1. Introduction

Diversity, equity and inclusion represent broad values and goals that businesses, institutions and communities seek to embody. This report addresses the state of knowledge about how diversity, equity and inclusion are considered in the competing, reviewing and awarding of research grants, and how the review process impacts the outcomes of scholarly research. While many academic institutions and funding bodies have expressed support for these values, little experimental and causal research exists regarding the effects of diversity, equity and inclusion programs on research. Thus, the report draws on larger bodies of scholarship about peer review practices, scientific production and diversity in group processes within corporate as well as academic programs. Considering these literatures together provides descriptive work about how different processes of review and deliberation shape outcomes. Studying peer review as a social process uncovers the mechanisms that contribute to inequality and that both advance and hinder the scientific process.

The report proceeds with five sections. First, it defines diversity, equity and inclusion with attention to how the concepts have been operationalized and used in the context of peer review and studies of science. Next, the report provides an overview of peer review, including its historical and ideational roots, which continue to shape how peer review is structured. The second section also provides four challenges to diversity, equity and inclusion present in standard peer review practices. The third section addresses benefits of diversity, equity and inclusion within the scientific enterprise, incorporating research on scientific excellence and group processes. The fourth section provides models of peer review and interventions to support diversity, equity and inclusion from two other major American based funding organizations, namely The National Science Foundation and the National Institute of Health, who have each reflected on their practices in support of these values. The final section offers best practices and recommendations for funding bodies based on a synthesis of the literature, as well as recognizes knowledge gaps to evaluate the effects of diversity, equity and inclusion in peer review on the scientific process.
2. Definitions and operationalization


diversity

Diversity is broadly defined as the range of human difference. Most commonly, diversity refers to difference based on race, ethnicity, gender, sexual orientation, age, social class, physical ability, religious background, national origin and/or political beliefs. The attributes that are salient in defining diversity depend on social context and group dynamics, specifically how a group or society constructs categories of sameness and difference. For example, a young female worker may find her gender identity most salient within a predominantly male work group, but her age more salient in an inter-generational, familial setting (Brekhus 2020). Within the context of peer review and science studies, diversity is commonly operationalized and studied as gender and disciplinary background, difference based on race and/or ethnicity, epistemic orientation, LGBTQ identification and institutional employment (Cech and Waidzunas 2021; Ginther et al. 2011; Guzzo and Salas 1995; Huutoniemi 2012; Lamont and Da Silva 2009; Laudel 2006; Luo et al. 2021; Smith-Doerr et al. 2016; Whittaker et al. 2015).

Equity

Equity refers to the quality of being fair and impartial while considering current and past realities. Equity differs from equality because it accounts for historical and situational factors that limit peoples’ ability to achieve equivalent levels of success. Diversity is linked to equity by considering the role of difference enroot to equal outcomes. Within the context of peer review and science studies, equity has mainly been considered in terms of different career outcomes with an emphasis on gender, as well as some work on race, class status and LGBTQ background. Career outcomes considered include academic hiring, promotion and citation rates and funding levels (Cardel et al. 2020; Ellemers et al. 2004; Hengel 2020; Isbell et al. 2012; Moss-Racusin et al. 2012; Sugimoto 2013; Teele and Thelen 2017; Van der Lee and Ellemers 2015).

Inclusion

Inclusion encompasses the involvement and empowerment of individuals within a collective by promoting a sense of belongingness, while also recognizing uniqueness (Shore et al. 2011). Inclusive groups recognize the value, worth and dignity of all members and their contributions to the collective. Within the context of peer review and science studies, inclusion has been studied as the incorporation of underrepresented groups within the scientific discipline (Cheryan et al. 2017; Griffith and Dasgupta 2018) and the relationship between inclusive work environments and work productivity, success rates and satisfaction (DeAro et al. 2019; Gurthrie et al. 2019; Hong and Page 2004).

3. Peer review

Peer review is a central practice within academia in which those working in similar disciplinary fields evaluate the research of colleagues. Peer review developed in the 17th century, emerging as a distinct practice with the proliferation of individual authorship in the publishing trade (Huutoniemi 2015). The creation of scientific journals redefined peers in terms of
specialized expertise and nonlocal colleagues, who then acted as gatekeepers in the scientific enterprise (Csiszar 2006). In the 19th century, with an increase in funding for scientific pursuits, the peer review model transferred to grant and selection committees, expanding reviewer qualifications beyond academics to include a range of researchers and administrators. Peer review thus became the marker of scientific legitimacy as expressed by expert peers, as well as a method of holding science accountable through review by funding bodies (Baldwin 2018).

