by an avid critic of the IAT (Liu & Jones, 2024). It is possible, even likely, that had an AC has been commissioned to study the IAT, the picture that emerged may have been less salutary. ACs can increase the likelihood that alternative explanations will be examined using a variety of methods and statistical tests, thereby elevating the robustness of whatever findings are ultimately actually obtained and reported.

ACs can further help science become more diverse by creating a *demand* for diverse perspectives. This is particularly important when the topic carries sociopolitical implications such as studies of affirmative action, gender bias, the efficacy of vaccines, abortion, diversity, equity, and inclusion attestations, implicit racism, and immigration. However, ACs—or variants of ACs such as adversarial teams of fact-checkers (Ceci & Williams, 2020), sequential editors to *Wikipedia*, contexts in which the wisdom of the crowd is relevant, and even editorializing—are all improved by having divergent thinkers challenging each other's interpretation (see Shi et al., 2019, for empirical demonstrations of the positive effect of ideological diversity of team members' sequential edits to *Wikipedia*'s political, social issues, and science articles).

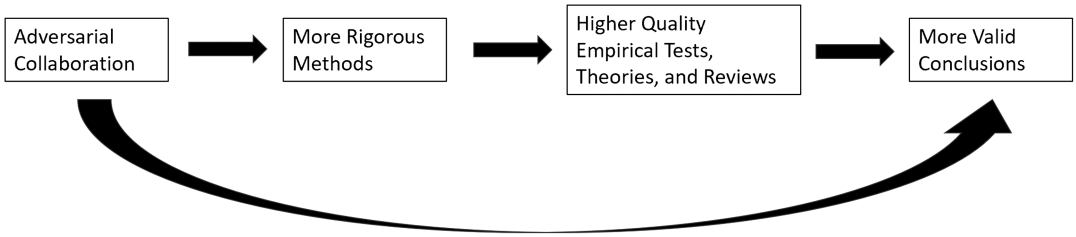
ACs may also help stem the decline of academic credibility, at least among the U.S. lay public. As mentioned above, trust in science has been declining among nonliberals for a long time (Gauchat, 2012) and, more recently, across the political spectrum (Brenan, 2023). The more Americans across the political spectrum become aware of the political skew of academia, the less credibility they ascribe to its claims (Marietta & Barker, 2019). And recent research suggests that when political values appear to influence scientific journals' and institutions' work, the public loses trust in those institutions and becomes less willing to defer to scientists' expertise (Clark, Isch, et al., 2023; Zhang, 2023).

Given the role of public support for research via federal grant funding and state funding of state universities, it would seem imperative to take steps to recover some of that lost trust. ACs could help stem the declining tide of credibility among the lay public and perhaps among scientists themselves. When a successful AC occurs on potentially politicized topics, scientists and the public may be assuaged by the belief that there is someone standing by the validity of the results whom they can identify with.

A Theoretical Model for Improving Validity Through ACs

Figure 2 is a sketch of how ACs can lead to more valid psychological science, one which suggests empirically testable predictions. It implies that ACs can improve validity via two routes. The first route is indirect, via the improved methodological rigor that is expected to result from negotiating differences with skeptical team members. Such negotiations are expected to produce more severe tests (Mayo, 2018) of alternative hypotheses (Platt, 1964) than the tests that are typically produced by teams of like-minded collaborators. This model also indicates that the more proximal causes of validity are not the composition of the research team but the rigor of the methods and quality of the empirical tests. Figure 2

Figure 2
A Model of How Adversarial Collaborations Can Improve Rigor and Validity



also specifies a direct route from AC to more valid conclusions. Our analysis predicts that ACs will be more likely to limit confirmation biases (Gauchat, 2012; Shi et al., 2019) and leap to unjustified and excessively broad and extravagant conclusions.

Vazire et al. (2022) outlined four key types of validity (Construct, Internal, External, and Statistical Inference) to consider for improving psychological science. ACs help address all of them. Construct validity refers to how well research claims map onto what was actually measured in a given study. For example, scholars have made claims about associations between political conservatism and racism using measures of "modern racism" (e.g., Kinder & Sears, 1981). However, measures of modern racism confound attitudes toward Black people with conservative values such as hard work and meritocracy (Sniderman & Tetlock, 1986; Wright et al., 2021). Consequently, this represents an invalid construct for making claims regarding associations between political conservatism and racism. Indeed, this measure better predicts political bias than racial bias, and those who score "low" on the metric show the most racial bias (and in favor of Black over White people; Wright et al., 2021). ACs would not permit such glaring flaws with construct validity.

As one example of ACs raising the bar for construct validity, consider recent research testing the hypothesis that political conservatives are more cognitively rigid than political liberals. An AC by Bowes et al. (2024) surveyed 44 constructs of "rigidity" and deemed them all flawed save for one: belief updating in response to evidence. In traditional collaborations, scholars often choose the measure with the best chance of confirming their hypothesis and hope reviewers do not notice, do not mind, or share their desired hypothesis. In ACs, flawed metrics are vetoed by one side or the other and only those deemed valid by both opposing sides remain.

Internal validity refers to the validity of causal inferences (Vazire et al., 2022). Many ongoing empirical debates in the social sciences specifically regard the existence or directions of causation (such as whether stereotypes cause differences or differences cause stereotypes, natural differences or social expectations cause gender differences, parenting or parental genes cause various child outcomes, to name a few). Here too, ACs would increase the likelihood of causal explanations being held to higher standards of evidence and would limit the inclusion of unwarranted causal explanations that ignore alternatives.

ACs also help improve external validity—the extent to which inferences drawn by the research extend or generalize beyond the sample, operationalizations, and procedures used in the research (Vazire et al., 2022). ACs limit scholars' freedom to make unwarranted generalizations (or inflated estimates of importance) regarding the significance of their findings in the real world (Clark et al., 2022a). Indeed, one AC specifically involved concerns about exaggerated claims (Martel et al., 2024). In this case, both sides ultimately agreed that the effect sizes for accuracy prompts improving the quality of news sharing decisions on social media are small and that the claim that these prompts did not work as well for Republicans/conservatives was only true in particular contexts (Martel et al., 2024).

