Is Blinded Review Enough?

How Gendered Outcomes Arise Even Under Anonymous Evaluation

Julian Kolev, Yuly Fuentes-Medel, and Fiona Murray[1]

October 30, 2019

Latest Version: http://bit.ly/2Y1mlDf

Abstract:

Blinded review is a direct and increasingly popular approach to reducing the impact of bias, yet its effectiveness is not fully understood. We take advantage of the blinded-review process to document the drivers of gender inclusion in a unique setting: innovative research grant proposals submitted to the Gates Foundation from 2008-2017. Despite blinded review, we find that female applicants receive significantly lower scores, which cannot be explained by ex-ante measures of applicant quality or applicants' choice of topic. By contrast, we show that the gender score gap is fully mediated after controlling for text-based measures of proposals' titles and descriptions, with female applicants tending to use narrow, topic-specific words that tend to receive lower reviewer scores. Importantly, the text-based measures that predict higher reviewer scores *do not* also predict higher *ex-post* innovative performance. Our results reveal that gender differences in writing and communication are a significant contributor to gender disparities in the evaluation and funding of science and innovation.

[1] Julian Kolev: jkolev@smu.edu; Yuly Fuentes-Medel: yuly@yfuentes-medel.com; Fiona Murray: fmurray@mit.edu.
We are grateful to the Bill and Melinda Gates Foundation for providing access to their internal data and procedures, and for their overall support. We thank Cecilia Testart Pacheco for her invaluable assistance with the text analysis portion of the paper. We also thank Pierre Azoulay, Scott Stern, Philippe Aghion, Ray Reagans, Michael Cima, Joshua Krieger, Daniel Fehder, Michael Bikard, Annamaria Conti, Donna Ginther, seminar participants at MIT and NBER, and conference attendees at the Academy of Management, the American Economic Association, and the REER conference for their helpful comments and suggestions.

I. Introduction

Diversity and inclusion are key goals for society at large, and specifically within the scientific and innovative communities across corporate, government, and academic organizations (Robinson and Dechant 1997, Bilimoria et al 2008, Østergaard et al 2011). Diversity of individuals and ideas leads to better outcomes (Azoulay et al 2011) and higher levels of productivity (Reagans and Zuckerman 2001). This is particularly important in the context of innovation-driven organizations: in such contexts, diversity and inclusion offer not only the usual benefits of greater efficiency as bias is eliminated, but also the potential to improve an organization's innovative capacity as individuals and teams benefit from the introduction of new ideas and perspectives (Freeman and Huang 2015, Tasheva and Hillman 2018). At the core of any attempt to increase diversity and inclusion lies the selection of projects and people, including hiring (Fernandez and Fernandez-Mateo 2006), promotion (Castilla and Benard 2010), and resource allocation (Boudreau et al 2016).

When seeking to maximize the selection of the best individuals, teams and ideas, organizations need to minimize or eliminate the tendency toward biased evaluation (especially bias based on ascriptive characteristics). However, eliminating the impact of bias is particularly challenging when selection involves both complexity and uncertainty; in such contexts, the use of heuristics is both prevalent and susceptible to the impact of bias (Bodenhausen and Wyer 1985, Bohnet et al 2015). While a wide range of interventions has been proposed for addressing this challenge[2], blinded review has often been considered as the 'gold-standard' process to remove opportunities for bias in evaluation and selection. In the emerging literature on 'blinded-ness," more diverse and

[2] Organizations and individuals have successfully reduced bias by emphasizing objective measures of candidates' ability and past performance (Reuben et al 2014), building institutional support for equality (Monroe et al 2008), mitigating the impact of differences in professional networks (Wold and Wenneras 2010), increasing the diversity of evaluators (Kunze and Miller 2017), and adopting accountability and transparency procedures (Castilla 2015).

inclusive outcomes have been predicated on the avoidance of patterns of bias in response to names (on CVs or scripts, as in McIntyre et al 1980), or in-person interactions (orchestra auditions in Goldin and Rouse 2000, entrepreneurial pitches in Brooks et al 2014). However, within the scientific community – the community which pioneered double-blinded review as a central process used to eliminate the role of biased perception in its own activities – evidence suggests that organizational practices do not in fact lead to unbiased evaluations (Bornmann et al 2007, Shen et al 2013, Ceci et al 2014, Witteman et al 2017). In light of this discrepancy, it is worth considering the impact of blinded review in the context of innovation and scientific activity, as it seems to be a necessary condition for the elimination of bias. We attempt to evaluate whether blinded review is *sufficient* to overcome all aspects of under-representation, or whether there are significant barriers to diversity and inclusion that remain after its implementation.