Peer review institutionalized norms and values of 17th century, enlightenment science. The practice built on sociologist of science, Robert Merton’s (1973) ideals of organized skepticism, communality, universalism and disinterestedness. Peer review relied on the belief that detached scrutiny of ideas existed as an objective criterion for determining their worth. The ideal of universalism relates to equality and the notion that all research and researchers should be treated and evaluated equally. He further defined a split between rational, universal factors that applied across cases and scenarios that were seen as objective, in opposition to particularistic or social factors that related to individual identities, processes and contexts that were defined as subjective (Reinhart 2009). Research shows that elements of 17th century peer culture persist in peer review practices, including the separation of universalism and particularism in defining bias based on identity categories (Langfeldt et al. 2020), norms of not speaking negatively about colleagues and upholding disciplinary specific standards of quality (Reinhart 2010).

Since the latter 20th century, research has challenged Merton’s idealized conception of science and belief in objective forms of knowledge production. Historical and comparative research shows how definitions of scientific objectivity change over time and across contexts. Objectivity has been defined variably as representations of normality, truth-to-nature representations, consensus among scientific elites or public demonstrations of scientific principles (Daston 1992; Lamont et al. 2011; Latour 1999). Divides between universal, objective facts and particularistic, subjective viewpoints rely on “boundary work” or the construction of categories and institutions that define science as distinct from individual, social and political activities (Gieryn 1995). Furthermore, some groups, historically white, upper-class men, have been able to speak more from a “universalist” standpoint, whereas the voices of those from historically underrepresented groups within the scientific enterprise have been labeled more “particularistic.” Miranda Fricker (2007) coined the term “epistemic justice” to refer to the structured ways in which some groups’ voices are not seen as objective due to a deflated sense of credibility based on identity prejudices that distort images of the social type of the speaker. For example, the labeling of theories from female scholars as “feminist” positions women as unable to produce generalist or universal knowledge (Bacevic 2021). Research continues to show that knowledge claims made by women and people of color are more frequently questioned and met with doubt and suspicion compared to claims made by white or male colleagues (Dupas et al. 2021; Dotson 2011; Petty et al. 1999).

In the following section, the report outlines critiques of peer review that contradict the process’s universalism and present challenges for incorporating diversity, equity and inclusion in science and academia. These critiques include 1) contending definitions of research excellence; 2) bias in decision-making practices and outcomes; 3) conservatism and risk-aversion; and 4) the contribution of peer review to inequality in career outcomes.

*Determining Research Excellence*
Within the review process, defining research excellence remains a contentious and debated standard, influenced by idiosyncratic and personal interests, as well as social dynamics (Laudel 2006). Peer review, studied as a social, multi-step process of evaluation, relies on ordering and ranking categories and recognizing, by oneself and others, the value of scholarship (Hirschauer 2010; Krueger and Reinhart 2018; Lamont 2012). Ranked categories and recognition are not innate qualities but negotiated between people who hold different interests and preferences (Bourdieu 1993). Research on peer review as a process uncovers heterogenous conceptions of research quality, limited convergence on outcome decisions and high levels of arbitrariness (Brezi and Birukou 2019; Esarey 2017; Guetzkow et al. 2004; Langfeldt et al. 2020). Reviewers may prioritize the originality of research, its plausibility, the importance of its contribution within academia or within applied fields or methodological soundness (Langfeldt et al. 2020). Idiosyncratic preferences and tastes intertwine with criteria of evaluation. For example, Lamont and Huutoniemi (2011) show how definitions of originality and interestingness relate to reviewers’ own areas of expertise, with reviewers defining work as more interesting that resonates with their own identities and specialties. Increasing review criteria results in more variance and arbitrariness in review outcomes, rather than more detailed reviews (Brezi and Birukou 2020).

Evaluators also draw on different epistemological styles or orientations towards knowledge production. In contrast to ideals of universalism, reviewers define fairness based on using epistemological styles that are seen as most appropriate to the field or discipline of the proposal under review (Mallard et al. 2009). Similarly, Reinhart (2010) shows how reviewers adjust their language to adhere to disciplinary norms, using the language of creativity to positively evaluate work that they find exciting, while using the language of rigidity or instrumentality in negative evaluations as to not directly offend or critique their peers. Panels containing reviewers from diverse disciplinary and career backgrounds show more diversity in notions of research quality, but also more flexibility in negotiating judgements amongst one another (Huutoniemi 2012; Langfeldt et al. 2020). In this way, peer review encompasses interactional, emotional and cognitive processes. The identity of the researcher, the norms of research evaluation, the language deemed appropriate within disciplines and the makeup of review panels influence how one defines excellence and evaluates research in contrast to the application of universal practices and standards.