Last, ACs bolster and build upon many of the recent scientific reforms for improving the validity of statistical inference. Consider some examples. ACs strongly incentivize the use of preregistration and registered reports (RRs) that lock adversaries into a contract regarding methods, analyses, competing hypotheses, and interpretations of the data, thereby limiting or eliminating the ability of the adversaries to enlist post hoc auxiliary assumptions to maintain their views. This is why preregistrations and RRs are both recommended for ACs (Clark et al., 2022b; Vlasceanu et al., 2022) and are frequently used in ACs (e.g., Bowes et al., 2024; Breznau et al., 2023; Clark, Pennycook, et al., 2024; Connor et al., 2024; Cowan et al., 2020; Martel et al., 2024; Matzke et al., 2015; Melloni et al., 2023; Stern & Crawford, 2021; Tasimi & Friedman, 2024; Van Dessel et al., 2017). ACs support a more expansive and balanced set of possible methodological procedures in integrative experimental designs, which specify and test numerous experimental approaches to a research question in consideration of multiple relevant theories (Almaatouq et al., 2024; Clark, Isch, et al., 2024). ACs encourage the use of multiverses (Steegen et al., 2016), which specify and test all (or at least many) valid statistical models to make sure conclusions do not exaggerate claims on the basis of just one (often nonrandom) statistical model (e.g., Breznau et al., 2023; Martel et al., 2024). And ACs reduce or eliminate

scholars' ability to file-drawer studies that do not support desired conclusions (Clark et al., 2022a), thus improving both the accuracy of original articles as well as meta-analyses of related phenomena. Additionally, ACs can reduce risks of motivated statistical errors that systematically support authors' hypotheses (Wicherts et al., 2011), either because neutral third parties conduct the analyses or because adversaries share data analysis responsibility so that both parties trust the outcome and are mutually accountable for the results.

ACs are predicted to be more likely to acknowledge the type of methodological and empirical limitations that can create a great deal of uncertainty about how to interpret findings from psychological research. When there should be a great deal of uncertainty about the justified conclusions from a line of research, teams that acknowledge that uncertainty reach more valid conclusions than do teams that erroneously report they have clear and compelling evidence for their conclusions. This can be seen in massive experiments involving 4,000 subjects in each of 80 different bipartisan communities sequentially rate the veracity of information; in such cases, crowd intelligence backfires due to initial false information provided by ideologically segregated communities is subsequently rated as accurate by less ideologically segregated ones (Stein et al., 2023). Even the process of adversaries articulating their competing perspectives with precision can often reveal that disagreements are much smaller than originally thought (likely because scholars are incentivized to frame their own findings as maximally novel and different from earlier ideas; Clark et al., 2022a). This facilitates the presentation of disagreements and findings with precise and accurate terminology.

Empirically testing and developing the above sketch into a formal model will probably require organizing many teams of researchers into adversarial versus nonadversarial teams. This will be difficult, but over the past 10 years, metascience advocates have organized very large numbers of independent teams of researchers to study replicability (Open Science Collaboration, 2015), variability in data analysis choices when testing the same hypothesis with the same data (Breznau et al., 2022), and the inferences researchers make on the basis of analysis (Starns et al., 2019). And recent researchers have assigned 4,000 people to one of 80 ideologically partisan conditions to assess the effect of ideological heterogeneity on the wisdom of the crowd (Stein et al., 2023). Empirically testing the effectiveness of ACs is clearly possible and desirable.

In principle, comparable efforts could organize many teams into ACs versus like-minded teams to study theoretically or politically controversial topics or the detection of misinformation. Such studies could then be compared along any dimensions related to rigor, quality, and validity, including but not necessarily limited to sample sizes, use of preregistration, open materials and data, use of overstatements (e.g., inferring cause from correlation), the success of any resulting applications or interventions, and contributions to the generation of later confirmed hypotheses. The quality of the studies produced by ACs versus studies produced by like-minded teams could also be evaluated using forensic methods, such as *p*-curves (Simonsohn et al., 2014).

The Platinum Standard for Resolving Controversies?

On Synergies Between AC and Varieties of Preregistration

Preregistration refers to the practice of creating a written, publicly available document that describes, for a particular study or set of studies, the hypotheses, methods, and planned, a priori analyses, and how results will be interpreted with respect to (dis)confirmation of these hypotheses. Preregistration has emerged from the science-reform movement (Nosek et al., 2018) as a critical practice capable of limiting *p*-hacking (Simmons et al., 2011), cherry-picking results based on the "garden of forking paths" (Gelman & Loken, 2013), and for clearly distinguishing a priori hypothesis testing from exploratory analyses. A cursory review of psychology journals shows that many studies published in the past few years have been preregistered.

Two variations on preregistration are RRs and Registered Replication Reports (RRRs). In RRs, authors submit, in essence, a preregistration to a journal or review platform (such as the Peer Community in RRs) for consideration. In this case, the journal has the opportunity to offer an *in-principle acceptance*, meaning that, if the authors conduct the study and interpret it as described, the journal will most likely publish the article no matter how the results turn out. RRs address the problem of publication bias because the journal more or less commits to publishing the research "no matter how it turns out, that is, whether the results are significant or not, and no matter whose theoretical or political ox is gored." RRRs are essentially the same as RRs, except that they are proposed replication studies. RRRs have proven invaluable for assessing the credibility of published conclusions in psychological science postreplication crisis (Simons et al., 2014).

Nonetheless, neither RRs nor RRRs *inherently* improve the validity and credibility of psychological science in the way that is possible with ACs. They provide some insurance against HARKing, *p*-hacking, forking paths, and publication bias. This is a major set of accomplishments. However, they are not designed to solve the same set of problems that ACs target. ACs target the following: (a) theoretical or political biases built into the topics studied, the specific hypotheses tested, and the very operationalizations and methods used to assess them; (b) biases in interpretations of results; (c) acknowledging and then testing predictions derived from alternative, sometimes competing, perspectives and giving both a fair shot; and (d) identifying a priori standards for evaluating which predictions were or were not supported by the data.

However, RRs and RRRs work well *with* ACs. ACs can help ensure that RRs and RRRs truly are the fairest and most rigorous tests of the hypotheses they purport to test. And RRs or RRRs help lock adversaries into an explicit research plan to avoid post hoc quibbling about a priori predictions. RRRs are almost *inevitably* adversarial in that they challenge an earlier finding, so AC approaches would almost always be appropriate for RRRs (assuming advocates of the findings in question are still active researchers). And given the risks of interpersonal conflict in ACs, ACs would almost always benefit from RRs. In other words, we see these approaches as mutually beneficial to each other.

A recent illustration of this mutual benefit is a team comprised proponents and opponents of extrasensory perception (ESP; Kekecs et al., 2023). Team members implemented various credibility-enhancing methodologies, including a multilaboratory replication of Bem's Experiment 1 on ESP, which was codesigned by both proponents and opponents of the validity of ESP, employing born-open data and real-time data sharing and data collection, and external research auditors to monitor research integrity. They found 49.89% successful guesses, with the chance level being 50% (using a 99.75% CI, the expectation of successful guesses in the population falls between 49.11% and 50.67% [computed using the final mixed logistic regression within the primary analysis]), thus the observed rate of correct guessing was no higher than chance. Thus, Bem's main claim of the validity of ESP was not replicated, and the *Royal Society Open Science* published the findings as per preagreement.

