SCIENTIFIC COMMUNITY

Topic choice contributes to the lower rate of NIH awards to African-American/black scientists

Travis A. Hoppe[1,2], Aviva Litovitz[1,2], Kristine A. Willis[3]*, Rebecca A. Meseroll[1,2], Matthew J. Perkins[1,2], B. Ian Hutchins[1,2], Alison F. Davis[4], Michael S. Lauer[5], Hannah A. Valantine[4], James M. Anderson[2], George M. Santangelo[1,2†]

Despite efforts to promote diversity in the biomedical workforce, there remains a lower rate of funding of National Institutes of Health R01 applications submitted by African-American/black (AA/B) scientists relative to white scientists. To identify underlying causes of this funding gap, we analyzed six stages of the application process from 2011 to 2015 and found that disparate outcomes arise at three of the six: decision to discuss, impact score assignment, and a previously unstudied stage, topic choice. Notably, AA/B applicants tend to propose research on topics with lower award rates. These topics include research at the community and population level, as opposed to more fundamental and mechanistic investigations; the latter tend to have higher award rates. Topic choice alone accounts for over 20% of the funding gap after controlling for multiple variables, including the applicant's prior achievements. Our findings can be used to inform interventions designed to close the funding gap.

INTRODUCTION

Despite ongoing efforts at the National Institutes of Health (NIH) to promote a diverse biomedical workforce (*1*, *2*), a 2011 study showed that applications from African-American/black (AA/B) scientists were significantly less likely to receive an R01 award than those submitted by white (WH) scientists, even after controlling for educational background, country of origin, training, previous research awards, and employer characteristics (*3*). Especially concerning was the finding that typical measures of scientific achievement (e.g., NIH-funded training, previous grants, publications, and citations) did not translate into an equal probability of funding across racial/ethnic groups, highlighting the need for further study to guide interventions aimed at closing the funding gap. No significant funding gap for applications from Hispanic scientists or women was identified by the 2011 study; however, a more recent study disaggregating race and gender showed that applications from African-American and Asian-American women were less likely to receive R01 awards, underscoring the possibility of an additive effect for women of color (*4*). These studies raised important questions about fairness in peer review because most of the funding gap for AA/B applicants remained unexplained. Here, we seek to answer those questions by examining the characteristics of applications submitted by AA/B and WH scientists.

The underlying causes of the funding gap have been difficult to identify, in large part because of the complex and multifaceted nature of the application and review process. To address this challenge, we identified six decision points at which differential outcomes might contribute to an overall difference in funding: how frequently applicants submit, whether an application was chosen for discussion by a study section, reviewer-assigned impact scores of discussed applications, final funding decisions made by NIH institutes and centers (ICs), resubmission if the application was not funded, and a previously unstudied factor—choice of topic. An analysis of both new (Type 1) and renewal (Type 2) R01 applications (N = 157,549; attributes summarized in table S1) shows that, although the award rate has dropped for all applicants over the past decade, the funding rate for WH scientists remains approximately 1.7-fold higher than for AA/B scientists [16.1% AA/B versus 29.3% WH in fiscal year (FY) 2000–2006 (*3*) and 10.7% AA/B versus 17.7% WH in FY 2011–2015; Fig. 1].

Complex problems such as this are frequently studied with multivariate regression analysis, which can account for the effect of many independent variables on a single dependent variable. However, interpreting multivariate regression data can be challenging. When one independent variable acts both directly on the outcome and indirectly on another variable, when variables presumed to be independent are highly correlated, or when two or more variables interact with each other in a feedback loop, it can be difficult to decipher which factors make the most significant contributions to an outcome. In addition, real-world data may not provide sufficient power to calculate statistical interactions when a large number of variables act on a relatively small population. For these reasons, we first did simple descriptive analyses to characterize each of our six decision points independently before using multivariate regression analysis to determine how the relevant variables might be interrelated.

RESULTS

Career stage and institutional resources influence the gap in the number of submissions by AA/B and WH scientists

One factor that might be expected to influence whether a scientist receives funding is how many applications he or she submits. From FY 2011–2015, AA/B scientists submitted R01 applications at 83.7% the frequency of WH applicants (Fig. 1 and fig. S1). However, AA/B applicants are unevenly distributed across institutional funding quintiles; 33.9% of all AA/B investigators are from institutions in the lowest quintile, compared with only

[1]Office of Portfolio Analysis, National Institutes of Health, Bethesda, MD, USA. [2]Division of Program Coordination, Planning, and Strategic Initiatives, National Institutes of Health, Bethesda, MD, USA. [3]National Institute of General Medical Sciences, National Institutes of Health, Bethesda, MD, USA. [4]Scientific Workforce Diversity, National Institutes of Health, Bethesda, MD, USA. [5]Office of Extramural Research, National Institutes of Health, Bethesda, MD, USA.
*Present address: National Cancer Institutes, National Institutes of Health, Rockville, Maryland, USA.
†Corresponding author. Email: george.santangelo@nih.gov

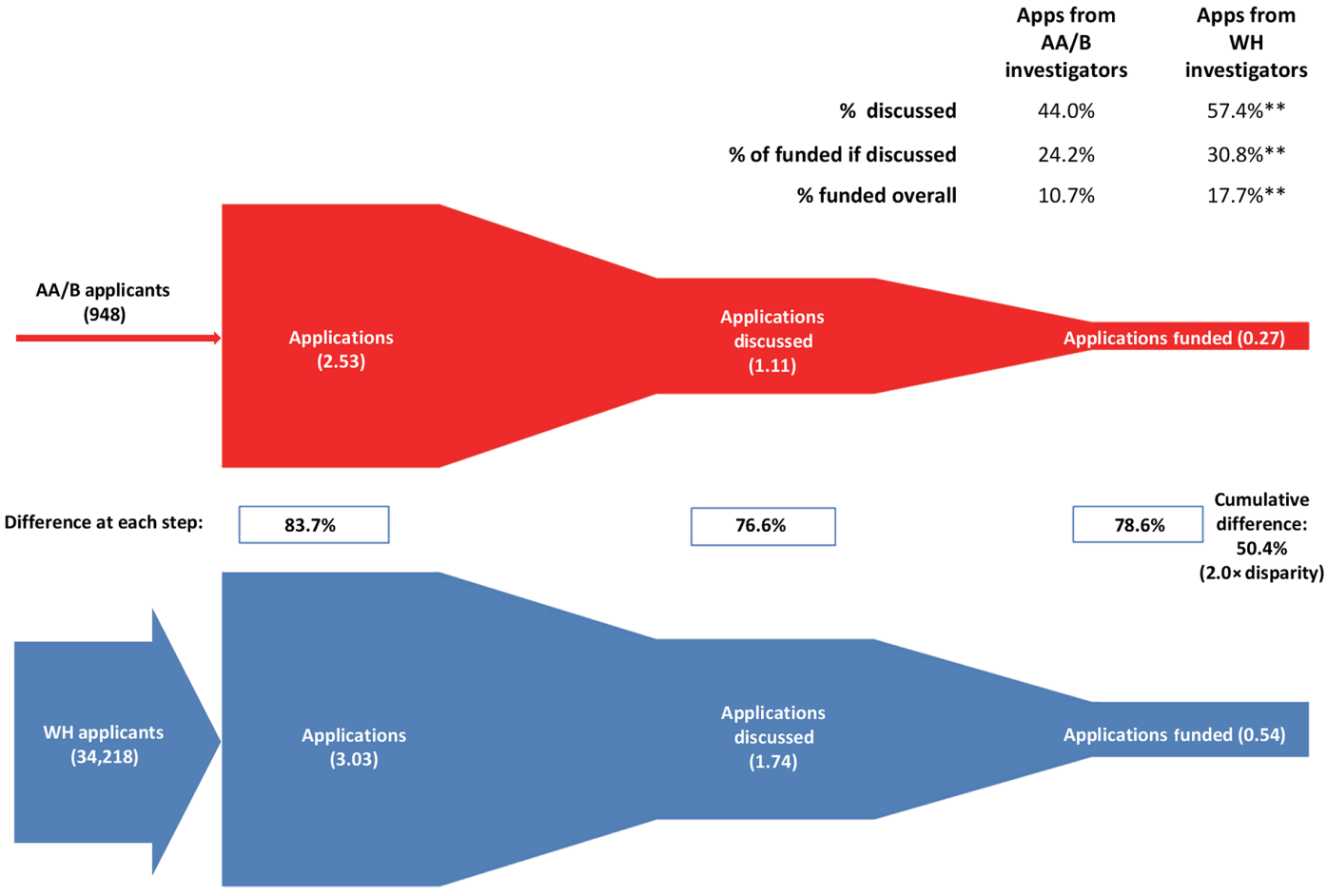

Fig. 1. Funding gap between AA/B and WH scientists at each stage of the R01 application and review process. Arrows on the left indicate the number of AA/B and WH R01 applicants in FY 2011–2015. The total number of applicants with a reported race/ethnicity is 45,998. Rocket charts depict the number of applications that were submitted, discussed, and funded per applicant. Comparative rates of discussion, funding of discussed applications, and overall funding rates are presented on the top right (**$P < 0.01$).