**Bias**

Research focused on the outcomes of peer review highlight biases that disadvantage underrepresented groups in academia based on reviewer background and the social dynamics of review panels (Guthrie et al. 2019; Shah 2021; Squazzoni 2021; Van der Lee and Ellemers 2015). Peer review requires making decisions under conditions of uncertainty to predict the significance of research before the research is completed. Classic work in psychology shows how decision-making under uncertainty results in the use of heuristics to predict future values, resulting in an array of biases such as the retrievability of past experiences and models and the imaginability of future options (Kahneman et al. 1974). Within uncertain situations, people often rely on stereotypes to reflect general expectations about groups in the evaluation process (Ellemers 2017). Ample research documents how gender stereotypes influence the evaluation of women in academia, disadvantaging them in hiring, grant review, teaching evaluations, group meeting dynamics and publication acceptance rates (Cardel et al. 2020; Correll 2017; Ellemers et
al. 2004; Rivera 2017; Severin et al. 2020). Application processes in which reviewers lack complete information about a proposal or are overburdened with work result in an increased reliance on stereotypes and biases to infer quality (Guthrie et al. 2018; Guthrie et al. 2019; Teplitskiy et al. 2018). In contrast, junior or new reviewers have been shown to be more engaged in the review process and produce reviews rated as higher quality than experienced reviewers who default to preformed judgments and modes of evaluation (Shah 2021).

Reviewer characteristics also impact individual biases and criteria used for evaluation. For example, studies show that male reviewers give higher scores to other male applicants, while women do not differ in scores given to women or men (Severin et al. 2020). Other research finds that male reviewers rely more on bibliometrics than female reviewers (Cruz-Castro and Sanz-Menendez 2021). Reviewers also have been shown to support their own self-interest through “cognitive particularism,” or preferences for research closer to their own area of expertise or by other scholars within their own networks (Laudel 2006; Teplitskiy et al. 2018). Panels that contain reviewers from similar research traditions show more bias as reviewers compete for authority by adhering more strongly to the paradigms within their area of expertise (Huutoniemi 2012). This particularly disadvantages interdisciplinary research or research done by less well known, younger or fringe researchers, limiting diversity and inclusion within academia. Other dynamics documented on review panels, such as strategic voting (giving a low rank to some proposals to increase the likelihood that others will win) or horse-trading (enabling other panelists’ objectives in the hopes that they will reciprocate), show how personal interests and preferences bias review outcomes (Lamont and Huutoniemi 2011). These dynamics perpetuate exclusions (Bacevic 2021; Elsass and Graves 1997; Whittaker et al. 2015) and raise the risk of reviewers converging on incorrect scientific assumptions, due to preferences for particular paradigms or networks (Park et al. 2014).

Conservatism

The tendency to support research closer to one’s own discipline, as well as the use of disciplinary specific criteria of evaluation result in the support of more conservative and incremental, rather than risky or novel, research. This trend disadvantages underrepresented groups in science whose claims are often seen as less legitimate and riskier (Bacevic 2021; Blair-Loy et al. 2017; Dupas et al. 2021; Petty et al. 1999), and limits diversity and inclusion. For example, Hofstra et al. (2020) found that underrepresented groups contribute to more innovative discoveries, defined as the first instance of linking discipline specific concepts in a thesis, however these contributions are taken up at a lower rate, less likely to contribute to successful scientific careers or result in positive recognition compared to findings by majority group members (Abir-Am 2020). The makeup of review panels can discourage support for innovative work. Panels in which reviewers are closely aligned in discipline trend towards rating works higher that resonate with a reviewer’s own research approach and objectives based on disciplinary standards (Huutoniemi 2012; Laudel 2006; Li 2017). In a simulated experiment of panels reviewing the same research projects, Brezi and Birukou (2020) found that the most innovative projects receive the highest variance of reviews, and in consequence, are accepted at the lowest rate. In another randomized controlled experiment, Boudreau et al. (2016) found that evaluators gave lower scores to research proposals that are highly novel, again defined as a new combination of field specific terms.
Furthermore, the standardization of grant funding procedures and field boundaries reduces funding for unconventional and cross-disciplinary projects (Laudel and Galser 2014). An emphasis on reviewer agreement (Gurthrie et al. 2018), strict standards and bibliometric scores has been shown to reduce support for projects that can lead to unexpected findings or use new approaches (Azoulay and Li 2020; Langfeldt et al. 2010). Researchers themselves may also alter their behavior, research focus and proposals as a reaction to being evaluated and to conform to disciplinary standards, produce standardized, measurable impacts and frame their research as less risky (Espeland and Sauder 2007; Martin 1997). Conservatism further retrenches established paradigms, disciplinary boundaries and metrics that define success, at the expense of more expansive, flexible or creative projects that support and recognize diverse scientific approaches.