Our paper explores the degree to which gender shapes outcomes even in a blinded setting, and seeks to evaluate the drivers of gender disparity in science and innovation, in terms of both access to key inputs and subsequent outputs. We take advantage of data covering a blinded grant-review process from the Bill and Melinda Gates Foundation that has several critical characteristics: reviewers evaluate anonymous proposals, the reviewing process enables multiple individual reviewers to provide independent scores, it provides scoring data at the reviewer-proposal level, and it allows us to trace individual applicants' later activities through scientific publications, subsequent NIH grant receipts, and other measures of innovation.

Using a sample of 6,794 proposals submitted to the Gates Foundation between 2008 and 2017, we analyze two major components of the interplay between diversity and innovation. First, we examine the determinants of diversity in innovative organizations, by examining the role of gender in explaining reviewer evaluations of innovative proposals. Second, we construct a difference-in-difference estimator to explore the interaction between funding and applicant gender, in order to

identify the differential impact of funding across applicants. In these analyses, we focus on a homogeneous sample of US-based life science researchers, and take advantage of the features of our setting to identify the causal impact of gender on both reviewer scores and subsequent outcomes.

Our results offer important new insights into the relationship between gender and innovation. We find that even in an anonymous review process, there is a robust negative relationship between female applicants and the scores assigned by reviewers. This disparity persists even after controlling for proposal topics, reviewer demographics, applicant publication histories, and applicants' prior applications. However, the gender disparity becomes insignificant after controlling for text-based measures of applicants' proposals. Specifically, building on the text analysis methods of previous studies (Schmader et al 2007, Magua et al 2017), we show that female applicants use fewer of the words favored by reviewers when describing their proposals, and more of the words associated with low reviewer scores. This finding persists after controlling for broad topic areas (e.g. HIV, malaria, tuberculosis), and is robust to finer-category controls, suggesting that our findings are not driven by the female applicants' selection of under-valued topics, but rather by the specific approach they take in the pursuit of funding. Exploring this pattern further, we show that female applicants have a tendency to choose more "narrow" (i.e. topic-specific), words and fewer "broad" words. Both of these tendencies lead to lower scores for female applicants, with their use of narrow words having the greater negative effect. After controlling for the impact of word choice, the gender-based score disparity is no longer significant, and its effect size drops by over 50%. Our findings therefore suggest that a focus on writing style, word choice, and the details of applicants' research approaches can offer insight into the drivers of the gender gap in science and innovation.

Having shown how differences in communication and word choice influence the evaluation of innovative proposals, we then turn our attention to follow-on innovative outcomes. Specifically, we explore how male and female applicants differ in their innovative output subsequent to their application, across measures including academic publications and NIH grant awards. We begin by establishing that the text-based measures which were a major driver of selection by (blinded) reviewers *do not* also predict an increase in follow-on innovation. Most prominently, we show that the use of the broad words favored by reviewers actually predicts a neutral-to-negative impact on ex-post outcomes, suggesting that reviewers may be overly credulous to broad descriptions that are likely to reflect style more than substance. Next, we find that across a range of subsequent outcomes, being selected by the Gates Foundation leads to a bigger impact for female applicants, primarily through "leveling the playing field" relative to male applicants. Specifically, female applicants are disadvantaged relative to male applicants if they are not selected; by contrast, successful female applicants generate innovative outcomes that are either indistinguishable from or better than those of successful male applicants. Indeed, the disappearance of disparities after being selected and receiving Gates Foundation funding suggests that from the perspective of impact, female applicants may well generate a greater "return" on Gates Foundation resources. This effect is strongest for the outcome of NIH grant funding, where successful female applicants not only catch up but significantly outperform their male counterparts.

Overall, our results identify two major areas where gender and innovation interact: first, the tendency of female applicants to use narrow words to describe their innovations can lead to a significant reduction in their perceived quality (even when the proposed innovations are actually high-quality!). Second, the failure to be selected for funding acts as a disproportionate barrier to the follow-on innovation of female applicants, while successful female applicants exhibit outcomes that are either indistinguishable from or superior to those of their male counterparts.

These findings suggest that there is significant scope for improvement at innovation-driven organizations, in terms of both increasing the selection of high-quality projects, and allocating resources to the innovators for whom they would have the greatest impact.