22.0% of WH investigators ($P < 0.0001$; table S2 and fig. S2). Irrespective of race, scientists from these lower resourced institutions submit fewer applications overall (2.97 versus 2.33 from scientists at institutions in the highest quintile of NIH funding, $P < 0.0001$). Furthermore, and consistent with a recent study (*5*), we found that applications from AA/B scientists were nearly twice as likely as those from WH scientists to be submitted by new investigators (defined as applicants who had not yet been awarded a competing NIH research grant; 47.4% versus 24.9%, respectively, $P < 0.01$; table S3). This difference is attributable to differences in the career age of AA/B and WH applicants, with AA/B applicants more likely to have completed their terminal degree in the past 15 years (fig. S3). AA/B applicants were also more likely to be early-stage investigators (ESIs) and therefore subject to special funding policies (*6*). Overall, scientists early in their career submit fewer applications than those with more experience (an average of 1.83 applications per person for ESIs over the period of our study versus 2.70 or 3.05 for investigators who are more than 10 or 20 years postdegree, respectively, $P < 0.0001$). Controlling for institutional resources and career age reduces the gap in the number of submissions by 39%; when additional variables are added, the difference is no longer statistically significant (see below).

Applications from AA/B scientists are less likely to be discussed and receive lower impact scores

After submission, R01 applications are reviewed following assignment to study sections, which are composed of subject matter experts (SMEs) recruited from the scientific community. Each application is typically assigned to three reviewers who provide the initial critiques used to inform which applications will be discussed and scored by the full study section. Following established policy, only the top-ranked 55% of applications assigned to each study section were discussed during FY 2011–2015 (table S1). In this time frame, applications from AA/B scientists were discussed 76.6% as frequently as those from WH applicants (Fig. 1).

If an application is discussed in the study section, it receives an impact score—a numerical representation of the application's scientific and technical merit as assessed by the reviewers. Impact score values range from 10 (high impact) to 90 (low impact) (*7*). At the discussion stage, applications from AA/B scientists receive poorer overall impact scores on average than those of WH scientists (38.4 ± 13.4 SD and 35.2 ± 12.6 SD, respectively, $P < 0.0001$; table S4). Cumulatively, the lower submission rates, lower average discussion rates, and lower impact scores result in applications from AA/B scientists receiving R01 funding at approximately half the rate (0.5-fold) of those from WH scientists (Fig. 1).

IC decisions do not contribute to funding gap

After study sections have provided scores for discussed applications and advisory councils have offered their recommendations, final funding decisions are made by IC directors with input from their program staff. In addition to impact score and advisory council recommendations, ICs consider a variety of other factors when making funding decisions, including public health burden, opportunities for scientific progress, and overall portfolio balance. Funds are generally awarded to applications with impact scores below a given percentile, which can differ both year to year and between ICs based on available funds; however, if an application is of particular relevance to the funding IC, it may still be awarded even if its score is above the typical percentile-based payline (i.e., the application receives discretionary funding). A higher fraction of applications from WH scientists received impact scores in the percentile range correlating with likely funding (Fig. 2A). However, below the 15th percentile, there was no difference in the average rate at which ICs funded each group (Table 1); applications from AA/B and WH scientists that scored in the 15th to 24th percentile range, which was just above the nominal payline for FY 2011–2015, were funded at similar rates (AA/B 25.2% versus WH 26.6%, $P = 0.76$; Table 1). The differences we observe at narrower percentile ranges (15 to 19, 20 to 24, 25 to 29, and 30 to 34) slightly favored either AA/B or WH applicants alternately but were in no case statistically significant ($P \geq 0.13$ for all ranges). These results suggest that final funding decisions by ICs, whether based on impact scores or discretionary funding decisions, do not contribute to the funding gap.

AA/B investigators are not less likely to resubmit an unfunded application

Previous analyses have shown that resubmitted R01 applications are more likely to be discussed and awarded than new applications (*8*), that initial impact score and resubmission are correlated (*9*), and that AA/B R01 applicants are less likely to revise and resubmit unfunded applications (*3*, *8*). Our analysis confirms that unfunded new (Type 1) R01 applications from AA/B scientists are revised and resubmitted less frequently than those from WH scientists (AA/B 36.8% versus WH 43.3% for FY 2011–2015). Note that this difference cannot be explained by hypothetical future resubmissions not present in our dataset (i.e., censoring); applicants who wish to resubmit unfunded applications are required to do so within 37 months of the original application (*10*, *11*), and analysis of actual resubmissions rates shows that 98% are received within 24 months (fig. S4).

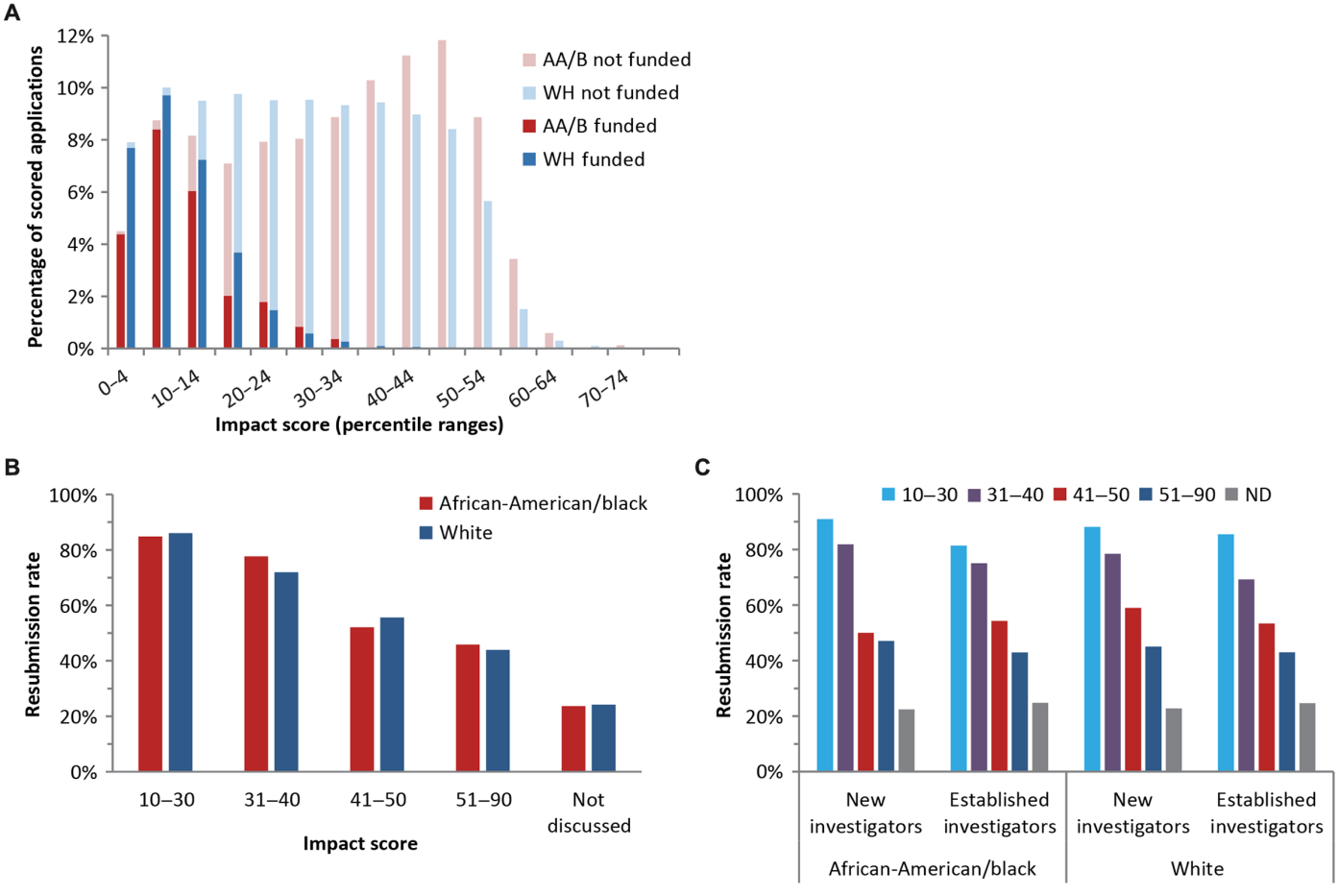

Fig. 2. Effect of impact score on discretionary funding and resubmission rates. (**A**) The distribution of percentile scores for funded and unfunded Types 1 and 2 R01 applications submitted by AA/B (red bars) and WH scientists (blue bars). (**B**) Resubmission rates by impact score range for unfunded, unsolicited Type 1 R01 applications (FY 2011–2015) from AA/B and WH applicants and (**C**) AA/B and WH applicants by career stage. ND indicates applications that were not discussed and therefore not scored. All pairwise comparisons between resubmission rates for AA/B and WH applicants within each impact score range in (B) and (C) are not statistically significant ($P > 0.07$).