**Compounded Inequality**

Ideals of universalism, in which everyone is treated equally and evaluated based on the same criteria, contribute to the “Matthew effect” in which those with access to resources, status and prestige continue to succeed, while those with fewer resources struggle to gain recognition and success (DiPrete and Eirich 2006; Merton 1973). Multiple studies show that past track record is the most predictive of peer review success (Bornmann and Daniel 2007; Severin et al. 2020; Shah 2021; Teplitskiy et al. 2018; Taffé 2021; Van der Lee and Ellemers 2015). However, women and people of color face barriers throughout their careers in hiring, promotions, publications, awards and recognition, which limit their ability to develop a robust track record compared to white and/or male colleagues (Sugimoto 2013; Whittaker et al. 2015). Furthermore, research shows that funding contributes to predicting project and career success, thus those that gain funding can continue to prosper, while those that struggle to gain funding continue to face challenges (Reinhart 2009). In this way, peer review, grounded in ideals of universalism and meritocracy, compounds inequality by continually directing resources to previously successful and high-status researchers at the expense of achieving diversity, inclusion and equity (Ginther et al. 2011; Shah 2021; Teplitskiy et al. 2018). Cycles of advantage and disadvantage can serve as self-fulfilling prophecies in which people come to expect certain positive or negative outcomes and modify their behavior accordingly (Ellemers 2017; Ellemers and Rink 2005; Kanter 1977). For example, the lack of representation of women in leadership positions on review committees affects perceptions of women’s adequacy and success, leading women to doubt their own abilities (Squazzoni et al. 2021). Compounded inequality contributes to the “leaky pipeline” in STEM in which women dropout of STEM careers, as well as perpetuates disadvantages for groups historically excluded from science (Cardel et al. 2020; Misra et al. 2017; Severin et al. 2020).

4. **Benefits of diversity, equity and inclusion**

Addressing the shortcomings of peer review requires active and purposeful interventions (Moss-Racusin et al. 2014; Whittaker et al. 2015). Before laying out models for interventions, this section documents the benefits of incorporating diversity, equity and inclusion to improve research outcomes and decision-making processes.

*Scientific quality and innovation*
Diversity promotes expanded ways of knowing in terms of both method and perspective, which strengthens research excellence and produces higher quality outcomes (Haraway 1991; Whittaker et al. 2015). While peer review has been critiqued for promoting conservatism and enforcing disciplinary boundaries, research consistently shows that more creative and collaborative work has a larger impact. Uzzi et al. (2013) found that papers worked on by teams that combined conventional ideas in unusual combinations showed higher impact factors, measured by citation networks, than narrow papers. Freeman and Huang (2014) found that papers produced by homogeneous research teams published in lower impact journals and received fewer citations than papers produced by diverse teams in terms of author ethnicity, location and reference history. Similarly, Campbell et al. (2013) found that gender-heterogenous teams produced publications with 34% more citations than publications produced on gender-uniform teams.

Review panels with scholars from multiple disciplinary backgrounds and approaches more frequently support diverse forms of research by extending definitions of quality beyond disciplinary norms (Langfeldt et al. 2020). Huutoniemi (2012) found that multi-disciplinary panels produced complementary judgements to recognize broader merits of proposals, such as environmental impacts, while panels of researchers from similar backgrounds competed to establish their expertise and authority using narrow criteria to advance specific fields. Multi-disciplinary review panels resulted in the support of more interdisciplinary research. Panels containing scholars from different research backgrounds and traditions also pay more attention to the process of evaluation itself, defining fairness based on negotiations between reviewers and evaluative criteria, rather than hold to “universalist” practices of considering standardized criteria of evaluation (Mallard et al. 2009). Combing criteria of evaluation and multiple viewpoints creates productive friction to reflexively consider new ideas and approaches (Stark 2011). Considering more diverse criteria of evaluation has been advocated to support innovative and risk-taking research (Azoulay and Li 2020; Dezso and Ross 2012; Hofstra et al. 2020; Valantine and Collins 2015).

**Decision-making processes**

Research finds that in general diverse and inclusive teams exchange a wider range of information, exhibiting more creativity, flexibility and thoughtfulness in decision making processes (Antonio et al. 2004; Elsass and Graves 1997; Hong and Page 2004; Sommers 2006). Those with access to a broader range of perspectives show more creativity in their thinking (Page 2010). This allows them to connect disparate ideas and produce and share information that is more highly regarded (Burt 2004). The benefits of diversity for decision-making extends beyond the inclusion of more voices. Majority group members also behave differently when interacting with diverse others. Sommers (2006) found that whites in mixed-race jury panels demonstrated more complex thinking and processed trial information more systematically in anticipation of encountering those different from oneself. This led to heterogenous groups deliberating longer and considering a wider range of information to come to their conclusions. Similarly, Dezso and Ross (2012) found that the presence of a women in a predominately male group stimulated broader and richer discussion. In the context of peer review, panels with mixed academic and non-academic reviewers produced longer, more concrete and detailed impact evaluations than
those by academic researchers who mainly focused on criteria of scientific excellence without considering broader impacts (Luo et al. 2021).

5. **Interventions to increase diversity, equity and inclusion**

Organizations that seek to support scientific research excellence have developed various models and interventions to counteract bias, trends towards conservatism and inequality within academia and promote diversity, equity and inclusion. This section outlines the review practices and programs at two other major American funding organizations. While there are limited published studies on the impacts of many of these interventions, the concluding section offers concrete recommendations drawing from research on review processes beyond funding bodies.