II. Theoretical Framework and Hypotheses

A. Heuristics and Biases Under Uncertainty

The starting point of our theoretical framework is the role of heuristics within behavioral decision theory: when asked to make a judgement in the face of uncertainty, decision-makers will often turn to heuristics in an attempt to determine the optimal choice with limited information (Tversky and Kahneman 1974, Maitland and Sammartino 2015). At their core, heuristics simplify complex decision processes by focusing on the most useful aspects of available information while ignoring less-reliable indicators; heuristics are widely-used in a range of decision-making contexts; moreover, in some cases, they are an efficient approach that can outperform more complex methods of analysis (Gigerenzer and Gaissmaier 2011, Mousavi and Gigerenzer 2014). However, heuristics also tend to generate systematic errors and well-documented biases, including the use of stereotypes over direct evidence (Bodenhausen and Wyer 1985). In light of the potential benefits of heuristics, recent work has attempted to identify interventions which guides the process toward useful information, and away from potential sources of bias: Hoffman et al. (2017) highlight test scores as superior to managers' subjective evaluations, while Bohnet et al. (2015) show that joint evaluation of multiple candidates reduces bias relative to evaluating candidates individually. This paper seeks to build on this prior work on the impact of information availability on the effectiveness of decision-making: specifically, we analyze the intervention of blinded review in the high-uncertainty context of innovation, to evaluate whether the elimination of demographic information can reduce gender disparity in reviewers' decisions.

B. Blinded Review and The Structure of Gender Disparity

Recent studies have documented the prevalence of gender disparity and bias in a wide range of contexts, including academia, business, and government (Bohnet et al 2015). Blinded review is a prominent intervention seeking to reduce or eliminate such disparity (Goldin and Rouse 2000); in addition, it has an important benefit from a research perspective: it can separate the direct drivers of bias and discrimination from indirect mechansisms. With direct mechanisms such as bias and discrimination, it is possible to observe that an individual belongs to a disfavored group (Castilla 2008, Brooks et al 2014), and this leads to a disparity in outcomes. This stands in contrast to studies focusing on indirect mechanisms, where a disparity in outcomes exists despite the lack of direct discrimination (Ginther et al 2016, Fernandez and Campero 2017). In this latter group, disparities can be driven by self-selection, conforming to stereotypes, or avoidance of competition, rather than the direct impact of ascriptive bias. Thanks to the blinded review process used in our empirical setting, we can not only estimate the magnitude of the gender disparities that persist after eliminating traditional forms of bias and discrimination, but also highlight the precise mechanisms which drive any remaining gender disparities.

Prior work indicates that bias can interact with structural elements of organizations (Murray and Graham 2007, Kelly et al 2010, Ding, Murray, and Stuart 2013) to generate significant disparities in outcomes across demographic groups. If such bias-driven mechanisms are the dominant driver of organizational decision-making, one would expect that the adoption of blinded review would lead to parity in inclusion for women and other under-represented groups. Specifically, in evaluations of innovative proposals, we predict that blinded review will make gender irrelevant after controlling for the quality and experience of applicants.

Hypothesis 1 (H1): *Under blinded review, there will be no gender gap in the scores applicants receive, after controlling for applicant characteristics.*

C. Potential Disparities Under Blinded Review

While blinded review is often seen as the ideal intervention for eliminating the impact of bias, either explicit or implicit, this may not be the sole driver of demographic disparities within organizations and throughout the economy. Differences in interests, experience, risk tolerance, or other factors may form the foundation of indirect mechansisms of gender disparity, even under a completely unbiased review process. One of the strengths of our empirical setting is that we can test for such indirect mechanisms, as the intervention of blinded review has ruled out direct mechansism such as bias and discrumination. To identify potential indirect mechansism of gender disparity, we rely on the rich literature of gender differences across a broad range of theoretical domains.

The first indirect mechanism that we explore is one that has been studied extensively in the gender wage gap literature: the tendency for women to self-select into lower-payoff specializations (O'Neill 2003, Goldin 2014). This tendency may well be related to the findings that women tend to be significantly less competitive than men (Niederle and Vesterlund 2007), with this difference explaining a substantial portion of gender differences in chosen fields of study (Buser et al 2014).[3] Translated to our context, we would expect that women will self-select into less-rewarding topic areas, and therefore receive lower scores as the result of their choices.

Hypothesis 2 (H2): *Women will tend to submit proposals to topic areas that are less likely to be valued by reviewers, leading to lower scores and generating a gender gap in evaluations.*

In addition to the tendency for women to avoid competition from an *ex-ante* perspective, recent research has also documented the potential for gender differences *ex-post*. Specifically, Buser (2016) shows that women are less likely to seek additional challenges in response to losses in

[3] It is worth noting that there is significant variation in the gender confidence gap across disciplines, with Kamas and Preston (2012) finding no significant gender differences in confidence for undergraduate students in STEM fields. At the same time, Correll (2001) finds that the gap in self-assessed competence in high-school students is significant in mathematics but not in verbal tasks.