Kekecs et al.'s (2023) methodology prespecified at the outset how they would interpret evidence for or against the predictions, which ought to be an essential aspect of ACs to maintain credibility. In the absence of such an agreement, the "losing side" in an AC could "move the goalposts" to save their pet theory. Thus, ACs incorporating RRs helps guarantee that the rules of engagement are made clear before data are collected. Chambers argued that ACs should almost always be RRs because, in his words, "by accepting articles in advance, the RR model provides a stronger incentive for researchers to engage in ACs in the first place, free from the risk that reviewers/editors will dislike the results and reject on that basis" (Chambers, 2023, personal communication).

Real Nonpolitical Examples

Cowan et al. (2020) described their experiences engaging in a long-run, multistudy adversarial collaboration, one that has produced several publications regarding findings and their theoretical interpretations with respect to short-term memory. As they described it, although this long-term collaboration has not led any of the adversaries to abandon their preferred theories of short-term memory, it has led to "understanding of others' views and presents to the field research findings accepted as valid by researchers with opposing interpretations" (p. 1011). After pointing out that even ACs do not always provide critical tests that completely falsify the theory preferred by at least one adversary, they continued:

> Even if theorists within the collaboration continue to disagree, the products of the collaboration do, we believe, help to indicate to the field what the true situation is, inasmuch as the opposing views are now applied to a common data set emerging from agreed-upon methods. (Cowan et al., 2020, p. 1013)

Although none of Cowan et al. (2020) adversaries *abandoned* their theory as a consequence of these efforts, all three groups did alter their theories in response to the data that they had communally agreed to produce. As another example of the synergistic effects of AC and preregistration, most of the studies were preregistered, meaning that (a) the adversaries could not generate post hoc apologia denying the relevance or appropriateness of the tests and (b) they could no longer ignore or dismiss hypothesis-disconfirming data because they had produced and published it. This alone is a major improvement over normal operating procedures in psychological science, wherein it is disturbingly common for published articles to simply ignore findings inconsistent with one's preferred narrative (Gambril, 2010; Honeycutt & Jussim, 2020; von Hippel & Buss, 2018).

An even more unequivocally successful AC was recently reported by Killingsworth et al. (2023). In this AC, the first and second authors were able to resolve their long-standing debate over whether there is an income threshold in the increase of emotional well-being with wealth. (The third author served as a neutral referee who also contributed to the design and writing.) They found a linear-log pattern in which average happiness rose consistently with log(income) but only for the least happy individuals; increased income was associated with systematic changes in the shape of the happiness distribution.

Even more important than whether the individual adversaries change their theories, we argue, is that the entire scientific community now has the opportunity to evaluate findings produced by an AC team expressly designed to pit competing predictions derived from those theories. A theory is plausibly viewed as debunked or a line of research degenerating (Lakatos, 1970), not when its most prestigious or aggressive opponents say so but when the community of diverse scholars reaches a consensus that the theory is some version of useless, of extremely limited value, or, in some cases, debunked entirely.

Areas Ripe for ACs

Clark et al. (2022a) listed forty topics of long-standing dispute that would be well-suited for ACs, to which we have added another 20 (see Table 1, albeit itself a woefully

incomplete list). Take, for example, the long-standing and sometimes acrimonious dispute over the validity of affective priming, a dispute that has gone back and forth between proponents and critics for decades. Several researchers have reported automatic priming of an emotional state by a cue. This has been shown to occur even in the absence of conscious awareness, such as when the cue is presented too fast to be perceived, yet it affects the individual's emotional state, or when participants walk slowly after exposure to words related to old age, such as "wrinkle," "bald," and "Florida" (e.g., Bargh et al., 1992; Förster & Liberman, 2007; Payne et al., 2007). A highly cited example of affective priming was provided by the experiments conducted by Chen and Bargh (1999). They showed that participants would pull a lever toward themselves faster when it was positively valenced and push the lever away from themselves when it was negatively valenced, and participants did this without any awareness of the push–pull association with positive–negative emotions. However, other researchers have called into question the validity of affective priming, arguing it is not replicable (Doyen et al., 2012) and subject to experimenter bias and is a "train wreck" as far as the science underpinning it (Kahneman, 2012) and has low statistical power (Rivers & Sherman, 2018; Rotteveel et al., 2015).

Social psychology has many topics ripe for such collaborations. Implicit bias is riddled with controversies, including but not restricted to how strongly it predicts discrimination, whether measures of implicit bias are distinct from explicit measures of prejudice, whether implicit trainings do more harm than good, and whether or not scores of 0 on the implicit association test actually correspond to egalitarian attitudes (for reviews of these controversies, see Cesario, 2022; Gawronski et al., 2022, or the forthcoming edited volume containing chapters by many of the adversaries, Krosnick et al., 2024). Although for the purposes of brevity, we do not review these issues in detail, similar controversies surround microaggressions, the effectiveness of gender-affirming care, the accuracy of stereotypes, the relative prevalence of biases among those on the political left versus right, the magnitude and importance of sex differences, the role of racism in police violence, the predictive validity of standardized achievement tests, and the importance of stereotype threat and growth mindsets in academic achievement, to name a few.

Neither ACs nor preregistrations are platinum standards in the sense that they are guaranteed to resolve these or any other empirical or theoretical controversies. However, the synergistic combination of ACs that are also preregistered, especially in the form of RRs or RRRs, is probably as close as the field can currently come to such a standard. And, as Cowan et al. (2020) so articulately argued, AC-Rs (preregistered ACs) are likely to advance psychological science more effectively, rigorously, and quickly than any known alternative.

Limitations to ACs

ACs, even if generally effective at improving psychological science, are not a silver bullet. Any particular team comprised theoretical or political adversaries may still have its own limitations, biases, or blind spots. Whatever the team members produce is still subject to evaluation by conventional scientific standards, including the wider community of scholars and scientists. ACs may often prevent failures to meet conventional scientific standards, but there is no guarantee. Moreover, one could reasonably argue that ACs will not work as well in political contexts, as some researchers in these areas may care more about ideologies than the truth, the latter being a precondition for ACs to succeed.

As shown in Figure 2, even though ACs may increase rigor, ultimately, methods may trump ACs. All else equal, large, and representative samples are more credible for most purposes than small unrepresentative samples. Preregistered studies with open materials are usually more credible than those without them, and studies with well-validated measures are usually more credible than those with measures concocted on the fly. Similarly, even for theory and review articles (and the theoretical reviews that often constitute introductory and discussion sections), articles that systematically acknowledge conflicting findings and claims are generally more rigorous and should be more persuasive than those that cherry-pick them to support a narrative.