Table 1. Effect of percentile score on award rate. Percentage of applications funded for each percentile range. n/a indicates that there were no applications in the given range.

Percentile range	% AA/B funded	% WH funded
0–4	97.4%	97.2%
5–9	95.9%	97.0%
10–14	73.9%	76.1%
15–19	28.3%	37.6%
20–24	22.4%	15.4%
25–29	10.3%	6.0%
30–34	4.0%	2.8%
35–39	0.0%	0.9%
40–44	0.0%	0.6%
45–49	0.0%	0.4%
50–54	0.0%	0.2%
55–59	0.0%	0.2%
60–64	0.0%	0.0%
65–69	n/a	0.0%
70–74	0.0%	0.0%
75–79	n/a	0.0%
80–84	0.0%	0.0%
85–89	n/a	0.0%

The finding that AA/B applicants appear to be less likely to revise and resubmit has important policy implications. However, the 2011 study that originally reported this result did not control for the influence of impact score on resubmission (*3*, *9*). After doing so, we found that there is no statistically significant difference in resubmission rates (Fig. 2B). AA/B and WH scientists who received lower (more favorable) impact scores (10 to 40) resubmitted applications at approximately the same rate; in the 41 to 50 (less favorable) impact score range, AA/B scientists were less likely to resubmit than WH scientists, but this difference is not statistically significant (Fig. 2B). Controlling for an applicant's prior funding history does not change this result; unfunded applications from AA/B scientists were not statistically less likely to be resubmitted than were those from WH scientists, regardless of whether the applicants were new investigators or established scientists who previously held an NIH research award (new: AA/B 50.0% versus WH 59.8%, $P = 0.16$; established: AA/B 56.7% versus WH 53.7%, $P = 0.85$; Fig. 2C). Although applicant-specific decisions about whether to resubmit may be an area for targeted intervention, those decisions do not contribute significantly to the gap in funding applications from AA/B or WH investigators.

Choice of the topic of study contributes to differences in funding outcomes

An understudied aspect of the R01 application process is the degree of correlation, if any, between funding outcomes and the topics that scientists propose to investigate. Since the R01 mechanism allows scientists to request support for research in their area(s) of interest (i.e., the projects are investigator initiated), these applications provide a unique window on the priorities of applicants and reviewers. To examine how topic choice might relate to funding outcomes in general, and to the gap in funding for AA/B investigators in particular, we used word2vec (*12*), an informatics approach that uses word embedding of text to build document vectors suitable for grouping applications into clusters based on the similarity of their content (see Materials and Methods for details). We used word2vec to divide the 157,549 R01 applications in our dataset (Types 1 and 2, FY 2011–2015) into 150 topic-based clusters. After testing a variety of options, the choice of 150 clusters seemed optimal, both because it is roughly equivalent to the number of standing study sections and because the resulting areas of science were well defined (see Materials and Methods). Within-cluster and within-study section variance in percentile scores are similar to each other but significantly lower than overall variance ($P < 0.0001$; fig. S5A), indicating that word2vec clusters are at least as cohesive as standing study sections. To confirm the cohesiveness of word2vec clusters, we compared the assignments made by the algorithm to the opinions of SMEs. Presented with 10 sets of 10 applications, representing increasing degrees of semantic overlap, SMEs reproduced the groupings generated by the computational method 97.6% of the time, indicating a very high degree of correlation between word2vec and human judgment (table S5).

Comparison of word2vec assignments to the 166 study sections administered continuously from FY 2011–2015 by the NIH Center for Scientific Review (CSR) reveals the lack of a one-to-one correspondence between study sections and scientific topics. Applications with similar content are assigned to multiple study sections, ranging from 1 to 49 study sections per topic (fig. S5B); conversely, between 1 and 27 topics are reviewed in any given study section (fig. S5C). While perhaps initially unexpected, this result is not difficult to interpret, since study sections are designed to provide complementary perspectives on areas of science that span a variety of related fields. For example, a study section with expertise in intestinal epithelial biology may review applications that span topics as diverse as basic cell biology, intestinal infections, and inflammatory bowel diseases.

This lack of a one-to-one correlation between study sections and scientific topics raised the possibility of unequal funding rates, meaning that, independent of the study section in which the reviews are conducted, some topics might be favored and others disfavored. Topic-based inequality in funding rates would not be controlled for by the use of percentile rankings, which are designed to normalize study sections that in the aggregate tend to score applications more harshly with those that tend to score more generously. Notably, the award rates of cluster-defined topics varied from a minimum of 7.5% to a maximum of 28.7% (see below). At both ends of this distribution, 56 clusters have an award rate that differs significantly from average (25 high and 31 low, $P < 0.01$; table S6). These results demonstrate the existence of topic preference, meaning that different topics are accorded different levels of acceptance and/or enthusiasm, which may reflect shared, broadly held views on the relative scientific value of different areas of research.

The discovery of topic preference next led us to ask whether AA/B and WH applicants tend as groups to study the same or different topics. As one high-level indicator of topic choice, we looked at applications that propose studies using human subjects, animal subjects, both, or neither and found very different results for AA/B scientists compared with scientists of other racial/ethnic groups (table S1). Applications from AA/B scientists were significantly more likely to involve human subjects (49.8% for AA/B versus 31.8% for WH applicants, $P < 0.0001$). Furthermore, they were distributed differently across study sections [$P < 0.001$, with 10 of 10 standing study sections

that receive the greatest number of applications from AA/B scientists falling under the Division of AIDS, Behavioral, and Population Sciences, versus 5 of 10 for WH applicants]. Together, these two observations suggest that AA/B scientists may be proposing to study a different distribution of topics than other applicants.

To test this hypothesis, we next mapped the applications from AA/B scientists onto our 150 word2vec clusters and found that they were highly skewed (Fig. 3, A and B). Notably, 37.5% of all applications from AA/B scientists mapped to only 8 of the 150 topic clusters (compared to a random distribution, $P < 0.0001$). Of those eight clusters, six had award rates that were significantly below the NIH average (table S6). There was therefore a trend among AA/B applicants to submit applications on topics that experience lower funding rates, irrespective of the study section to which they were assigned (Fig. 3C and fig. S6). WH applicants also experienced lower award rates in these clusters, but the disparate outcomes between AA/B and WH applicants remained, regardless of whether the topic was among the higher- or lower-success clusters (fig. S6).

The marked skew in topic choice by AA/B applicants led us to investigate whether those areas of science share commonalities or are instead broadly distributed across the biomedical landscape. This is most easily visualized by generating word clouds of each of the eight clusters identified in Fig. 3A. Consistent with the more frequent use of human subjects (table S1), applications from AA/B scientists tend to describe research on health disparities and patient-focused interventions (Fig. 4A). Defining words in the eight clusters with the highest percentage of applications from AA/B applicants include socioeconomic, health care, disparity, lifestyle, psychosocial, adolescent, and risk; these clusters had funding levels ranging from 11.2 to 17.2% (table S7). In contrast, frequently used words in the eight clusters without any AA/B applicants (see Fig. 3A) include osteoarthritis, cartilage, prion, corneal, skin, iron, and neuron; these clusters had funding levels ranging from 12.5 to 28.7% (Fig. 4B and table S7). For all applicants, the cluster with the lowest award rate, 7.5%, is characterized by the words ovary, fertility, and reproductive, while the cluster with the highest award rate, 28.7%, is characterized by the words odor, olfactory, and chemosensory. Topics that focus on fundamental and mechanistic questions are distributed across the entire range of award rates (table S6).

Grouping related topic clusters in a network markedly illustrates the tendency of applications from AA/B scientists to focus on disease prevention and intervention, much more than on any other area of science (Fig. 4C). It should be noted that the largest circles in Fig. 4C represent the clusters with the largest number of applications from AA/B scientists, which are not equivalent to the clusters with the largest overall number of applications. Furthermore, there is no correlation between award rate and field size, as represented by the total number of applications in a cluster (fig. S7). This is directly explained by three factors: applications are assigned percentile scores relative to others in their study section, CSR limits the number of applications that a standing study section can consider, and any individual topic is spread across multiple study sections. Therefore, higher or lower award rates can only occur if a topic is systematically favored or disfavored, respectively, across multiple study sections.

It could be argued that study sections are able to appropriately discern that some topics receive fewer proposals that will go on to produce influential results and give scores beyond the nominal payline to those applications predicted to be subpar. This would be unexpected, since numerous analyses have shown that study section–assigned scores do not discriminate between grants that go on to produce work of higher versus lower influence (*13*–*19*). A reanalysis of data from the sole report claiming otherwise shows that, in fact, only 1% of the observed variance in the number and influence of papers produced by a funded grant can be accounted for by percentile ranking, confirming that the correlation between reviewer judgment and project outcomes is poor (*14*, *20*). If reviewers are unable to make accurate predictions at the level of individual grants, then it seems unlikely that they would be able to do so for particular topics. To rule out the possibility that separating applications into topic areas reveals a previously unidentified predictive power of percentile score, we asked whether publications resulting from R01 awards in higher- and lower-success clusters differ in their scientific influence, as measured either by the Relative Citation Ratio [RCR; an article-level metric that measures the influence of an individual publication relative to its cocitation network (*21*)] or by the number of raw citations they receive per year. We found that the lowest-success topics produced papers that were typically more influential (higher median RCR) than those from the highest-success topics (fig. S8). Furthermore, for topics in either the highest or lowest quintile of award rates, plotting the percentile score of each award against the median RCR of all papers it produced shows the complete absence of a correlation between study section assessment and future productivity (fig. S9).