tournament settings where outcomes are based on relative performance. Indeed, such differences in responses to setbacks can account for a large portion of gender performance gaps in experimental settings (Gill and Prowse 2014). These findings are consistent with social-cognitive models of behavior, where women are tend to exhibit higher levels of rejection sensitivity: they perceive negative feedback as an indication of the futility of their efforts, leading to the phenomenon of "self-silencing" where an individual suppresses their own beliefs if they are perceived as non-conforming (London et al 2012). Indeed, recent work suggests that one of the potential consequences of gender rejection sensitivity is that women will face particularly large barriers to success in STEM fields, particularly in response to negative academic experiences (Ahlqvist et al 2013). Combining the predictions of these literatures, we would expect that female applicants in our setting will be less likely to re-apply if their initial application is rejected, and that this is likely to generate a gender gap in evaluations if repeat applicants tend to receive higher scores:

Hypothesis 3 (H3): *Women will be less likely to reapply after rejection, and in combination with stronger performance by repeat applicants, this will generate a gender gap in evaluations.*

Beyond the potential for gender disparities in whether and where to apply, recent work on gender differences suggests that women may differ from men in *how* they approach the application process. Specifically, recent studies document that there are significant differences in writing and communication between the genders; moreover, like the mechanisms described above, these differences in communication often stem from underlying differences in confidence. Ibarra and Obodaru (2009) identify the lack of a "presumption of competence" as one reason women tend to adopt a defensive posture and gravitate to safe choices, leading them to be judged as less visionary, even as they outperform men in all other dimensions of leadership. A similar pattern emerges in the innovation-focused setting of entrepreneurial pitch competitions: Kanze et al (2018) highlight

how women are pushed to speak defensively while men are pushed to speak to growth opportunities when interacting with potential investors. Women's propensity to adopt a cautious and defensive posture leads to significant gender differences in writing style in academic contexts: the bolder and more risk-taking writing of male academics leads to higher assessments of quality, particularly in argument-based rather than fact-based writing (Earl-Novell 2001). Indeed, when attempting to publish their research, female academics are often held to a higher standard of writing and proactively seek to address anticipated gender biases in their initial written submissions (Hengel 2019). This body of prior work differs from our context in two important ways: first, none of the above studies are able to implement blinded review; second, none of them address the specific challenges of writing and communication in the context of science and innovation. Despite this, we would expect that the tendency for women to adopt a defensive posture to persist even in a blinded setting; moreover, within the context of scientific writing, one might expect that caution would lead to a narrowing of expected results and applications, whereas confidence would entail a broader and more expansive view of scientific potential. Within our context of scientific grant proposals, we would therefore expect female applicants to use narrower language than male applicants, and for reviewers to perceive this writing style in a negative light:

Hypothesis 4 (H4): *Women will tend to submit proposals whose communication style is not received favorably by reviewers, due to a lower level of confidence and a tendency to use narrower language.*

Taken together, hypotheses H2-H4 generate a counterpoint to hypothesis H1: while organizations should certainly aim to remove bias from their evaluations, the intervention of blinded review may not be sufficient if their goal is to eliminate demographic disparities in the fields of science and technology.

D. The Interaction of Demographics and Innovation

A number of recent studies have focused on the under-representation of women, racial minorities, and other demographic groups in innovative fields (Cook and Kongcharoen 2010, Lincoln et al 2012, Bell et al 2018, Marschke et al 2018) and knowledge-focused organizations (Fernandez and Campero 2014, Gompers and Wang 2017). The academic attention has been complemented by discussion of workforce diversity in the popular press, leading top technology companies like Google and Apple to begin disclosing the demographic statistics of their employees. While such reports can spur greater awareness of the lack of diversity in knowledge-based organizations, these group-level statistics do not offer an opportunity to identify the mechanisms behind the patterns of diversity, because they do not capture outcomes at the individual level. By adopting the perspective of individual innovators, we add to the growing literature seeking to track the presence and impact of demographic disparities over the course of individuals' entire careers (McKown and Weinstein 2002, Lerchenmueller and Sorenson 2018, Hengel 2019). Importantly, demographic inclusiveness and diversity are not only desirable end-goals, but can also be important inputs in the context of innovation: Freeman and Huang (2015) demonstrate that collaborations across demographic categories are associated with higher-impact research.

Even as the value of inclusiveness and diversity is elevated in the context of innovation, this setting also offers additional barriers to developing and maintaining demographic diversity and inclusiveness in the first place. Relative to other fields, the pursuit of innovaiton has a number of features which make it difficult for standard efficiency-based effects to push out discriminatory tendencies.[4] Specifically, the following features of innovation combine to make it more likely to harbor a persistent lack of diversity:

- Outcomes are unpredictable, and there are long lags before success is realized

[4] For a full discussion of the traditional perspective on discrimination and (in)efficiency, see Becker (2010).