These standards for practices and methods apply whether or not a project is an AC. Therefore, although we expect that ACs will tend to produce more rigorous methods and practices, we fully acknowledge that this will not *always* occur and that plenty of non-ACs use high-quality methods. ACs will primarily improve the validity of scientific outputs when they lead to the use of more rigorous procedures, but might not otherwise. Nonetheless, we do expect ACs will produce more rigorous science in most cases.

In addition, ACs are very difficult to conduct and can result in failed collaborations when diverse team members are unable to come to an agreement, a situation that *PSPI* has experienced on a number of occasions, with team members sometimes resigning out of frustration and anger (Ceci & Williams, 2022). Recently, one journal (*Advances in Methods and Practices in Psychological Science*) put out a call for ACs but reported that there was no response. Some of this may be due to the lack of awareness of what ACs are, but some of it may be the result of the perception (often correct) that teaming up with adversaries can be unpleasant and can put one's pet hypotheses in jeopardy.

However, the potential value of ACs is so great that it behooves social scientists to endeavor to surmount such challenges when possible. Diverse teams do more innovative, impactful research. Among other venues, this has been shown to be true of sequential editing, detection of misinformation,

and wisdom of the crowd (e.g., Shi et al., 2019; Stein et al., 2023). Moreover, the normalization of ACs has the capacity to alter scientific norms and culture such that changing one's mind in response to new evidence is viewed as a mark of integrity, not failure. However, there are currently few formal mechanisms to foster ACs, and without such mechanisms, the concept may not take off.

Creating Institutional Support for ACs

As noted earlier, few requests for funding ACs have been reported, and few "Calls for AC Proposals" have been met with submissions. We propose the establishment of an infrastructure to support and encourage conflicting perspectives among team members, emphasizing ACs but also any other mechanism that infuses a study with opposing voices such as red teams or RRRs. Particularly in cases when the research question concerns a topic on which there have been major disagreements, or if the sociopolitical aspects of a question are obvious, the representation of competing views or ideological positions should be reflected by members of the AC team.

If large professional organizations were interested in this issue, they might explore creating a central exchange in which certain research questions could be generated by editors, policymakers, and others based on the importance of the question for theory or policy, then announce these questions broadly so that psychological scientists can self-nominate or nominate others to be team members to study them. In their nomination, they might be asked to explain the prior work on the research question by themselves or others, and members of the AC would be selected to reflect diverse views with respect to the question. Journals could dedicate space in some issues for publishing the findings of such ACs and establish dedicated RR-AC article tracks. Increasingly, journals now make prepublication commitments for preregistered studies, so committing to publish ACs is consistent with the same goal of maximizing reliable science. Additionally, ACs could be treated favorably in the peer review process and not penalized for forwarding nuanced and complex findings rather than strong one-sided perspectives (as journals currently incentivize).

Ideally, AC-specific funding tracks might be established for particularly controversial or high-stakes research questions. We are currently aware of one initiative launched by the Templeton World Charity Foundation to encourage and support ACs on consciousness. We hope this project will bear fruit, so other funders might follow their lead, but if not, this might warrant little loss for hope given the selection of perhaps the most mysterious phenomenon in all of the human sciences. In general, it seems any organization interested in identifying truth should be interested in supporting adversarial approaches. Research agencies might implement a funding bonus on top of regular grants for authors who accept the risk of working "across the aisle" with rivals/opponents. Universities could lower the transaction costs of ACs by not double-dipping on overhead fees when funds need to be regranted from one institution to another for shared data collection responsibilities.

And, finally, part of the infrastructure should be a means of commissioning retrospective metastudies to determine whether ACs do in fact score higher on validity and reliability and are more likely to move a field forward.

Among the myriad questions that might be examined are as follows: Under what conditions does including an adversary improve methods? Do OS practices improve the quality of ACs and the interpersonal outcomes of the adversaries? How much more effortful are ACs compared to traditional collaborations and are they more or less likely to make it to the publication stage? Do ACs produce more "boring" findings than traditional collaborations? Do early career researchers (ECRs) benefit more from ACs? Are ECRs more willing to participate in ACs? And do established scholars or ECRs risk more from participating in them? What are the different varieties of ACs (e.g., an AC could include a three-part study, where each party initially conducts a study independently, then they come together to run a joint study; or parties could run a joint study first and then consider how their thinking and preferred methods have changed), and what are the costs and benefits of different models?

In contrast to OS procedures, ACs situate definitional and methodological issues front and center. If a research team is truly diverse with respect to the question being examined, it will at the outset of collaboration do something that is not part of procedures designed to increase reproducibility: Team members will have to negotiate with each other about how to frame their hypotheses, operationalize constructs and definitions, agree on the most suitable methods and DVs, and how to interpret the resulting data. AC team members must commit to conditions of falsifiability, explicitly stating which outcome each member of the collaboration will accept as evidence against her or his position, which ideally should be done a priori in a preregistration. None of these negotiations is integral to OS. In our experience as both contributors and commissioning editors of ACs, the negotiation of such issues can be very difficult and sometimes result in dissolving the AC if team members fail to come to an agreement—indeed, sometimes adversaries cannot even agree on what they are disagreeing about. ACs have been dissolved on several occasions for the journal *PSPI* (Ceci & Williams, 2022). RRs would be particularly helpful in this regard, as adversaries may be reluctant to walk away from a project that is already accepted in principle because doing so would deprive the AC of their input.

Conclusion

If research teams included scholars with opposing research agendas, it would likely lead to teams that make more use of strong inference and engage in more severe testing,

thereby optimizing methods and, ultimately, producing better science. OS initiatives, although very important, were not designed to deal with these issues. Even when a team preregisters their hypotheses, sample size, choice of stats, and so forth, it can lead to different findings than an AC team studying the same issue because the latter may define constructs differently, frame hypotheses differently, not accept the evidence as persuasive, and so forth.

The commissioning of ACs could lead to a much-needed cultural shift in psychological science. However, some might argue that this goal is not feasible for psychological science at this time. The only way we will know is if an effort is made to try. Because ACs have the potential to upend researchers' preferred theory, thus entailing reputational/career consequences, we close by suggesting ways to encourage scholars to participate in them.

Christopher Chambers (personal communication, 2023), a leader in the European Open Science movement and cofounder of RRs and the Transparency and Openness guidelines, recently commented that he encounters a similar problem when he gives talks on RRs. He has observed that ECRs:

> Often report that their PI/boss forbids RRs in their lab because they eliminate the PI's ability to selectively publish results that support the "lab brand". This problem may be potentially exacerbated with ACs because at least with RRs the researcher can always try to frame the research question in line with their prior beliefs. To overcome these counter-incentives and feared loss of control over the narrative, there needs to be a powerful incentive for authors to participate in ACs.