Multivariate analysis of factors influencing the funding gap

The above analyses make definite predictions about the underlying causes of the funding gap between WH and AA/B applicants. However, the NIH application and award process is complex, and at least some of the factors that contribute to success are likely to be correlated. We therefore used multivariate regression to analyze the relationship between race and award rate while controlling for confounding variables (see Materials and Methods for full details). For each investigator, we controlled for the FY of the application, whether the application was a resubmission, number of years since the applicant's last degree, ESI status, number of past applications, and number of past awards. We also controlled for evidence of past accomplishment on the part of the applicant by including both the median RCR for papers listed in the biosketch and the number of those papers that fall in the top decile of RCR values (*21*). To account for differences in research environment, we included controls for type of applicant organization (e.g., research institute versus degree-granting university), geographical region, and institutional resources. When considering binary outcomes, such as whether an application is discussed or awarded, we used probit models estimated with maximum likelihood. Since these models are nonlinear, we calculated the difference between AA/B and WH applicants as average marginal effects (AMEs). In this case, the marginal effects measure the change in probability of a given event as a function of the change in a particular explanatory variable while holding all other covariates constant. Individual-level marginal effects are averaged to provide a population-level metric.

In total, our model accounts for 42% of the observed difference in the rates at which AA/B and WH scientists receive funding (see the Supplementary Materials for full details of regression analyses). It confirms our earlier finding that AA/B scientists are not statistically less likely to resubmit a previous application in response to similar percentile scores (1.6 percentage points difference, $P = 0.551$). When percentile score is not taken into consideration, applications

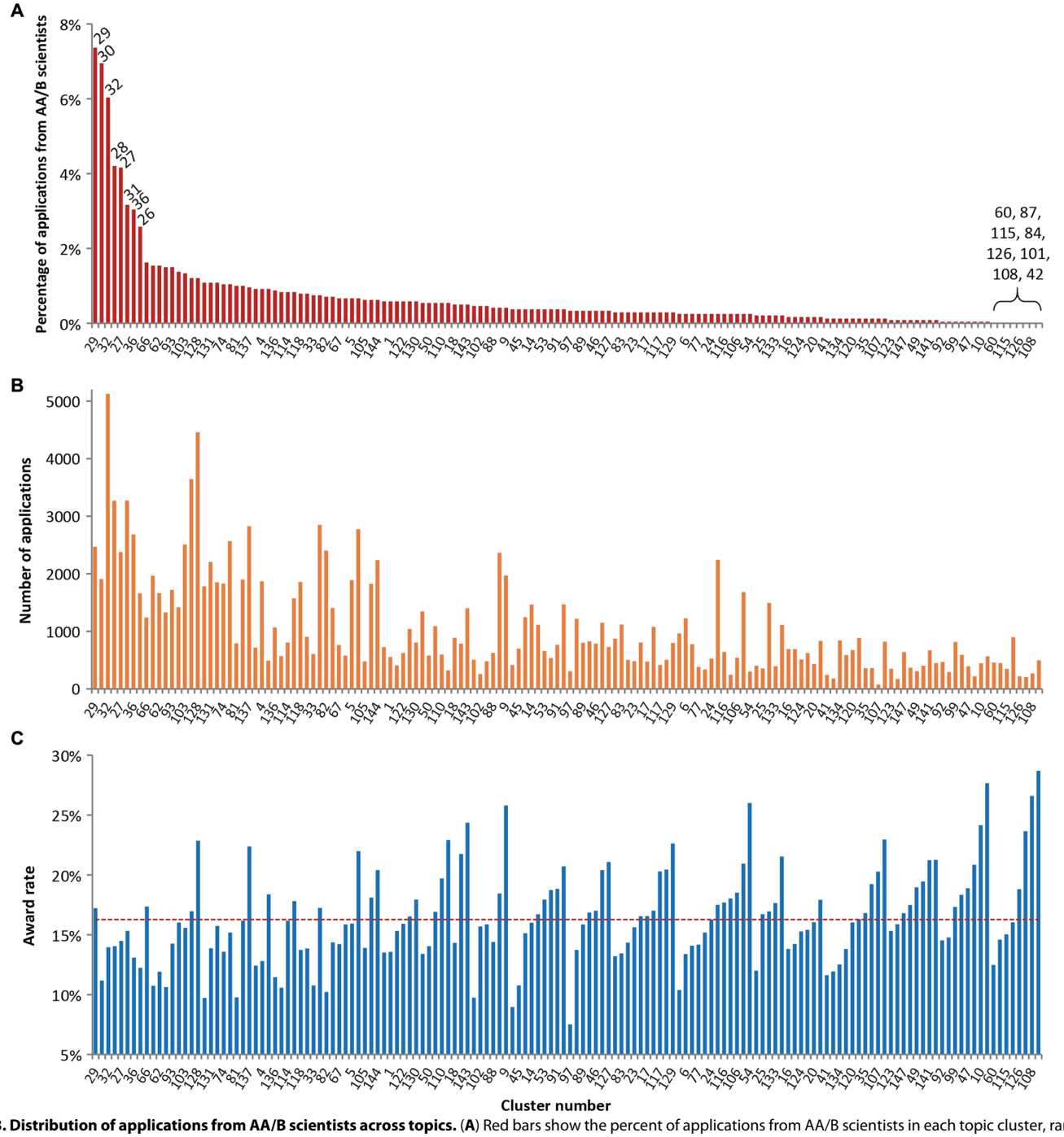

Fig. 3. Distribution of applications from AA/B scientists across topics. (**A**) Red bars show the percent of applications from AA/B scientists in each topic cluster, ranked from highest to lowest. Clusters were initially defined based on content similarity; thus, clusters that are numerically close also tend to have relatively similar content. Of all applications from AA/B scientists, 37.5% belong to the first eight clusters; at the other end of the distribution, eight clusters contain no applications from AA/B scientists. Because of space constraints, every other cluster number is reported on the *x* axis; cluster numbers for the first and last eight clusters are highlighted on the graph. (**B**) Number of applications in (orange bars) and (**C**) award rate for (blue bars) each topic cluster, ranked by percentage of applications from AA/B scientists in each cluster [i.e., same ranking as in (A)]. The dashed red line represents the overall R01 award rate (16.3%). In the 25 clusters with a significantly above average award rate (see table S6), the number of applications from AA/B scientists was too small to determine how they fared relative to applications from WH scientists.

from AA/B scientists appear to be 7.3 percentage points less likely to be resubmitted ($P < 0.001$). It also shows that after sufficiently controlling for applicant and organizational variables, the racial difference in number of applications submitted becomes statistically insignificant (0.03 applications, $P = 0.643$).

The decision point that makes the largest single contribution to the funding gap is the selection of applications for discussion. WH applicants have a 54% probability of being discussed; relative to this value, the AME of race corresponds to an 8.2 percentage point reduction in the likelihood of discussion for AA/B applicants ($P < 0.001$).