- Outcomes are only observable for funded or attempted innovations
- An individual's contribution is hard to separate, especially in cumulative innovation
- Teamwork is often required to push past the current state of the art, and some team members might prefer working with those similar to them

These challenges are significant, and innovative fields do often generate greater disparities than other sectors (Magua et al 2017). At the same time, diversity, while lacking, may be uniquely valuable in the context of innovation (Robinson and Dechant 1997, Østergaard et al 2011). Innovative organizations would stand to gain not only the usual efficiency benefits from a reduction in discrimination, but they would also have the potential to introduce new ideas, or recombine existing ideas in new ways. We therefore seek to evaluate both sides of this question in our context, as we focus specifically on the dimension of gender: what mechanisms can lead innovative organizations to become more inclusive, and what are the potential benefits of such an increase in inclusiveness?

III. Data and Methods

A. Empirical Strategy and Regression Specifications

In analyzing the relationship between gender and innovation-related outcomes, there are a number of challenges that normally interfere with attempts to estimate causal effects. To capture an organization's ability to evaluate innovations, an ideal empirical design would take ideas of comparable ex-ante quality, and randomly assign each idea to multiple applicants across a range of demographic groups. These applicants, also of comparable ex-ante quality, would develop and submit proposals based on the ideas to the organization. If the proposals were submitted anonymously, it would be possible to remove the direct elements of bias and identify any indirect or underlying causes of gender disparity. Ideally, these proposals would then be (independently) evaluated by a diverse set of reviewers, in order to examine the relative impacts of applicant and

reviewer demographics. We can approximate this ideal empirical design in our setting by attempting to control for idea and applicant quality while comparing the scores that different reviewers give to a single proposal. Thus, the regression specification we would like to estimate is:

$$Reviewer\_Score_{ijk} = \beta_0 X_{ijk} + \beta_1 ApplicantGender_j + \beta_2 ReviewerGender_k + \beta_{12} Interaction_{jk} + \epsilon_{ijk}$$

In the above equation, we estimate the score of idea *i*, from applicant *j*, as evaluated by reviewer *k*. The vector of covariates X_{ijk} varies by specification, and can include fixed effects for the time of submission and the subject area of the idea, text-based measures of the idea's title and description, and a range of applicant and reviewer characteristics. The demographic variables capture the gender for both applicants *j* and reviewers *k*, and the interaction term identifies the impact of shared demographic characteristics between applicants and reviewers (e.g. a female reviewer evaluating a proposal from a female applicant). The primary coefficient of interest in these specifications is β_1, the impact of applicant gender on the score received from reviewers; specifically, we will track how this coefficient changes based on the inclusion of controls for covariates that might be correlated with both applicant gender and reviewers' evaluations.

Turning to the second portion of our empirical design, we encounter the additional challenge of analyzing the relationship between gender and innovative outcomes. In an ideal research setting, we would like to take a single idea and assign it to "twin" applicants of identical quality and from the same demographic group, and randomly assign funding to one but not the other. We would then want to compare the funded idea with the one that did not receive funding, and perform a difference-in-differences analysis across demographic groups whose ideas also receive the same random assignment of funding. In this context, we approximate this ideal design through a

regression-discontinuity approach. Specifically, we compare funded proposals to proposals that received high scores from reviewers but did not receive funding. Our estimated equation becomes:

$$InnovativeOutcome_{ij} = \beta_0 X_{ij} + \beta_1 Funding_i + \beta_2 ApplicantGender_j + \beta_{12} Interaction_{ij} + \epsilon_{ij}$$

In the above equation, we analyze our sample at the level of idea *i* from applicant *j*, and X_{ij} captures key covariates including fixed effects for the time of submission and the subject area of the idea, as well as key applicant characteristics such as innovative output during the ex-ante period. While the direct effects of funding and applicant gender are valuable for interpreting the overall pattern of results, the primary coefficient of interest is β_{12}, which captures the differential impact of funding across applicant gender.[5] In effect, this is a difference-in-differences estimator of the impact of diversity on innovation, and allows us to draw conclusions regarding the efficiency of the innovative process.

B. Empirical Setting: The Gates Foundation's GCE Program

In the previous section, we identified a set of ideal experimental settings that would allow researchers to estimate the impact of diversity on innovation, both in terms of understanding the drivers of diversity in innovative organizations, and in terms of the impact of diversity itself on innovative outcomes. While the expectation of random assignment of funding is not likely to be satisfied in any well-run organization, our empirical setting does offer a number of valuable features that allow us to estimate the relationship between diversity and innovation. Our empirical setting is the Global Challenges: Exploration (GCE) Program at the Bill and Melinda Gates Foundation (subsequently, the Gates Foundation), providing a sample 6,794 anonymous proposals submitted by US-based researchers from 2008-2017. While this program offers valuable internal

[5] In some specifications, we focus solely on applicants who successfully obtained funding; in this setting, rather than a difference-in-differences approach, we instead draw conclusions from the direct effects of gender on innovative outcomes.

information on the decisions of individual reviewers, it also differs from more "traditional" grant review institutions (e.g. the NIH or NSF): these organizations use a non-blind review process and send proposals only to reviewers within the proposal's narrow subfield. Further, traditional grant-review institutions often engage in consensus-based collective decisions, where reviewers discuss proposals together and at a collective evaluation representing the views of the entre reviewer panel (NIH 2008). By contrast, the GCE Program implements a review process with the following key characteristics: diverse panels of reviewers, anonymous proposals, and champion-based funding decisions.