We agree that the incentive structure in science is currently misaligned with ACs. Editors and peer reviewers reward big, broad, novel, and surprising claims (i.e., those very likely to be exaggerated, missing important nuance, and that contradict the scientific canon). Because we—scholars—are these editors and peer reviewers, we have the potential to change that. We can instead reward sincere efforts to put competing theories to rigorous tests, even when the results are not earth-shattering. Funders who sincerely want to solve societal problems (rather than promote a particular agenda) can and should encourage and in some cases insist upon ACs as a condition for receiving funding. And employers and awards committees can value signals of earnest truth seeking—such as participation in ACs and use of OS procedures—above suspiciously perfect research agendas (and especially so when those research agendas contradict other scholars' work). The status quo has created the replication crisis and the validity crisis. But luckily, norms can change. And we urge scholars to make ACs the new normal way of engaging in scientific debate.

References

Almaatouq, A., Griffiths, T. L., Suchow, J. W., Whiting, M. E., Evans, J., & Watts, D. J. (2024). Beyond playing 20 questions with nature: Integrative experiment design in the social and behavioral sciences. *Behavioral and Brain Sciences*, *47*, Article e33. https://doi.org/10.1017/S0140525X22002874

Azoulay, P., Fons-Rosen, C., & Zivin, J. S. G. (2019). Does science advance one funeral at a time? *The American Economic Review*, *109*(8), 2889–2920. https://doi.org/10.1257/aer.20161574

Bargh, J. A., Chaiken, S., Govender, R., & Pratto, F. (1992). The generality of the automatic attitude activation effect. *Journal of Personality and Social Psychology*, *62*(6), 893–912. https://doi.org/10.1037/0022-3514.62.6.893

Bateman, I., Kahneman, D., Munro, A., Starmer, C., & Sugden, R. (2005). Testing competing models of loss aversion: An adversarial collaboration. *Journal of Public Economics*, *89*(8), 1561–1580. https://doi.org/10.1016/j.jpubeco.2004.06.013

Bowes, S. M., Clark, C. J., Conway, L. G., III, Costello, T. H., Osborne, D., Tetlock, P., & van Prooijen, J. W. (2024). *An adversarial collaboration on the rigidity-of-the-right, rigidity-of-extremes, or symmetry: The answer depends on the question*. PsyArXiv. https://osf.io/preprints/psyarxiv/4wmx2

Brenan, M. (2023, July 11). Americans' confidence in higher education down sharply. *Gallup News*. https://news.gallup.com/poll/508352/americans-confidence-higher-education-down-sharply.aspx

Breznau, N., Auspurg, K., Brüderl, J., Holzmeister, F., Nilsonne, G., Aczél, B., Clark, C. J., & Uhlmann, E. L. (2023). *Is the inter-researcher variability in social scientific results explicable? An adversarial collaboration and joint effort to parse model and estimate dispersion. Pre-registered analysis plan*. MetaArXiv. https://osf.io/preprints/metaarxiv/h9t5c

Breznau, N., Rinke, E. M., Wuttke, A., Nguyen, H. H. V., Adem, M., Adriaans, J., Alvarez-Benjumea, A., Andersen, H. K., Auer, D., Azevedo, F., Bahnsen, O., Balzer, D., Bauer, G., Bauer, P. C., Baumann, M., Baute, S., Benoit, V., Bernauer, J., Berning, C., … Żółtak, T. (2022). Observing many researchers using the same data and hypothesis reveals a hidden universe of uncertainty. *Proceedings of the National Academy of Sciences of the United States of America*, *119*(44), Article e2203150119. https://doi.org/10.1073/pnas.2203150119

Buss, D. M., & von Hippel, W. (2018). Psychological barriers to evolutionary psychology: Ideological bias and coalitional adaptations. *Archives of Scientific Psychology*, *6*(1), 148–158. https://doi.org/10.1037/arc0000049

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Nave, G., Nosek, B. A., Pfeiffer, T., Altmejd, A., Buttrick, N., Chan, T., Chen, Y., Forsell, E., Gampa, A., Heikensten, E., Hummer, L., Imai, T., … Wu, H. (2018). Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour*, *2*(9), 637–644. https://doi.org/10.1038/s41562-018-0399-z

Casselman, B. (2021, February 23). Doubts feed economics gender gap. *The New York Times*, Section B, p.1. https://www.nytimes.com/2021/02/23/business/economy/economics-women-gender-bias.html

Cech, E. A., & Blair-Loy, M. (2019). The changing career trajectories of new parents in STEM. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(10), 4182–4187. https://doi.org/10.1073/pnas.1810862116

Ceci, S. J., Kahn, S., & Williams, W. M. (2023). Exploring gender bias in six key domains of academic science: An adversarial collaboration. *Psychological Science in the Public Interest*, *24*(1), 15–73. https://doi.org/10.1177/15291006231163179

Ceci, S. J., & Williams, W. M. (2020, October 25). The psychology of fact-checking. *Scientific American*. https://www.scientificamerican.com/article/the-psychology-of-fact-checking1/

Ceci, S. J., & Williams, W. M. (2022). The importance of viewpoint diversity among scientific team members. *Journal of Applied Research in Memory and Cognition*, *11*(1), 35–40. https://doi.org/10.1037/mac0000007

Cesario, J. (2022). What can experimental studies of bias tell us about real-world group disparities? *Behavioral and Brain Sciences*, *45*, Article e66. https://doi.org/10.1017/S0140525X21000017

Chen, M., & Bargh, J. A. (1999). Consequences of automatic evaluation: Immediate behavioral predispositions to approach or avoid the stimulus. *Personality and Social Psychology Bulletin*, *25*(2), 215–224. https://doi.org/10.1177/0146167299025002007

Clark, C. J., Connor, P., & Isch, C. (2023). Failing to replicate predicts citation declines in psychology. *Proceedings of the National Academy of Sciences of the United States of America*, *120*(29), Article e2304862120. https://doi.org/10.1073/pnas.2304862120

Clark, C. J., Costello, T., Mitchell, G., & Tetlock, P. E. (2022a). Keep your enemies close: Adversarial collaborations will improve psychological science. *Journal of Applied Research in Memory and Cognition*, *11*(1), 1–18. https://doi.org/10.1037/mac0000004

Clark, C. J., Costello, T., Mitchell, G., & Tetlock, P. E. (2022b). The road less traveled: Understanding adversaries is hard but smarter than ignoring them. *Journal of Applied Research in Memory and Cognition*, *11*(1), 50–53. https://doi.org/10.1037/mac0000020

Clark, C. J., Fjeldmark, M., Lu, L., Baumeister, R. F., Ceci, S., Frey, K., Miller, G., Reilly, W., Tice, D., von Hippel, W., Williams, W. M., Winegard, B. M., & Tetlock, P. E. (2024). Taboos and self-censorship among U.S. psychology professors. *Perspectives on Psychological Science*. Advance online publication. https://doi.org/10.1177/17456916241252085

Clark, C. J., Isch, C., Connor, P., & Tetlock, P. E. (2024). Assume a can opener. *Behavioral and Brain Sciences*, *47*, Article e36. https://doi.org/10.1017/S0140525X2300239X

Clark, C. J., Isch, C., Everett, J. A., & Shariff, A. (2023). *Even when ideologies align, people distrust politicized institutions*. PsyArXiv. https://osf.io/preprints/psyarxiv/sfubr

Clark, C. J., Pennycook, G., Bhargava, P., Ditto, P. H., Haidt, J., Rand, D., & Tetlock, P. E. (2024). *Do social concerns influence reasoning? An adversarial collaboration* [Manuscript in preparation].