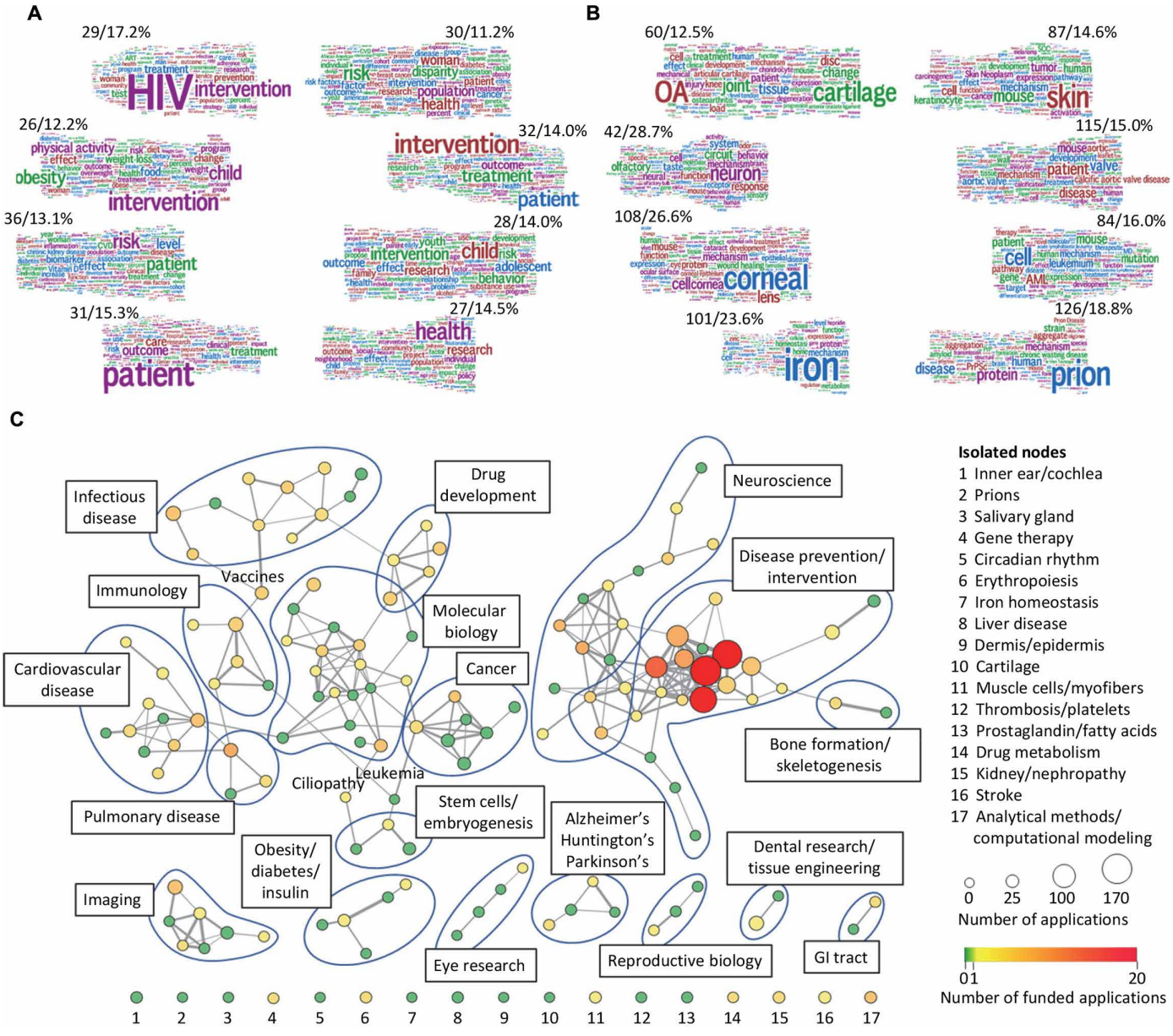

Fig. 4. Topics most and least commonly proposed by AA/B scientists. (**A**) Topic clusters with the highest percentage of applications from AA/B scientists. (**B**) Topic clusters with no applications from AA/B scientists. Word clouds are placed in a clockwise orientation relative to the order shown in Fig. 3A. Cluster numbers are presented alongside overall award rate (cluster number/award rate). (**C**) Distribution of applications and awards for AA/B scientists across topics in the NIH portfolio. Each node in the network represents a topic cluster, and related topic clusters are grouped inside blue borders and labeled (rectangles). Node size correlates with the number of applications from AA/B scientists, and nodes are heat mapped by the number of funded applications from AA/B scientists in each cluster. GI, gastrointestinal.

At this stage, the inclusion of topic choice as a variable in the model makes a relatively small contribution (8.2 versus 8.6 percentage point reduction, $P = 0.03$). In contrast, among applications that reach the discussion stage, controlling for topic choice substantially changes the probability of award for AA/B relative to WH scientists. After discussion, the probability of award for WH applicants is 28%; including topic choice reduces the AME of race on the likelihood of award from 3.2 to 2.5 percentage points, narrowing the funding gap by 21.1% ($P = 0.005$; table S8).

Once an application reaches the point of being discussed, only the relative influence of publications listed by applicants in their biosketch has a larger effect than topic choice on the gap in award rates (26.1% versus 21.1%, respectively; table S9). This result is consistent with a recent study that found publication history, as reported in an applicant's biosketch, to be a significant factor contributing to the funding gap between AA/B and WH scientists (*22*). Controlling for the number of prior applications and awards reduces the funding gap by 20.3%. Together, these three factors account for 43.2% of the modeled difference at this stage. That prior applications and awards act as important drivers of the funding gap is consistent with the well-known "Matthew effect," which describes how past success determines future success in a manner that cannot be ascribed solely

to differences in merit (*23*, *24*). More recent work has demonstrated that the diffusion of ideas and accumulation of influence are driven heavily by indicators of prestige, as opposed to relying exclusively on quality (*25*–*27*). Viewed together, our data lead us to speculate that the funding gap between AA/B and WH scientists may be driven by a vicious cycle, beginning with AA/B investigators' preference in the aggregate for topics less likely to excite the enthusiasm of the scientific community, leading to a lower probability of award, which in turn limits resources and decreases the odds of securing funding in the future. Mathematical modeling of the NIH review process has found that subtle depressions in score—the equivalent of a three-quarter point reduction on a scale of 1 to 9 by the three reviewers who provide the initial critiques used to inform which applications will be discussed—are sufficient to substantially bias the number of funded applications in favor of a preferred class of investigators (*28*).

To examine this possibility further, we used RCR (*21*) to compare the influence of papers listed by AA/B and WH new investigators in their biosketches with the influence of papers produced by their first award. Like all other metrics that rely on citation counts, RCR is a measure of influence and is therefore a good proxy for prestige. Although influential work is often valuable, it is erroneous to equate either influence or prestige with impact, importance, or quality (*29*). Prior to award, new AA/B investigators had fewer papers in the top decile of RCR values and a lower median RCR than new WH investigators. Following receipt of an award based on those biosketches, the gap in median RCR decreased and the gap in the number of top decile papers closed significantly ($P = 0.002$; fig. S10). Therefore, given equal opportunity, AA/B new investigators quickly reduce the gap in production of influential papers relative to WH new investigators.

DISCUSSION

Of the six initial decision points we chose to study in the NIH R01 application pipeline, three make a significant contribution to the funding gap between AA/B and WH applicants. First, applications from AA/B scientists are less likely to be discussed by study sections. Second, when discussed, they receive poorer impact scores. CSR is testing one possible cause—implicit bias in peer review—by anonymizing applications from AA/B and WH applicants (*30*). A recent study did not detect any evidence of race or gender bias in the initial evaluation of R01 applications by reviewers, but the sample size was small ($n = 48$), limited to funded proposals, and did not cover the full range of topics supported by the NIH; only applications from the National Cancer Institute, National Institute of General Medical Sciences (NIGMS), National Heart, Lung, and Blood Institute, and National Institute of Allergy and Infectious Diseases were included (*31*). Another possible cause, the Matthew effect, is not mutually exclusive with implicit bias and may reflect a broader challenge faced by peer review in meeting the goal of identifying the most meritorious applications. The previously reported effect of resubmission is attributable to the fact that applications from AA/B applicants are less likely than those from WH applicants to be discussed (*3*); the better the score they receive, the more likely scientists of both races are to resubmit. There is also no evidence that IC score–based or discretionary funding decisions correlate with an applicant's race. Last, we discovered that once an application reaches the point of being discussed, controlling for topic choice reduces the funding gap by 21%. Given the complexity of the application and review process, this is a considerable contribution for a single decision point. Reviewers seem to prefer certain topics over others, despite the fact that awards in lower-success clusters did not produce less influential science than those in higher-success clusters. This observation deserves further investigation, especially since it appears to be a prevalent feature of the NIH peer review process.

Our analysis shows that all three of the factors that underlie the funding gap—preference for some topics over others, assignment of poorer scores, and decision to discuss an application—revolve around decisions made by reviewers. It is generally accepted that decision-making processes can be influenced by the people who are invited to participate in them; for example, it has been observed that including at least one woman on the committee that organizes an academic symposium both correlates with a significantly higher proportion of invited female speakers and reduces the likelihood of an all-male roster (*32*). We therefore examined the racial composition of the study sections that reviewed the R01 applications in our dataset. We found that 2.4% of reviewers were AA/B scientists (table S10), which is very similar to the percentage of applicants who are AA/B (2.1%; Fig. 1). While not underrepresented relative to applicants, the absolute number of AA/B reviewers is still quite small, and it is conceivable that a more demographically diverse group of reviewers might have different opinions on the significance of some grant applications.

Together, our findings point to the salient factors for which targeted interventions could be considered in future attempts to address the funding gap. The first and most fundamental of these is to encourage a more diverse applicant pool. As has been previously observed (*8*), our data show that there is a marked difference in the number of AA/B and WH applicants for NIH funding (Fig. 1). Mathematical modeling indicates that this discrepancy is not due to an insignificant number of AA/B Ph.D. graduates, but rather to a dearth of postdoctoral fellows transitioning into faculty positions (*33*). Targeted funding opportunities such as the NIGMS MOSAIC program (*34*), which is designed to enhance postdoctoral career transitions to promote faculty diversity in the biomedical research workforce, may help address this. The next level of intervention is to develop and implement mentoring programs that provide all new and ESIs with quality guidance on navigating the NIH system. Not only does this have the potential to make the application process more fair and meritocratic overall, but it may also be of particular benefit to AA/B applicants, a higher proportion of whom have not yet received a major NIH award (table S3). Last, our data suggest that ICs may wish to consider establishing a policy that directs discretionary funding to meritorious applications on topics that are underappreciated by review but align well with their strategic priorities. In combination, the active management of these three phases of the grant life cycle—upstream of the application process, at the time submissions are being prepared, and after review—may help NIH move closer to its goal of a diverse workforce.