*The National Science Foundation*

The National Science Foundation (NSF) conducts peer review, typically with three reviewers per panel. Panelists send recommendations to a program officer who recommends a final funding decision before a division director reviews the decision for support. Diverse funding opportunities, including standard grants, as well as special programs for exploratory and high-risk research seek to provide flexible funding opportunities (Langfeldt and Scordato 2016). NSF attempts to select reviewers broadly, including reviewers with knowledge of the subfield under study, wider knowledge of the disciplinary field, insight into the institutional workings of science and technology and from diverse backgrounds. Efforts to diversify review panels include allowing more flexible reviewing opportunities, such as remote work, which was shown to increase the participation of female reviewers (Pinholster 2016).

Review panels provide one overall score based on a 5-point scale (poor to excellent) that focus on two main criteria, intellectual merit and broader impacts (Langfeldt and Scordato 2015). NSF introduced the broader impacts criteria in 1997 to improve public accountability (Watts et al. 2015). Prior to 1997, review scores focused on prior researcher performance, intrinsic merit of proposal, societal relevance and contribution to research infrastructure. However, the broader impacts criteria have been critiqued as vague, poorly understood and reliant on reviewer discretion, particularly in comparison to the greater detail provided about the intellectual merit criteria (Bozeman and Youtie 2017). NSF has made efforts to clarify the criteria to evaluators. Initially defined by three categories-- teaching and training, research dissemination and broadening participation in science-- changes to review guides in 2013 sought to define “broader impacts” in terms of novelty, impact and feasibility (Watts et al. 2015). Nonetheless, analysis of NSF reviews and awards indicates that both reviewers and PIs tend not to comment on many broader impact components, and when they do, they mostly emphasize teaching (Watts et al. 2015). The broader impacts criteria have also been criticized as incompatible with conventional peer review practices, based on scientific and technical expertise, which exclude non-experts who may be able to make more informed judgements about broader social impacts (Bozeman and Youtie 2017). Broader impacts are also difficult to evaluate. NSF requires PIs to comment on broader impacts in annual progress reports and often rely on case studies of broader impacts rather than quantitative metrics (Langfeldt and Scordato 2015).

In the last two decades, NSF has created specific programs aimed at advancing diversity, equity and inclusion in science. The ADVANCE program is one of the most well documented
programs to address equity in STEM and support a more diverse and capable science and engineering workforce. Started in 2001, the program has invested over $270 million to more than 200 academic and non-profit institutions to implement evidence-based, systematic change strategies. The program builds on evidence that diversity in backgrounds and perspectives is a powerful resource for scientific production. Originally focused on gender equity and outcomes, the program has expanded to include concerns about racial and ethnic disparities related to institutional and professional policies, practices, cultures and climates. Grants focus on institutional transformation by addressing systemic changes to organizations, as well as creating adaptable strategies that can be evaluated and shared, building partnerships to support change across fields and catalyzing a range of partners to undertake institutional self-assessment. Typical programs incorporate interventions focused on mentoring, networking, professional development, work-life balance, departmental culture, hiring, promotion, retention and leadership policies. Interventions use data-driven techniques applied to academic and scientific institutions. For example, implicit bias remains a core ADVANCE concept because it is measurable and demonstratable, actionable and seen as impartial and grounded in social-scientific research (Nelson and Zippel 2021).

In the review process, ADVANCE guidelines consider a project’s potential for impact on institutions, and the ways in which it widens opportunities to produce scientific knowledge. The separation of intellectual merit and broader impact criteria consider diversity, equity and inclusion in terms of “broader impacts,” which is distinct from other funding organizations that also consider how diversity, equity and inclusion impact the subjects of supported projects (Zippel and Ferree 2019). The program prioritizes self-reflection on the part of institutions applying for ADVANCE grants, as well as within the ADVANCE program. All ADVANCE NSF review panels participate in implicit bias trainings. Throughout its 20-year existence, ADVANCE adapted interventions through a model of test, apply, evaluate and refine. The program has changed over time using this self-reflective strategy, for example, by adopting in 2016 an intersectional approach that considers race and ethnicity alongside gender inequity (Nelson and Zippel).

ADVANCE programs have been linked to increases in job satisfaction, hiring and retention of women by reformulating mechanisms of evaluation to reward expanded forms of knowledge production, considering feminist and experiential understandings to promote inclusion and empower gender, racial and ethnic minority groups (Zippel and Ferree 2019). In this way, ADVANCE has been able to produce networks of actors with gender expertise both within and across research and academic institutions. Additionally, the structural focus of ADVANCE has been an influential model for programs at other funding organizations. However, case studies of institutions that adopted ADVANCE programs show that structures must be kept in place even after ADVANCE funding to retain improvements in hiring, promotion and leadership equity (Stepan-Norris and Kerrissey 2016).

NSF continues to prioritize diversity, equity and inclusion in science, unveiling 10 “Big Ideas” in 2017. Two priorities explicitly relate to diversity and inclusion in science, namely an emphasis on transdisciplinary and convergence research and the NSF INCLUDES program. A focus on convergence research aims to address specific challenges, whether a deep scientific question or social need. The program supports research that combines expertise from different disciplines to pursue common research, create new frameworks and sustain interactions across communities. It seeks to broaden participation in STEM by empowering a range of stakeholders in the scientific process and recognizing the importance of multiple forms of expertise. In this
way, the convergence research track applies expanded criteria of evaluation and aims to avoid conservatism to support riskier and more innovative projects. The aim of NSF INCLUDES matches much of the ongoing ADVANCE program by building networks of researchers and collaborative infrastructure that include partnerships with private and corporate philanthropy, federal agencies and professional societies.