- Anonymous proposals - reviewers have no information on the candidate beyond the proposal details
- Diverse pool of reviewers - reviewers are drawn from a wide range of scientific fields and may be from non-academic backgrounds such as the private sector or government
- Independent evaluations - reviewers do not confer with each other when assigning scores
- "Champion-based" review - rather than relying on consensus across multiple reviewers, strong support from a single reviewer can greatly increase the odds of being funded[6], while strong negative reviews are treated as identical to neutral reviews

The above features reflect the priorities of the Gates Foundation in its search for solutions to major challenges in global health and development. At the same time, they introduce a richness of variation in both reviewers and proposals, allowing for detailed analysis of reviewer decisions and project outcomes.

C. Data Sources and Key Variables

[6] While a single reviewer can greatly increase the funding odds for any single project, their highest level of endorsement can only be given to one of the approximately 100 proposals that they review during each round of the program.

The data for our analysis comes from the first nineteen rounds of the GCE program, which includes a total of 17,311 proposals focused on infectious disease research submitted between 2008 and 2017. In order to ensure a homogeneous pool of applicants, we focus on the 6,794 proposals submitted by applicants who have both an academic or non-profit research affiliation and a US contact address.[7] These proposals, submitted by 5,058 unique applicants, were allocated across subsets of 132 innovation-panel reviewers, leading to a total of 21,453 reviewer-proposal pairs. From the Gates Foundation, we obtain identifying information on the identities of proposers and reviewers, the substance of the proposals, and the reviewer scores and funding outcomes resulting from the program. We calculate probabilistic measures of gender for both applicants and reviewers using the techniques of Sood and Laohaprapanon (2018). Beyond this, we calculate a number of applicant and proposal characteristics: for applicants, we track all publications listed in the Scopus database. We then merge these records to PubMed to identify top-journal publications (defined as being in the top 10% by journal impact factor), publications where the applicant is the last author, and publications which include a new journal, coauthor, or Medical Subject Heading (MeSH) term when compared against the applicant's prior publications. For proposals, we calculate word counts, average word lengths, parts of speech, number of MeSH terms, and measures of the text's grade level using the Flesch-Kincaid, Gunning-Fog, and SMOG formulas. Finally, we analyze the 500 most frequent proposal words in our sample (after dropping a standard set of "stop words" including "and," "not," "or," "with," and others). One of our key measures is the division between broad and narrow words: using the ten topic areas depicted in Figures 1 and 2, we calculate the standard deviation of each word's log-use-rate across proposal topics. We then split this set of 500 words at the median, which in our sample is approximately 0.75, corresponding to a relative

[7] In the interest of analyzing a homogeneous sample, we also drop the small fraction applicants that we identify as members of under-represented minorities.

standard deviation of over 100% in linear use rates. Words with a below-median standard deviation across topics are categorized as broad words, while words with an above-median standard deviation are categorized as narrow words.

Reviewer evaluations in our sample are simple but highly skewed: from an average portfolio size of over 100 proposals per round, reviewers can choose a single proposal to receive their highest level of support (categorized as a "Gold" rating by the GCE program), and can award up to five other proposals with a high level of support (categorized as a "Silver" rating).[8] Our empirical approach employs an ordered-logit specification to capture this decision process, with the outcome variable of *Reviewer Score* equal to 1 for a gold rating, 0.2 for a silver rating, and 0 otherwise.[9] Based on these reviewer evaluations, 635 of the 6794 proposals (~9.4%) in our sample received grants of $100,000 each during the first nineteen rounds of the GCE program.

We use the identifying information on reviewers and applicants to obtain information on institutional affiliation, gender, country of residence, and area of expertise. Notably, even at the proposal stage, we find a significant gender gap: our sample of US-based academic applicants is 66% male. In addition to obtaining demographic information, we obtain applicants' full publication histories in order to estimate research productivity both before and after the peer review process. Table 1 lists summary statistics for these and other key variables, while Figure 1 presents the distribution of proposals and ratings across topic areas. Using this data, we proceed to investigate the role of gender in the peer review process, and the ability of peer review to both evaluate and select promising research projects and innovations.

TABLE 1 & FIGURE 1 HERE

[8] It is worth noting that this rating system does not offer the opportunity for reviewers to send negative signals: proposals not receiving support from a given reviewer all receive the same rating (effectively, a score of zero), regardless of the reviewer's degree of disapproval.