MATERIALS AND METHODS

Data sources

Data analyzed in this study were extracted from the Information for Management, Planning, Analysis, and Coordination (IMPAC II) database, which is used by the NIH staff to track and manage research grants and contracts. These data are publicly available through the NIH Commons (https://era.nih.gov/), except for personal identifying information, including race, ethnicity, and sex of applicants,

per NIH policy. The data used in this study include application texts, demographics of the principal investigator (PI), impact and percentile scores, whether an application was discussed by a study section, and the ultimate funding decision. We identified whether applicants were new investigators using the new investigator flag in IMPAC II and considered all other applicants as established investigators. With the exception of the resubmission data in Fig. 2 and the multivariate regression analysis, which were limited to Type 1 (new) R01s, all other analyses considered both Type 1 and Type 2 (renewal) R01s submitted between FY 2011 and 2015. For descriptive analyses, race of the contact PI was used to group applications. For multivariate regression, multi-PI applications were excluded. To visually represent the differing proportions of applications submitted, discussed, and funded for AA/B and WH applicants, we produced a flow diagram that we termed a "rocket chart" because of its shape (see more details in the "Rocket charts" section below).

To analyze the effect of the initial impact score on the funding gap, we extracted the overall impact score data from IMPAC II. Impact and percentile scores were only assigned to discussed applications and were available for 100 and 91%, respectively, of discussed R01 applications.

Rocket charts

We created rocket charts in Excel by producing two lines to identify the outline of each rocket (for example, with the thickness of the rocket scaled to the number of applications per applicant). The data we gathered for the group of interest (AA/B) and the comparator group (WH) to generate these charts were numbers of applicants, applications, applications discussed, and applications funded. We used these to calculate the percentage of applications that were discussed and funded. To visually compare the AA/B data with the much larger numbers in the WH dataset, we used rates per applicant (e.g., dividing the number of applications from AA/B scientists by the number of AA/B applicants). We calculated the percentage difference at each step based on the rate of change for applications from AA/B scientists over the rate of change for applications from WH scientists (e.g., applications from AA/B scientists are discussed at a rate of 44.0%, compared with 57.4% for WH scientists: 44.0/57.4% = 76.6%). We obtained the combined percentage difference by dividing the applications funded per submission for AA/B scientists by the same ratio for WH scientists (e.g., 0.27/0.54 = 50.4%). To analyze the effect of institution type on the funding gap, we extracted all R01 applications (Types 1 and 2) in FY 2011–2015 from the IMPAC II database, grouped applications by institution, and then ranked institutions from the highest to lowest aggregate amount of funding. We created quintiles and checked them to ensure that a roughly equal number of applications was present in each quintile. We used these quintiles in the rocket charts in fig. S2 to compare the highest- and lowest-funded institutions.

word2vec and clustering

To analyze the effect of application topic on the funding gap, we used the word2vec embedding method (*12*) to create topic clusters for the applications analyzed in this study. The goal of word embedding is to project the sparse high-dimensional features of individual words to a rich lower-dimensional space. When successful, this allows words that are close to each other semantically to be close to each other in the embedding space. For example, in this dataset, the words closest to insulin included glucose, diabetes, insulin secretion, hyperglycemia, and beta cell islet. We ultimately used document vectors, built from the individual word vectors, as the representative objects for clustering the applications into content areas. The clustering grouped areas of science in a way that was defined by the content of the applications but was agnostic to the underlying administrative boundaries. We binned clusters into quintiles from the lowest to the highest funding rates and displayed the distribution of AA/B and WH applicants across these quintiles. We ranked topic clusters by funding rates, i.e., the proportion of applications funded from quintile 1 (highest success) to quintile 5 (lowest success), to obtain equal numbers of topics in each quintile (n = 30 topics).

The text of interest, which consisted of all competing Types 1 and 2 R01 applications in FY 2011–2015, was preprocessed through a custom natural language processing (NLP) pipeline to handle internal data quality issues, identify domain-specific noun phrases, and optimize the input for the word2vec embedding (*35*). From each document, we first concatenated the text of the title, abstract, and specific aims. If it was not already present, we added a period to delineate the fields. Next, we converted the multiple encodings to standard American Standard Code for Information Interchange (ASCII) by transliterating any symbols to their closest equivalents. For example, the phrase "Aβ peptide" would be mapped to "Ab peptide." We then corrected for internal data quality issues by dehyphenating words that were split across lines and removed boilerplate text (e.g., "Abstract," "Specific Aims," or "Description provided by applicant"). Next, we substituted and replaced noun phrases as single tokens that were present in the entire corpus. Specifically, we first substituted for acronyms that were defined in the text via parentheticals and present in at least five documents. In this way, in the preceding sentences, we would replace both "NLP" and "natural language processing" with a single token: "natural_language_processing." In addition, we substituted for known terms and phrases in the National Library of Medicine's Medical Subject Heading (MeSH) vocabulary. After noun-phrase replacement, we removed all parenthetical statements and enforced a standard capitalization for each unique token in the corpus. Last, we only selected a subset of all possible parts of speech (POS). We removed syntactic POS-like determiners and conjunctions and semantic POS, including verbs and adverbs, and kept only nouns and adjectives. The rationale was twofold: First, we found that the resulting word embeddings used in independent downstream classification tasks had improved precision and recall, and second, the restricted vocabulary resulted in more interpretable clusters.

With the text preprocessed, we trained the word2vec word embedding over the corpus using the implementation in the program gensim (*36*, *37*). We selected the standard hyperparameter settings: 300 dimensions, a window of five, negative skip-gram sampling rate of 10^{-5}, 80 training epochs over the corpus, and a minimum of 10 words to be included in the dictionary. Once the training stabilized, we computed document vectors by taking the sum of the TF-IDF (term frequency–inverse document frequency) times the embedding vector for each unique word in the document. The resulting vector was renormalized onto the unit sphere. Since each document vector lies on the same manifold as the word embeddings, interdocument comparisons were made by taking the cosine similarity.

Converting a document (i.e., a collection of word tokens) into a single vector results in a loss of information. However, this transformation is necessary to make interdocument comparisons. Effectively, taking the average vector of the word tokens gives the average semantic sense of the document. In general, this is satisfactory, but often, individual datasets require domain-specific words that need to be ignored. In our particular analysis, we ignored words like "scientist,"

"plethora," "collaboration," "university," and "multitude" that reflect superfluous expository speech and that fail to describe a topic, as well as words like "collaboration" that describe meta-aspects of a grant; these words dominate and naturally form nondescriptive clusters. While the typical approach is to enumerate a list of stop-words to be excluded, we leveraged the power of word2vec to de-emphasize the importance of these words and all semantically similar words in a more nuanced fashion. We did this by applying a series of Gaussians centered on each superfluous word. The contribution of each word vector to a document vector was fractionally down-weighted to the proximity of these distributions.

With each document scored, we grouped the documents by using spectral clustering. Simple *k*-means clustering was not appropriate, since it ignores the spherical manifold of the embedding dimension, and a direct application of spectral clustering was too computationally expensive. Instead, we randomly sampled multiple subsets of 10,000 document vectors, computed the spectral clustering of the subset, and determined the centroids. An additional application of spectral clustering was used to collapse these centroids to "meta" centroids, e.g., centroids of the samples. Assignment of a document to a cluster was made by finding the closest meta centroid. The applications were clustered into three differently sized partitions, $k = 30$, 150, and 300. We found all three partitions to be informative: $k = 30$ reflected broad themes in the NIH portfolio and roughly approximated the distribution across the ICs; $k = 150$ produced well-defined areas of science; and $k = 300$ further refined these clusters into smaller, more specific topics, sometimes reduced to the level of an individual disease or treatment. Ultimately, we chose 150 clusters, ranging in size from 54 applications per cluster to 5125 applications per cluster. We viewed this as the best means of separating the entire NIH landscape into well-defined areas of research.

word2vec versus SME coding

We determined the level of agreement between word2vec clustering and human SMEs at NIH by presenting the SMEs with a series of binary questions derived from the word2vec partitions. Five SMEs were given a list of 10 applications from two word2vec clusters (five labeled as "group A" and five labeled as "group B"). They were allowed to use the title, abstract, and specific aims sections of each application to determine how the applications in each group belong together. The SMEs were then asked to partition a randomized set of 10 applications to the two groups, again using the title, abstract, and specific aims. This experiment was repeated by each SME for 10 sets of applications, with increasing semantic overlap (i.e., overlap group 0.0 had very little semantic similarity, and overlap group 0.9 had very high semantic similarity, as determined by word2vec), expecting that there would be less agreement between SMEs and word2vec at higher levels of semantic overlap. SMEs reproduced the groupings generated by the computational method 97.6% of the time, indicating that there is a very high degree of correlation between word2vec and human judgment. Median agreement between annotators was determined using Cohen's Kappa ($\kappa = 0.96$).