Research has not documented the effects of individual programs on overall review and funding award outcomes. As of 2019, NSF showed higher funding rates for women (31%), compared to men (28%), however men still submit more than double the number of research proposals as women. Similarly, as of 2018 the funding rates of white PIs was 26% compared to rates of 23% for Hispanic/Latino PIs, 19% for Black/African American PIs and 17% for Asian PIs, however white PIs submitted more than double the number of applications as any other group and 25 times as many applications as Black/African American PIs, the lowest submitting group. Program officers, who make final recommendations about review decisions, still skew towards a white (71%) and male (53%) population. NSF’s strategic plans continue to emphasize a commitment to diversity and inclusion as a strength for America’s research and innovation ecosystem as the organization strives to support underrepresented scholars.

**National Institute of Health**

The National Institute of Health (NIH) claims to structure its review process on the values of fairness, equity, timely and bias free practices. Projects go through a two-tiered review system. First, a group of non-federal scientists with expertise in the specific discipline of the project under review prepare a written critique based on judgments of merit, technical soundness and the protection of human subjects. These reviewers are instructed to pay particular attention to submitters’ publication record, research funding history, scientific achievement and colleague recommendations. They give individual scores, and those with the highest average are selected for panel discussion to revise score marks and send to the second tier of review. Secondly, institute and center national advisory councils or boards that contain both scientific and public representatives review projects based on general interest and relevance for matters of health and disease. These reviewers are instructed to consider previous tier scores, as well as the broader ways in which the research fits into the institute’s goals. Final funding decisions are made by institute center directors taking into consideration recommendations made at each stage of review.

In 2007, NIH initiated “Enhancing Peer Review,” a program to address concerns over the administrative burden of the review process, review quality, low funding rates among new investigators and a declining NIH budget (Erosheva et al. 2020; Fang and Casadevall 2009). Reviewers now provide scores from 1 (exceptional) to 9 (poor) based on five criteria: significance, investigator(s) background, innovation, approach and environment. Reviewers take each criterion into consideration to provide one overall impact score, weighing each criterion as they see fit. The average of preliminary impact scores determines if a proposal is selected for discussion by a review group, which calculates a final score for the institute panel. These changes aimed to improve transparency and decrease disparities by making review criteria less ambiguous.

Despite these changes, research documents continued disparities in NIH funding, prompting reflection on the part of the institute. Multiple studies document a persistent funding gap between Black and White PIs, with White PIs about twice as likely to receive funding (Ginther et al.
2018; Hayden 2015). These disparities have been attributed to multiple characteristics of the review process. Hoppe et al. (2019) found that 20% of the funding gap could be attributed to topic choice. Black scholars proposed research on topics with lower overall funding rates centered on community and population level health as opposed to more fundamental and mechanistic investigations. Other research also found that Black scholars were more likely to investigate health disparities and use research designs based on community and behavioral interventions, which are persistently underfunded areas of research (Carnethon et al. 2020). Additionally, Black scholars more commonly came from lower resourced institutions with fewer applications submitted overall and were more likely to be early-stage researchers (Hoppe et al. 2019). Erosheva et al. (2020) found that preliminary criterion scores account most for racial disparities, with Blacks averaging 0.350 points lower, which disadvantages them in the later review stages. Research also documents disparities based on gender, with women receiving more positive linguist comments, but lower numerical rankings and less overall funding, especially when considering the track record of PIs in the review process (Kaatz 2016; Magua et al. 2017; van der Lee and Ellemers 2015).

To address these continuing inequities, NIH has engaged in a systematic review of its practices and invested in programs to support diversity. In 2012, NIH invested $500 million in training and mentorship programs for minority scientists (Wilder et al. 2013; Valantine and Collins 2015). They also continue to study their own review practices (Pinholster 2016; Reardon 2011). In 2013, NIH formed a Diversity Working Group Subcommittee on Peer review. The group, made up of eight scholars with expertise in the social sciences, unconscious bias, stereotyping and faculty development, aimed to provide advice on interventions to reduce unconscious bias related to disparities in research awards. The panel instituted implicit bias and awareness trainings for review and program officers, conducted experiments to determine the effects of application anonymization (double-blind review) and analyzed successful versus unsuccessful grant applicants to spot trends in language used in reviews between races. While NIH has not published results from these experiments, other experimental studies indicate that double-blind review in academic journals show inconclusive results. Some arguing that single-blind review favors more well-known scholars from prestigious institutions (Tomkins et al. 2017), while others find no differences in the quality of the review process or equity of review outcomes (Chung et al. 2015; Cox and Montgomerie 2019).