[9] We choose these numerical values of gold and silver ratings to reflect the number of such ratings that reviewers can award in each round. Note that the choice of values in has no impact on results in an ordered logit specification.

IV. Analysis and Results

A. Determinants of Reviewer Evaluations: Evidence of Disparity

Our results begin with Table 2, which examines the impact of gender on the scores received by applicants' proposals. Columns 1 through 4 evaluate the overall impact of applicant gender across all reviewers in our sample, while column 5 decomposes the effect between male and female reviewers by adding an interaction between applicant and reviewer genders. All specifications include fixed effects by GCE round; specification 2 adds reviewer characteristics, and specifications 3 through 5 include reviewer fixed effects. In addition, specifications 4 and 5 include topic area fixed effects. In all specifications, we find consistent negative disparities for female applicants: such applicants are significantly less likely to receive high scores from reviewers. The effect size we estimate reflects our ordered-logit specification, reflecting a log-odds ratio that is approximately 16% lower than male applicants. In separate calculations, we find that this overall effect is driven by female applicants being approximately 15% less likely to receive a "silver" rating and 20% less likely to receive a "gold" rating from reviewers. In specification 5, we also include the interaction of female applicants and female reviewers; we find a strong positive effect for this variable. Importantly, this should not be interpreted to conclude that female reviewers have an affirmative preference for proposals from female applicants: the sum of the direct effect and the interaction is positive, but not significantly different from zero.[10] Effectively, the direct effect captures the lower scores that female applicants receive from male reviewers, while the interaction term indicates that female reviewers do not exhibit a similar disparity in reviewer evaluations across gender.

[10] Only 20 of 132 (~15%) reviewers in our sample are female, and in combination with the low proportion of female applicants, we lack the statistical power to precisely estimate the interaction between female reviewers and applicants.

Since the impact of applicant gender in Table 2 is robust to including topic-area fixed effects, our results suggest that the gender disparities we observe are not driven by applicants choosing less-valued areas of study. We explore this question further in Figure 2, which compares the prevalence of female applicants against average proposal scores across the topic areas in our sample. We find that there is significant variation across topics in both dimensions; further, we find a weak positive relationship between the rate of female applicants and the average of reviewer scores across topic. The patterns in Figure 2 therefore lend additional support that the choice of topic is not a significant driver of the gender disparities in our sample.

TABLE 2 & FIGURE 2 HERE

In Table 3, we continue our analysis of demographic disparities in reviewer scores, adding explicit controls for applicant quality using the measure of pre-period publications.[11] Panel A explores a range of publication-based metrics, and finds that the female applicants in our sample are at a disadvantage relative to male applicants across all measures. We begin with career length in specifications 1 and 2, where we find that female applicants are approximately 30% less senior than male applicants, as measured by the number of years from an applicant's first academic publication. This result is robust to adding topic fixed effects in specification 2. We perform similar analyses for the number of publications, and the share of top-journal publications[12] and last-author publications, all in the three years prior to applying to the GCE program. Across all three measures, we find that female applicants are at a significant disadvantage, though this effect diminishes significantly after controlling for topic-area fixed effects and applicant career length. Overall, these results suggest the possibility that within our sample, female applicants have weaker publication histories, and may therefore have less ability to generate a high-scoring proposals.

[11] Specifically, we define the "pre-period" as the three years prior to the proposal's application year.
[12] We define a top journal as one falling in the top 10% of journals by impact factor.

We evaluate the impact of this difference in prior publications in Panel B, where we repeat the analysis of Table 2 while controlling for the details of applicants' publication history. We find a significant positive impact of prior publications on reviewer scores in specification 2; after including additional measures, we find that the share of top-journal publications is the strongest predictor of reviewer scores in our setting. Overall, we find strong evidence for the hypothesis that applicants with a superior publication history tend to generate higher-scoring proposals. However, even after controlling for a range of measures of applicants' prior publications, we continue to find a significant disadvantage for female applicants in all specifications. This pattern of results suggests that the ex-ante differences in publication patterns across gender do not explain a significant portion of the outcome disparities we find in the proposal-evaluation process.