Variance of topic cluster percentile scores by study section

The median percentile score for each study section in each topic cluster was computed. Because the membership of study sections was not uniform for each cluster (i.e., clusters were often dominated by a small set of study sections), we calculated the weighted variance of the median percentile scores (within-cluster variance). Within-study section variance was similarly computed by sampling from all applications belonging to the topic clusters occupying a given study section. The distributions of within-cluster and within-study section variance were not significantly different ($P > 0.05$, using a two-sampled Kolmogorov-Smirnov statistic). Without context, the meaning of these variances is difficult to interpret. Thus, the same procedure was repeated after shuffling the cluster labels across the applications, randomizing the dataset. The shuffled distribution differs from both the within-cluster and within-study section distributions ($P < 0.001$), but the median difference between the pairs is large, indicating that the percentile scores within the clusters is consistent, and much more so than for a random clustering.

Number of study sections per cluster

To quantify the number of study sections that represented by a cluster, we ranked the study sections by the fraction of applications belonging to the cluster and then counted the study sections that compose 80% of the applications in the cluster. This threshold excluded study sections with very few applications from any given cluster and therefore do not represent the topic robustly. The number of clusters per study section was thresholded using the same method.

Topic cluster network

Topic cluster relatedness was determined using content analysis. Application content was analyzed using word2vec as described above, and intercluster relatedness ("distance") was measured between the median of one cluster and another. For ease of interpretability, this was calculated using a modified cosine similarity that ranged from [−0.3,1.0]. In the topic cluster network, nodes were considered connected if their median distance was at least 0.725.

Multivariate regression analysis

Our dataset includes Type 1 R01 applications submitted by WH and AA/B applicants between FY 2011 and 2015, restricted to applications for which there was only a single applicant to avoid the complications of applications with multiple applicants of different races. Additional data from FY 2006–2010 Type 1 R01s were used to construct some of the control variables, as detailed below.

We used probit models estimated with maximum likelihood when considering binary outcomes, such as whether an application is discussed or awarded. To consider the outcome of an application's percentile score, we used a linear model. Analysis was done at the application level with robust standard errors clustered by applicant. If an application is not awarded, it can be resubmitted. Since both the original and resubmitted application had the opportunity to be awarded, each is treated as a separate entry, with a control indicator for resubmissions. Analysis was done with Stata 14.

In addition to the race of the applicant, we controlled for a set of individual- and application-level parameters along with a set of organization-level parameters. Individual- and application-level parameters included the FY of the application, a binary indicator if the applicant is an ESI, a binary indicator if the application is a resubmission, and a continuous variable to describe the number of years since the applicant's last degree (linear and quadratic terms). To assess past success, we used continuous variables to describe both the number of prior R01 applications and awards per applicant. We also included two continuous variables related to the applicant's publication history as presented in the application's

biosketch. The biosketch of an NIH application represents the applicant's experience and includes a set of relevant publications selected by the applicant. We parsed biosketches to extract the relevant publications and generated two publication-based metrics for inclusion in the regression analysis: median biosketch RCR and number of biosketch publications in the top RCR decile. Because the biosketch contains only a selected group of publications, these features are proxies for publication influence rather than quantity. We transformed both RCR controls using an inverse hyperbolic sine (IHS) transformation. The IHS transformation behaves similarly to a logarithmic transformation and allows for transformation of zero values. Because the IHS transformation more closely approximates a logarithmic transformation when the numbers are not close to zero, we multiplied the median RCR by 100 before subjecting it to the transformation, effectively using the percentage of the field citation rate rather than the fraction of the field citation rate (*21*). We included a binary indicator for applications for which biosketch RCR data were missing (8.7%).

Organizational-level controls included the applicant organization's Carnegie classification [R1: doctoral universities (highest research activity); R2: doctoral universities (higher research activity), medical school, and other], the applicant organization's type in IMPAC II (higher education, hospital, research organization, and other), and the applicant organization's geographic region as defined by the U.S. Census (northeast, midwest, south, west, and outside the United States), all treated as categorical variables. We also included controls for the total amount of R01 funding and total number of applicants for the cognate organization in the prior period of 2006–2010 (both IHS transformed and treated as continuous variables). We used the prior period to avoid a deterministic relationship to award status for organizations with no funding.

As an independent method of controlling for organization-level characteristics, we used separate binary indicators for each applicant organization to more directly compare AA/B and WH applications from the same organization. Restricting data to organizations with more than 100 total applications and more than 10 applications from AA/B scientists [49 organizations, 30,664 (45%) of the full dataset], the estimated award rate gap between AA/B and WH applicants is 3.9 percentage points. Adding the 89 topic superclusters (see below) to the model reduced this gap by 9%. These estimates are quite similar to those obtained using the full sample, despite the fact that the subsample used was substantially smaller, limited to large organizations with both higher award rates (15.3% versus 14.2% outside the subsample) and a higher percentage of AA/B applicants (3.4% versus 2.1% outside the subsample).

We used a probit model to evaluate the probability an application was resubmitted, considering resubmissions in the FY of initial submission and the two subsequent FYs. We included all controls described above. We restricted this analysis to unawarded applications that are themselves not resubmissions. Restricting to the FY + 2 resubmission window should result in little censoring, as we found that over 98% resubmissions are submitted within this window for initial applications submitted in FY 2011–2014.

In assessing the number of applications submitted by each applicant, we used a Poisson model with analysis conducted at the applicant level. Quasi-maximum likelihood estimation with robust standard errors compensates for the restrictive assumptions of the Poisson model. To collapse application-level data to the applicant level, we used the mean for the RCR variable and years since degree variable, and the mode for the organization-level variables. Other application-level variables like FY were dropped.

We tested a variety of different topic area parametrizations to control for topic, beginning with the full set of 150 clusters generated with word2vec. Because some of these clusters contained a small number of applications, we merged them into various sets of superclusters as alternative topic parametrizations. Starting from the full set of 150 clusters, we iteratively merged clusters together in order of word2vec similarity under the constraint that no merged clusters comprised more than 5% of the application totals. We used 89 superclusters as our base case, since it most closely models the original 150 clusters.

In all cases, we reported the relationship between the independent and dependent variables as an AME, rather than reporting regression coefficients. The AME represents the average value of the marginal effect of the independent variable (e.g., AA/B applicant) on the dependent variable (e.g., probability the application is awarded). Because the regression models are not linear, the marginal effects differ depending on the values of the other independent variables. The AME was constructed by first calculating the marginal effect of interest for each observation in the sample at each observation's values for the other independent variables and then averaging these marginal effects.

SUPPLEMENTARY MATERIALS

Supplementary material for this article is available at http://advances.sciencemag.org/cgi/content/full/5/10/eaaw7238/DC1

Table S1. Attributes of R01 submissions by race of applicant.
Table S2. Distribution of AA/B and WH PIs across institutional funding quintiles.
Table S3. Comparison of the proportion of applications from new investigators by race (R01 Types 1 and 2, FY 2011–2015).
Table S4. Comparison of impact scores for R01 applications from AA/B and WH applicants (Types 1 and 2, FY 2011–2015).
Table S5. Comparison of word2vec topic assignment to human SME annotation.
Table S6. Statistically significant variation in the award rates of topic clusters.
Table S7. Topics favored by AA/B applicants compared to topics with no AA/B applicants.
Table S8. Effect of topic choice on funding gap throughout the R01 application process.
Table S9. Effect of removing variables from regression models (awarded given discussed).
Table S10. Reviewer demographics for all study section meetings that considered R01 applications (FY 2011–2015).
Fig. S1. Racial differences in the number of unique R01 applications submitted per applicant (Types 1 and 2, FY 2011–2015).
Fig. S2. Funding gap between AA/B and WH scientists at each stage of the R01 application and review process at institutions in the highest and lowest NIH funding quintiles.
Fig. S3. Comparison of career age distributions for AA/B and WH scientists.
Fig. S4. Time between initial application submission and resubmission.
Fig. S5. Distribution of topics across study sections.
Fig. S6. Comparison of topic choice variation by race.
Fig. S7. Award rates by topic cluster size.
Fig. S8. Scientific influence for higher and lower success topics.
Fig. S9. Comparison of scientific influence for publications linked to awards from higher and lower success topics by percentile score.
Fig. S10. Scientific influence of publications by new investigators before and after receiving their first award.
Regression analysis results for all variables

REFERENCES AND NOTES

1. National Institute of General Medical Sciences, *MARC Undergraduate Student Training in Academic Research (U-STAR) Awards*; www.nigms.nih.gov/Training/MARC/Pages/USTARAwards.aspx.
2. U.S. Department of Health and Human Services, *Minority Biomedical Research Support*; www.benefits.gov/benefits/benefit-details/696.
3. D. K. Ginther, W. T. Schaffer, J. Schnell, B. Masimore, F. Liu, L. L. Haak, R. Kington, Race, ethnicity, and NIH research awards. *Science* **333**, 1015–1019 (2011).
4. D. K. Ginther, S. Kahn, W. T. Schaffer, Gender, race/ethnicity, and national institutes of health r01 research awards: Is there evidence of a double bind for women of color? *Acad. Med.* **91**, 1098–1107 (2016).