Researchers studying disparities in NIH funding also suggest encouraging a more diverse applicant pool, targeting funds for topics underappreciated by reviewers, redefining scientific excellence to take into consideration engagement with professional organizations, public health influence in community and holistic definition of qualifications to include more diverse scholars and providing mentoring throughout the review process for early-stage researchers. Additionally, researchers suggest diversifying the pool of reviewers. As of 2019, only 2.4% of NIH reviewers were Black (Carnethon et al. 2020). A lack of diverse reviews perpetuates ingroup bias and favoritism for the status quo, continually disadvantaging researchers from underrepresented groups whose research commonly lays outside of reviewers’ areas of expertise (Hayden 2015). While the NIH has not published data that can identify the effects of any single intervention, overall, awards to African American/Black principal investigators has increased by 219% between 2013 and 2020, reducing the funding gap from 10% to 8%. Similarly, the funding rates for male and female scholars have equalized, however women still submit 55% fewer overall grant applications to the organization, thus compose significantly less of the final supported population of scholars.
6. Conclusion

Drawing on research about the advantages of diversity and inclusion for scientific excellence and models from organizations seeking to advance diversity, equity and inclusion in their workforces, the report concludes with four suggestions to address shortcomings for peer review in funding bodies.

Diversity on peer review panels

Incorporating diversity at all stages of the scientific process, including on peer review panels, supports innovation, as well as a wider range of scholars and research. Particularly given the expansion of science beyond the academy and clear disciplinary boundaries, review panels must consider the full range of stakeholders involved in the scientific process (Huutoniemi 2015; Langfeldt et al. 2020). Review panels that incorporate scholars from diverse disciplines who use a variety of approaches consistently fund more diverse research (Boudreau et al. 2016; Teplitskiy et al. 2018). Diverse groups, in terms of race, ethnicity and research background, are less likely to fall prey to “groupthink,” encouraging debate to counteract preformed preferences and biases (Antonio et al. 2004; Esarey 2017; Laudel 2006). Prioritizing different points of view encourages people to learn from each other, rather than hold to their beliefs and biases (Shore et al. 2011). Furthermore, incorporating diverse reviewers can combat stereotypes and elevate the status of underrepresented groups. For example, including women on peer review committees and in prestigious positions has been shown to affect women’s perceptions of adequacy and success (Squazzoni 2021), encouraging them to apply for opportunities and diminishing preformed judgements about other women (Faniko et al. 2020; Ellemers 2004). However, efforts to create more diverse review committees must be cognizant of equitably distributing service load, without overburdening women, people of color, junior, queer and working-class faculty members who often devote more time to mentorship and administrative tasks (Cardel et al. 2020; Social Sciences Feminist Network 2017).

Diversity coupled with inclusion

Within review panels, diversity must be coupled with inclusion to harness the benefits of diverse voices and the structural significance of empowering the perspectives of stigmatized groups. Structural or representational diversity, focused on matching the demographics of a group with a larger population, risks essentializing difference, tokenizing minority members and reinforcing, rather than combating stigma and bias (Kanter 1977; Smith-Doerr et al. 2017; Elsass and Graves 1997). Shallow discourses about diversity mask controversial discussions with positive language and fail to address structural issues by commodifying difference as multiculturalism or competitive advantage (Berry 2015; Bell and Hartmann 2007). Members from nondominant groups must be fully integrated to reap the benefits of information exchange and balanced power. This requires incorporating diversity as a “critical mass,” beyond a few symbolic members (Pfeffer 1983; Whittaker et al. 2015). Research suggests that it takes the combined voices of 25 percent of a group to shift dynamics and give weight to new perspectives (Centola et al. 2018). Furthermore, fully integrating diversity requires cultural changes that
recognize the value of diverse group members (Cheryan et al. 2017; Ellemers and Rink 2005; Weissmann et al. 2019). This includes engaging in processes of destigmatization, such as credibly and conclusively advocating for diverse members by those in high status and visible positions and pointing out the advantages of equity (Clair et al. 2016). Inclusion can also be practiced in groups by encouraging the sharing of information, participation in decision-making processes and expressing one’s viewpoint (Shore et al. 2011). Recognizing the perspective and contributions of underrepresented groups on panels contributes to advancing epistemic justice and equity by breaking reinforcing cycles of stigma and inequality (Abir-Am 2020; Misra et al. 2017).

**Review Criteria**

Incorporating diversity in review processes involves reevaluating review criteria. Bias can be amplified by both overly ambiguous criteria of evaluation, resulting in people filling in missing information using preformed beliefs, as well as narrow criteria that draw from attributes of groups currently in positions of power (Correll 2017). Combining more narrow criteria, such as methodology and research design, with opportunities for reviewers to express subjective opinions on the research, such as their agreement with the conclusions and originality of the topic under study, help avoid groupthink and the perpetuation of bias or incorrect assumptions (Guetzkow et al. 2004; Park et al. 2014). Rather than focus on averaging review scores, other ways of considering review comments, such as the range of opinions present, can be used as markers for creative potential and innovation (Azoulay and Li 2020; Gurthrie et al. 2018). This can help counteract tendencies towards conservative and support the work of underrepresented scholars less commonly identified with the mainstream of their fields. Specific funding programs with criteria geared toward innovation, novelty, early-career researchers, inter-disciplinary and non-mainstream work that allow flexible budgets and time horizons can also support more innovative, diverse and ground-breaking work while empowering a wider range of scholars (Laudel and Galser 2014).