Clark, C. J., & Tetlock, P. E. (2023). Adversarial collaboration: The next science reform. In C. L. Frisby, R. E. Redding, W. T. O'Donohue, & S. O. Lilienfeld (Eds.), *Ideological and political bias in psychology: Nature, scope, and solutions* (pp. 905–927). Springer. https://doi.org/10.1007/978-3-031-29148-7_32

Connor, P., Hahn, A., Clark, C. J., Axt, J., Vianello, M., Lahey, J., Petty, R., Mitchell, G., Costello, T., Tetlock, P. E., & Uhlmann, E. (2024). *On the relationship between automatic associations and discrimination: An adversarial collaboration* [Manuscript in preparation].

Costello, T. H., Clark, C. J., & Tetlock, P. E. (2022). Shoring up the shaky psychological foundations of a micro-economic model of ideology: Adversarial collaboration solutions. *Psychological Inquiry*, *33*(2), 88–94. https://doi.org/10.1080/1047840X.2022.2065130

Cowan, N., Belletier, C., Doherty, J. M., Jaroslawska, A. J., Rhodes, S., Forsberg, A., Naveh-Benjamin, M., Barrouillet, P., Camos, V., & Logie, R. H. (2020). How do scientific views change? Notes from an extended adversarial collaboration. *Perspectives on Psychological Science*, *15*(4), 1011–1025. https://doi.org/10.1177/1745691620906415

Doyen, S., Klein, O., Pichon, C. L., & Cleeremans, A. (2012). Behavioral priming: It's all in the mind, but whose mind? *PLOS ONE*, *7*(1), Article e29081. https://doi.org/10.1371/journal.pone.0029081

Duarte, J. L., Crawford, J. T., Stern, C., Haidt, J., Jussim, L., & Tetlock, P. E. (2015). Political diversity will improve social psychological science. *Behavioral and Brain Sciences*, *38*, Article e130. https://doi.org/10.1017/S0140525X14000430

Förster, J., & Liberman, N. (2007). Knowledge activation. In A. W. Kruglanski & E. T. Higgins (Eds.), *Social psychology: Handbook of basic principles* (pp. 201–231). Guilford Press.

Fortunato, S., Bergstrom, C. T., Börner, K., Evans, J. A., Helbing, D., Milojević, S., Petersen, A. M., Radicchi, F., Sinatra, R., Uzzi, B., Vespignani, A., Waltman, L., Wang, D., & Barabási, A.-L. (2018). Science of science. *Science*, *359*(6379), Article eaao0185. https://doi.org/10.1126/science.aao0185

Gambril, E. (2010). Evidence-informed practice: Antidote to propaganda in the helping professions? *Research on Social Work Practice*, *20*(3), 302–320. https://doi.org/10.1177/1049731509347879

Gauchat, G. (2012). Politicization of science in the public sphere: A study of public trust in the United States, 1974 to 2010. *American Sociological Review*, *77*(2), 167–187. https://doi.org/10.1177/0003122412438225

Gawronski, B., Ledgerwood, A., & Eastwick, P. W. (2022). Implicit bias ≠ bias on implicit measures. *Psychological Inquiry*, *33*(3), 139–155. https://doi.org/10.1080/1047840X.2022.2106750

Gelman, A., & Loken, E. (2013). The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time. *Department of Statistics, Columbia University*, *348*, 1–17. https://api.semanticscholar.org/CorpusID:198164638

Gigerenzer, G. (2022). Simple heuristics to run a research group. *PsyCh Journal*, *11*(2), 275–280. https://doi.org/10.1002/pchj.533

Gruber, J., Mendle, J., Lindquist, K. A., Schmader, T., Clark, L. A., Bliss-Moreau, E., Akinola, M., Atlas, L., Barch, D. M., Barrett, L. F., Borelli, J. L., Brannon, T. N., Bunge, S. A., Campos, B., Cantlon, J., Carter, R., Carter-Sowell, A. R., Chen, S., Craske, M. G., … Williams, L. A. (2021). The future of women in psychological science. *Perspectives on Psychological Science*, *16*(3), 483–516. https://doi.org/10.1177/1745691620952789

Halpern, D. F., Benbow, C. P., Geary, D. C., Gur, R. C., Hyde, J. S., & Gernsbacher, M. A. (2007). The science of sex differences in science and mathematics. *Psychological Science in the Public Interest*, *8*(1), 1–51. https://doi.org/10.1111/j.1529-1006.2007.00032.x

Hedges, L. V. (1984). Estimation of effect size under nonrandom sampling: The effects of censoring studies yielding statistically insignificant mean differences. *Journal of Educational Statistics*, *9*(1), 61–85. https://doi.org/10.3102/10769986009001061

Hill, C., Corbett, C., & Rose, R. (2010). *Why so few? Women in science, technology, engineering, and mathematics*. American Association of University Women.

Honeycutt, N., & Freberg, L. (2017). The liberal and conservative experience across academic disciplines: An extension of Inbar and Lammers. *Social Psychological & Personality Science*, *8*(2), 115–123. https://doi.org/10.1177/1948550616667617

Honeycutt, N., & Jussim, L. (2020). A model of political bias in social science research. *Psychological Inquiry*, *31*(1), 73–85. https://doi.org/10.1080/1047840X.2020.1722600

Hull, D. L., Tessner, P. D., & Diamond, A. M. (1978). Planck's principle: Do younger scientists accept new scientific ideas with greater alacrity than older scientists? *Science*, *202*, 717–723. https://doi.org/10.1126/science.202.4369.717

Inbar, Y., & Lammers, J. (2012). Political diversity in social and personality psychology. *Perspectives on Psychological Science*, *7*(5), 496–503. https://doi.org/10.1177/1745691612448792

Ioannidis, J. P. (2012). Why science is not necessarily self-correcting. *Perspectives on Psychological Science*, *7*(6), 645–654. https://doi.org/10.1177/1745691612464056