5. S. Nikaj, D. Roychowdhury, P. K. Lund, M. Matthews, K. Pearson, Examining trends in the diversity of the U.S. National Institutes of Health participating and funded workforce. *FASEB J.* **32**, fj201800639 (2018).
6. Office of Extramural Research, National Institutes of Health, *Early Stage and Early Established Investigator Policies*; https://grants.nih.gov/policy/early-investigators/index.htm.
7. Office of Extramural Research, National Institutes of Health, *Peer Review: Scoring*; https://grants.nih.gov/grants/peer-review.htm#scoring2.
8. Working Group on Diversity in the Biomedical Research Workforce (WGDBRW), The Advisory Committee to the Director (ACD), National Institutes of Health, *Draft Report of the Advisory Committee to the Director* (2012); https://acd.od.nih.gov/documents/reports/DiversityBiomedicalResearchWorkforceReport.pdf.
9. J. E. Boyington, M. D. Antman, K. C. Patel, M. S. Lauer, Toward independence: Resubmission rate of unfunded National Heart, Lung, and Blood Institute R01 research grant applications among early stage investigators. *Acad. Med.* **91**, 556–562 (2016).
10. Office of Extramural Research, National Institutes of Health, *Resubmission Applications*; https://grants.nih.gov/grants/policy/amendedapps.htm.
11. Office of Extramural Research, National Institutes of Health, *NOT-OD-12-128: Time Limit on NIH Resubmission Applications*; https://grants.nih.gov/grants/guide/notice-files/NOT-OD-12-128.html.
12. T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in Vector Space. arXiv:1301.3781 (2013).
13. J. R. Kaltman, F. J. Evans, N. S. Danthi, C. O. Wu, D. M. DiMichele, M. S. Lauer, Prior publication productivity, grant percentile ranking, and topic-normalized citation impact of NHLBI cardiovascular R01 grants. *Circ. Res.* **115**, 617–624 (2014).
14. F. C. Fang, A. Bowen, A. Casadevall, NIH peer review percentile scores are poorly predictive of grant productivity. *eLife* **5**, e13323 (2016).
15. J. M. Doyle, K. Quinn, Y. A. Bodenstein, C. O. Wu, N. Danthi, M. S. Lauer, Association of percentile ranking with citation impact and productivity in a large cohort of de novo NIMH-funded R01 grants. *Mol. Psychiatry* **20**, 1030–1036 (2015).
16. N. S. Danthi, C. O. Wu, D. M. DiMichele, W. K. Hoots, M. S. Lauer, Citation impact of NHLBI R01 grants funded through the American Recovery and Reinvestment Act as compared to R01 grants funded through a standard payline. *Circ. Res.* **116**, 784–788 (2015).
17. N. Danthi, C. O. Wu, P. Shi, M. Lauer, Percentile ranking and citation impact of a large cohort of National Heart, Lung, and Blood Institute-funded cardiovascular R01 grants. *Circ. Res.* **114**, 600–606 (2014).
18. J. Berg, *Productivity Metrics and Peer Review Scores* (2011); https://loop.nigms.nih.gov/2011/2006/productivity-metrics-and-peer-review-scores/.
19. J. Berg, *Productivity Metrics and Peer Review Scores, Continued* (2011); https://loop.nigms.nih.gov/2011/2006/productivity-metrics-and-peer-review-scores-continued/.
20. D. Li, L. Agha, Big names or big ideas: Do peer-review panels select the best science proposals? *Science* **348**, 434–438 (2015).
21. B. I. Hutchins, X. Yuan, J. M. Anderson, G. M. Santangelo, Relative Citation Ratio (RCR): A new metric that uses citation rates to measure influence at the article level. *PLOS Biol.* **14**, e1002541 (2016).
22. D. K. Ginther, J. Basner, U. Jensen, J. Schnell, R. Kington, W. T. Schaffer, Publications as predictors of racial and ethnic differences in NIH research awards. *PLOS ONE* **13**, e0205929 (2018).
23. T. Bol, M. de Vaan, A. van de Rijt, The Matthew effect in science funding. *Proc. Natl. Acad. Sci. U.S.A.* **115**, 4887–4890 (2018).
24. R. K. Merton, The Matthew effect in science. The reward and communication systems of science are considered. *Science* **159**, 56–63 (1968).
25. M. J. Salganik, P. S. Dodds, D. J. Watts, Experimental study of inequality and unpredictability in an artificial cultural market. *Science* **311**, 854–856 (2006).
26. H. Liao, R. Xiao, G. Cimini, M. Medo, Network-driven reputation in online scientific communities. *PLOS ONE* **9**, e112022 (2014).
27. A. C. Morgan, D. Economou, S. F. Way, A. Clauset, Prestige drives epistemic inequality in the diffusion of scientific ideas. *EPJ Data Sci.* **7**, 40 (2018).
28. T. E. Day, The big consequences of small biases: A simulation of peer review. *Res. Policy* **44**, 1266–1270 (2015).
29. G. M. Santangelo, Article-level assessment of influence and translation in biomedical research. *Mol. Biol. Cell* **28**, 1401–1408 (2017).
30. H. A. Valantine, F. S. Collins, National Institutes of Health addresses the science of diversity. *Proc. Natl. Acad. Sci. U.S.A.* **112**, 12240–12242 (2015).
31. P. S. Forscher, W. T. L. Cox, M. Brauer, P. G. Devine, Little race or gender bias in an experiment of initial review of NIH R01 grant proposals. *Nat. Hum. Behav.* **3**, 257–264 (2019).
32. A. Casadevall, J. Handelsman, The presence of female conveners correlates with a higher proportion of female speakers at scientific symposia. *MBio* **5**, e00846-13 (2014).
33. K. D. Gibbs Jr., J. Basson, I. M. Xierali, D. A. Broniatowski, Decoupling of the minority PhD talent pool and assistant professor hiring in medical school basic science departments in the US. *eLife* **5**, e21393 (2016).
34. National Institute of General Medical Sciences, *Notice of Intent to Publish a Funding Opportunity Announcement for Maximizing Opportunities for Scientific and Academic Independent Careers (MOSAIC) Institutionally-Focused Research Education Award to Promote Diversity (UE5);* https://grants.nih.gov/grants/guide/notice-files/NOT-GM-19-019.html.
35. Natural Language Preprocessing (NLPre); https://github.com/NIHOPA/NLPre.
36. RaRe-Technologies, *gensim: Topic Modelling in Python*; https://github.com/RaRe-Technologies/gensim.
37. R. Řehůřek, P. Sojka, Software framework for topic modelling with large corpora, in *Proceedings of the Workshop New Challenges for NLP Frameworks* (2010).

Acknowledgments: We gratefully acknowledge the contributions of many NIH staff: J. Lun and T. Flock of the NIH Scientific Workforce Diversity Office for assistance in writing and coordination; R. Harriman of the NIH Office of Portfolio Analysis for data acquisition and analysis; and J. Wang, L. Roberts, and staff members in the Statistical Analysis and Reporting Branch of the NIH Office of Extramural Research for assistance with statistical analyses. **Funding:** The authors received no specific funding for this work, but all are employees or contractors for the NIH. **Author contributions:** Conceived and designed the analysis: T.A.H., A.L., K.A.W., R.A.M., M.J.P., B.I.H., M.S.L., H.A.V., J.M.A., and G.M.S. Project management/oversight: M.S.L., H.A.V., J.M.A., and G.M.S. Data management and acquisition: T.A.H., A.L., B.I.H., M.S.L., and G.M.S. Analyzed the data: T.A.H., A.L., K.A.W., R.A.M., M.J.P., B.I.H., M.S.L., and G.M.S. Tool/code development: T.A.H. Contributed to the writing/editing to the paper: T.A.H., A.L., K.A.W., R.A.M., M.J.P., A.F.D., M.S.L., H.A.V., J.M.A., and G.M.S. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials. Additional data related to the paper may be requested from the authors. However, some of the data are protected under the Privacy Act of 1974 (5 U.S.C. § 552a).

Submitted 18 January 2019
Accepted 14 September 2019
Published 9 October 2019
10.1126/sciadv.aaw7238

Citation: T. A. Hoppe, A. Litovitz, K. A. Willis, R. A. Meseroll, M. J. Perkins, B. I. Hutchins, A. F. Davis, M. S. Lauer, H. A. Valantine, J. M. Anderson, G. M. Santangelo, Topic choice contributes to the lower rate of NIH awards to African-American/black scientists. *Sci. Adv.* **5**, eaaw7238 (2019).