Review formats can also include components of self-reflection to increase awareness about personal biases and minimize their impact. Studies show that a conscious acknowledgement of potential biases and a person’s positionality can encourage efforts to assess, monitor and disrupt bias in evaluation processes (FitzGerald et al 2019; Maxfield et al. 2020; Wong and Vinsky 2020). For example, participants in a lab experiment who learned about the tendency for people to exhibit implicit racial biases immediately before performing an implicit association test of bias showed less bias than groups that were not primed to think about biases in general before the task (Lai et al. 2016). Thus, review criteria can include questions that encourage reviewers to reflect on their own research, background and paradigms and how they may inform their reviews.

**Organizational Programs**

Diversity programs are most effective when they institute structural, rather than individual level change (Kalev et al. 2006; Stepan-Norris and Kerrissey 2016). Implicit bias training is the most common diversity and inclusion intervention, instituted on review panels at both NSF and NIH (Pinholster 2016). While implicit bias training can combat stereotyping and bias by slowing down cognitive processes to rethink preformed assumptions (Dupas et al. 2021;
Correll 2017; Moss-Racusin et al. 2014), critiques of implicit bias training argue that a focus on individual, internal and static cognitive processes do not address the full range of factors that shape and reinforce stereotypes, including cultural messages, organizational contexts and status hierarchies (Lamont et al. 2017; Nelson and Zippel 2021). Additionally, trainings can lead to backlash when they are presented as blaming individuals for structural inequalities, leading to practices that increase, rather than decrease equity and representation (Deschamps 2020; Kalev et al. 2006).

Establishing responsibility for diversity programs through the creation of diversity officers, leaders and accountability mechanisms better support the benefits from diversity trainings, networking and mentoring (Ellemers 2017; Stepan-Norris and Kerrissey 2016). Accountability structures include practices such as designating a specific taskforces or manager in charge of diversity programs, creating transparent lines of communication and regularly making available information on diversity programs and outcomes (Kalev et al. 2006). Additionally, creating structures that encourage reflexivity and program evaluation increase transparency and morale (Correll 2017). While the effects of diversity evaluations have not been directly assessed, work shows that evaluating managers decreased bias in assigning jobs (Kalev et al. 2006). Evaluations and setting targets help organizations understand areas that require attention and improvement, as well as show progress, enforce accountability and awareness about bias or inequities that otherwise dominant group members may refute (Handley et al. 2015; Stephans et al. 2021). In constructing evaluative measures, attention must be made to not build-in bias, which risks reinforcing stereotypes and naturalizing difference between groups (Correll 2017; Epstein 2007). To ensure equitable and meaningful evaluation criteria, criteria must be based on concrete skills, actions or results, rather than characteristics common among high status group members (Stephens et al. 2021).

Taken together, this research points to several concrete steps funding bodies may adopt to support diversity, equity and inclusion within academia and scientific excellence overall.

- Increase diversity on review panels, including a critical mass (~25%) of scholars from underrepresented groups, paying consideration to gender, race, topic of study, discipline, career stage and research institution

- Support practices of inclusion, such as group deliberation, to incorporate diverse members in the decision-making process

- Create limited evaluation criteria based on concrete skills and actions coupled with opportunities for more general comments

- Consider range of evaluation scores rather averages to identify promising work outside of the mainstream

- Create targeted and diverse funding streams to support under-funded topics, disciplines or groups and/or research seen as high risk

- Incorporate self-reflection about biases as part of the review process
- Establish accountability mechanisms through the creation of designated groups or managers in charge of diversity initiatives and providing transparent communication with relevant parties

- Continuously self-evaluate review processes and outcomes

As review panels adapt their practices, knowledge of the effects of diversity, equity and inclusion programs would be enhanced with more comparative work on how different interventions impact review and research outcomes. Additionally, while current suggestions focus on changing practices to support more inter-disciplinary and potentially high-risk research, it is likely that a diversity of review strategies and programs will best support a diversity of research once currently under-funded areas become more robust. Lastly, current research has not addressed the implications of expanding peer review beyond the realm of academia. This could include incorporating stakeholders from professional, policy, corporate or community arenas, who have insights about the implications of scholarship, particularly beyond intellectual merit. As the boundaries between science and society continually adjust, so too must the ideals of organized skepticism to make science that is seen as transparent, inclusive and productive to broader experts and audiences.

References


Shah, Nihar B. 2021. “Systemic Challenges and Solutions on Bias and Unfairness in Peer Review (Draft).”