Jussim, L., Crawford, J. T., Anglin, S. M., Stevens, S. T., & Duarte, J. L. (2016). Interpretations and methods: Towards a more effectively self-correcting social psychology. *Journal of Experimental Social Psychology*, *66*, 116–133. https://doi.org/10.1016/j.jesp.2015.10.003

Jussim, L., Finkelstein, D., & Stevens, S. T. (2023). Uncertainty, academic radicalization and the erosion of social science credibility. In J. P. Forgas, W. Crano, & K. Fiedler (Eds.), *The psychology of insecurity* (pp. 329–348). Routledge. https://doi.org/10.4324/9781003317623-22

Kahneman, D. (2003). Experiences of collaborative research. *American Psychologist*, *58*(9), 723–730. https://doi.org/10.1037/0003-066X.58.9.723

Kahneman, D. (2012). A proposal to deal with questions about priming effects. *Nature*. https://www.nature.com/news/polopoly_fs/7.6716.1349271308!/suppinfoFile/Kahneman%20Letter.pdf

Kekecs, Z., Palfi, B., Szaszi, B., Szecsi, P., Zrubka, M., Kovacs, M., Bakos, B. E., Cousineau, D., Tressoldi, P., Schmidt, K., Grassi, M., Evans, T. R., Yamada, Y., Miller, J. K., Liu, H., Yonemitsu, F., Dubrov, D., Röer, J. P., Becker, M., … Aczel, B. (2023). Raising the value of research studies in psychological science by increasing the credibility of research reports: The transparent Psi project. *Royal Society Open Science*, *10*(2), Article 191375. https://doi.org/10.1098/rsos.191375

Killingsworth, M. A., Kahneman, D., & Mellers, B. (2023). Income and emotional well-being: A conflict resolved. *Proceedings of the National Academy of Sciences of the United States of America*, *120*(10), Article e2208661120. https://doi.org/10.1073/pnas.2208661120

Kinder, D. R., & Sears, D. O. (1981). Prejudice and politics: Symbolic racism versus racial threats to the good life. *Journal of Personality and Social Psychology*, *40*(3), 414–431. https://doi.org/10.1037/0022-3514.40.3.414

Krosnick, J., Stark, T., & Scott, A. (2024). *The Cambridge handbook of implicit bias and racism*. Cambridge University Press.

Lakatos, I. (1970). History of science and its rational reconstructions. In A. Fine, M. Forbes, & L. Wessels (Eds.), *PSA: Proceedings of the biennial meeting of the philosophy of science association* (pp. 91–136). Cambridge University Press.

Lakens, D. (2020). Pandemic researchers—Recruit your own best critics. *Nature*, *581*(7807), 121–122. https://doi.org/10.1038/d41586-020-01392-8

Lakens, D. (2024). When and how to deviate from a preregistration. *Collabra Psychology*, *10*(1), Article 117094. https://doi.org/10.1525/collabra.117094

Liu, G., & Jones, P. (2024). Understanding implicit bias: Insights and innovations. *Daedalus*, *153*(1).

Löckenhoff, C. E., Chan, W., McCrae, R. R., De Fruyt, F., Jussim, L., De Bolle, M., Costa, P. T., Jr., Sutin, A. R., Realo, A., Allik, J., Nakazato, K., Shimonaka, Y., Hřebíčková, M., Graf, S., Yik, M., Ficková, E., Brunner-Sciarra, M., Leibovich de Figueora, N., Schmidt, V., … Terracciano, A. (2014). Gender stereotypes of personality: Universal and accurate? *Journal of Cross-Cultural Psychology*, *45*(5), 675–694. https://doi.org/10.1177/0022022113520075

Marietta, M., & Barker, D. C. (2019). *One nation, two realities: Dueling facts in American democracy*. Oxford University Press. https://doi.org/10.1093/oso/9780190677176.001.0001

Martel, C., Rathje, S., Clark, C. J., Pennycook, G., Van Bavel, J. J., Rand, D. G., & van der Linden, S. (2024). On the efficacy of accuracy prompts across partisan lines: An adversarial collaboration. *Psychological Science*, *35*(4), 435–450. https://doi.org/10.1177/09567976241232905

Matzke, D., Nieuwenhuis, S., van Rijn, H., Slagter, H. A., van der Molen, M. W., & Wagenmakers, E. J. (2015). The effect of horizontal eye movements on free recall: A preregistered adversarial collaboration. *Journal of Experimental Psychology: General*, *144*(1), e1–e15. https://doi.org/10.1037/xge0000038

Mayo, D. G. (2018). *Statistical inference as severe testing: How to get beyond the statistics wars*. Cambridge University Press. https://doi.org/10.1017/9781107286184

Mellers, B., Hertwig, R., & Kahneman, D. (2001). Do frequency representations eliminate conjunction effects? An exercise in adversarial collaboration. *Psychological Science*, *12*(4), 269–275. https://doi.org/10.1111/1467-9280.00350

Melloni, L., Mudrik, L., Pitts, M., Bendtz, K., Ferrante, O., Gorska, U., Hirschhorn, R., Khalaf, A., Kozma, C., Lepauvre, A., Liu, L., Mazumder, D., Richter, D., Zhou, H., Blumenfeld, H., Boly, M., Chalmers, D. J., Devore, S., Fallon, F., … Tononi, G. (2023). An adversarial collaboration protocol for testing contrasting predictions of global neuronal workspace and integrated information theory. *PLOS ONE*, *18*(2), Article e0268577. https://doi.org/10.1371/journal.pone.0268577

National Academies of Sciences, Engineering, and Medicine. (2020). *Promising practices for addressing the underrepresentation of women in science, engineering, and medicine: Opening doors*. The National Academies Press. https://doi.org/10.17226/25585

National Academy of Sciences, National Academy of Engineering, & Institute of Medicine. (2007). *Beyond bias and barriers: Fulfilling the potential of women in academic science and engineering*. National Academies Press. https://doi.org/10.17226/11741

Nelson, L. (2018). *How many studies have not been run? Why we still think the average effect size does not exist*. Data Colada: Thinking about evidence and vice versa. https://datacolada.org/70

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(11), 2600–2606. https://doi.org/10.1073/pnas.1708274114

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), Article aac4716. https://doi.org/10.1126/science.aac4716

Payne, B. K., McClernon, F. J., & Dobbins, I. G. (2007). Automatic affective responses to smoking cues. *Experimental and Clinical Psychopharmacology*, *15*(4), 400–409. https://doi.org/10.1037/1064-1297.15.4.400

Planck, M. (1950). *Scientific autobiography and other papers*. Philosophical library.

Platt, J. R. (1964). Strong inference: Certain systematic methods of scientific thinking may produce much more rapid progress than others. *Science*, *146*(3642), 347–353. https://doi.org/10.1126/science.146.3642.347

Rivers, A. J., & Sherman, J. W. (2018). *Experimental design and the reliability of priming effects: Reconsidering the "train wreck."* PsyArXiv. https://doi.org/10.31234/osf.io/r7pd3

Rotteveel, M., Gierholz, A., Koch, G., van Aalst, C., Pinto, Y., Matzke, D., Steingroever, H., Verhagen, J., Beek, T. F., Selker, R., Sasiadek, A., & Wagenmakers, E. J. (2015). On the automatic link between affect and tendencies to approach and avoid: Chen and Bargh (1999) revisited. *Frontiers in Psychology*, *6*, Article 335. https://doi.org/10.3389/fpsyg.2015.00335

Schaerer, M., du Plessis, C., Nguyen, M. H. B., van Aert, R. C. M., Tiokhin, L., Lakens, D., Giulia Clemente, E., Pfeiffer, T., Dreber, A., Johannesson, M., Clark, C. J., Gender Audits Forecasting Collaboration, & Uhlmann, E. L. (2023). On the trajectory of discrimination: A meta-analysis and forecasting survey capturing 44 years of field experiments on gender and hiring decisions. *Organizational Behavior and Human Decision Processes*, *179*, Article 104280. https://doi.org/10.1016/j.obhdp.2023.104280

Shi, F., Teplitskiy, M., Duede, E., & Evans, J. A. (2019). The wisdom of polarized crowds. *Nature Human Behaviour*, *3*(4), 329–336. https://doi.org/10.1038/s41562-019-0541-6

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632

Simons, D. J., Holcombe, A. O., & Spellman, B. A. (2014). An introduction to registered replication reports at perspectives on psychological science. *Perspectives on Psychological Science*, *9*(5), 552–555. https://doi.org/10.1177/1745691614543974

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-curve and effect size: Correcting for publication bias using only significant results. *Perspectives on Psychological Science*, *9*(6), 666–681. https://doi.org/10.1177/1745691614553988

Sniderman, P. M., & Tetlock, P. E. (1986). Symbolic racism: Problems of motive attribution in political analysis. *Journal of Social Issues*, *42*(2), 129–150. https://doi.org/10.1111/j.1540-4560.1986.tb00229.x

Starns, J. J., Cataldo, A. M., Rotello, C. M., Annis, J., Aschenbrenner, A., Bröder, A., Cox, G., Criss, A., Curl, R. A., Dobbins, I. G., Dunn, J., Enam, T., Evans, N. J., Farrell, S., Fraundorf, S. H., Gronlund, S. D., Heathcote, A., Heck, D. W., Hicks, J. L., … Wilson, J. (2019).

Assessing theoretical conclusions with blinded inference to investigate a potential inference crisis. *Advances in Methods and Practices in Psychological Science*, *2*(4), 335–349. https://doi.org/10.1177/25152459 19869583

Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science*, *11*(5), 702–712. https://doi.org/10.1177/17456916 16658637

Stein, J., Frey, V., & van de Rijt, A. (2023). Realtime user ratings as a strategy for combatting misinformation: An experimental study. *Scientific Reports*, *13*(1), Article 1626. https://doi.org/10.1038/s41598-023-28597-x

Stern, C., & Crawford, J. (2021). Ideological conflict and prejudice: An adversarial collaboration examining correlates and ideological (a)symmetries. *Social Psychological & Personality Science*, *12*(1), 42–53. https://doi.org/10.1177/1948550620904275

Tasimi, A., & Friedman, O. (2024). An adversarial collaboration on dirty money. *Social Psychological & Personality Science*, *15*(3), 255–263. https://doi.org/10.1177/19485506231167231

Tetlock, P. E., & Mitchell, G. (2009). Adversarial collaboration aborted but our offer still stands. *Research in Organizational Behavior*, *29*, 77–79. https://doi.org/10.1016/j.riob.2009.06.012

Urry, M. (2015). Science and gender: Scientists must work harder on equality. *Nature*, *528*(7583), 471–473. https://doi.org/10.1038/528471a

Van Dessel, P., Gawronski, B., Smith, C. T., & De Houwer, J. (2017). Mechanisms underlying approach–avoidance instruction effects on implicit evaluation: Results of a preregistered adversarial collaboration. *Journal of Experimental Social Psychology*, *69*, 23–32. https://doi.org/10.1016/j.jesp.2016.10.004

Vazire, S., Schiavone, S. R., & Bottesini, J. G. (2022). Credibility beyond replicability: Improving the four validities in psychological science. *Current Directions in Psychological Science*, *31*(2), 162–168. https://doi.org/10.1177/09637214211067779

Vlasceanu, M., Reinero, D. A., & Van Bavel, J. J. (2022). Adversarial collaborations in behavioral science: Benefits and boundary conditions. *Journal of Applied Research in Memory and Cognition*, *11*(1), 23–26. https://doi.org/10.1037/mac0000002

von Hippel, W., & Buss, D. M. (2018). Do ideologically driven scientific agendas impede understanding and acceptance of evolutionary principles in social psychology? In J. T. Crawford & L. Jussim (Eds.), *The politics of social psychology* (pp. 7–25). Psychology Press.

Wicherts, J. M., Bakker, M., & Molenaar, D. (2011). Willingness to share research data is related to the strength of the evidence and the quality of reporting of statistical results. *PLOS ONE*, *6*(11), Article e26828. https://doi.org/10.1371/journal.pone.0026828

Witteman, H. O., Hendricks, M., Straus, S., & Tannenbaum, C. (2019). Are gender gaps due to evaluations of the applicant or the science? A natural experiment at a national funding agency. *Lancet*, *393*(10171), 531–540. https://doi.org/10.1016/S0140-6736(18)32611-4

Witze, A. (2020). Three extraordinary women run the gauntlet of science—A documentary. *Nature*, *583*(7814), 25–26. https://doi.org/10.1038/d41586-020-01912-6

Wixted, J. T., & Wells, G. L. (2017). The relationship between eyewitness confidence and identification accuracy: A new synthesis. *Psychological Science in the Public Interest*, *18*(1), 10–65. https://doi.org/10.1177/1529100616686966

Wright, J. D., Goldberg, Z., Cheung, I., & Esses, V. M. (2021). *Clarifying the meaning of symbolic racism* [Unpublished manuscript]. https://www.researchgate.net/publication/349681203_Clarifying_the_meaning_of_symbolic_racism

Zhang, F. J. (2023). Political endorsement by Nature and trust in scientific expertise during COVID-19. *Nature Human Behaviour*, *7*(5), 696–706. https://doi.org/10.1038/s41562-023-01537-5

Received March 21, 2024
Revision received May 23, 2024
Accepted June 2, 2024 ■