TABLES 3A & 3B HERE

B. Mechanisms Driving Disparity: Repeat Applications

Having presented evidence for consistent gender disparities in reviewer scores, we now proceed to evaluate the mechanisms through which they might emerge even under a blinded-review process. The first potential mechanism is that of persistence: might female applicants be more easily discouraged if their first proposal is rejected? We evaluate this dimension in Panel A of Table 4, where we find that female applicants are significantly less likely to reapply after an initial rejection. This negative association becomes insignificant in specifications 3 and 4 after controlling for career length, suggesting that experience can lead to increased persistence that mitigates this aspect of gender disparity. We then proceed to evaluate the impact of this difference in Panel B; we show that repeat applicants receive significantly higher scores than first attempts. This effect is both strong and highly robust, persisting in all specifications as we include topic fixed effects, publication characteristics, and interaction effects. Our results therefore serve as a reminder of the value of persistence in the face of rejection; in light of the results in Panel A, this

is advice that is particularly important for female researchers and innovators at the early stages of their career.[13] However, controlling for this repeat-applicant effect does not meaningfully reduce the gender disparity we identified in Tables 2 and 3. Thus, while repeat applicants are expected to receive significantly higher scores from reviewers, this dimension does not explain a significant portion the female score disparity in our sample.

TABLES 4A & 4B HERE

C. Mechanisms Driving Disparity: Word Choice

A second potential mechanism to explain demographic disparities in an anonymous review process is the set of words that applicants use to describe their proposals. Specifically, we analyze the words present in the title and descriptions of applicants' proposals, after removing standard conjunctions, pronouns, and linking words. In Figures 3 and 4, we plot words based on their relative rates of use by male and female applicants, and include 45-degree lines to clearly separate "male" from "female" words in our data. We present the 100 most frequent of the remaining words in Figure 3, in order to show that word use rates are strongly correlated across applicant gender. While the overall correlation is strong, there is nevertheless significant variation in the gender use rates of some words. Figure 4 highlights this trend by focusing on words that have a significant correlation with reviewer scores, and also identifies the difference between "narrow" words (those which appear significantly more often in some topics than others), and "broad" words (which appear at similar rates in all topic areas). The latter figure suggests that male applicants tend to favor broad words, while female applicants have a tendency to use narrow words. Overall, these figures identify the similarities and difference in word choice across applicant gender, and set the stage for analyzing whether the differences we find can explain the gender score gap in our sample.

[13] While repeat applicants receive a significant boost in terms of their expected scores from reviewers, the interaction between female applicants and repeat applicants in specifications 5 and 6 are insignificant. Thus, there is no additional boost to female applicants who re-apply after rejection, relative to male applicants who do the same.

FIGURES 3 & 4 HERE

In addition to disparities in use rates across applicant gender, the frequent words in our sample also have significant disparities in their tendencies to appear within high- and low-scoring proposals. We highlight these score disparities in Figure 5, plotting frequent (i.e. top-100) words that are disproportionately associated with male or female applicants based on their rates of use within high- and low-scoring proposals. While there is a reasonable amount of variation across words, the overall pattern suggests that words falling well below the 45-degree line (i.e. those with strongly negative score disparities) are much more likely to be used disproportionately by female applicants. By contrast, most "male" words are likely to be near or above the 45-degree line, indicating a positive score disparity. We follow up the binary analysis of gender in Figure 5 with a full two-dimensional exploration of gender- and score-based disparities in Figure 6. This figure highlights the positive relationship between words used more often by male applicants, and words associated with high-scoring proposals. Importantly, while most "male" words have a positive score disparity, there are "female" words with both positive and negative score disparities. This effect seems to be associated with the difference between broad and narrow words: the broad words in Figure 6 seem to be driving the high scores of male applicants, while the narrow words seem to be associated with the lower scores of female applicants. This suggests that there is significant scope for female applicants to improve their scores by altering the words they use to describe their proposals.

FIGURES 5 & 6 HERE

In light of the patterns described above, we now proceed to explore the impact of word choice on reviewer scores using a regression framework in Table 5. We begin with Panel A, which highlights some of the basic patterns of word choice in our sample, focusing on the top 1,000 most frequent words used in the proposals in our sample. We examine the differences in word use between male

and female applicants, and in specification 1, we show that the total number of frequent words does not differ across gender. By contrast, we find significant differences across applicant gender in all remaining word-choice measures: female applicants use fewer high-scoring words and more low-scoring words in specifications 2 and 3, and this result is robust to identifying high- and low-scoring words using only male applicants' proposals in specifications 4 and 5. Finally, in specifications 6 and 7, we classify words as "broad" and "narrow" based on whether they are used at similar or different rates across topic areas, and show that female applicants use fewer broad words and more narrow words when describing their proposals.

Having established these basic patterns, we next construct text-based measures of each proposal's quality as perceived by reviewers. Specifically, we predict a proposal's reviewer scores based only on the presence or absence of the top 500 most score-influencing words in each proposal's title and description. Importantly, to avoid circular reasoning, we calibrate our text-based reviewer score predictions using only the proposals from *male* applicants. Thus, the high-and low-scoring words we identify are those with which male applicants receive high and low scores from reviewers, with no information regarding the use of these words by female applicants. We normalize score predictions to a mean of zero and unit standard deviation, and proceed to use them as dependent variables in Panel B, and explanatory variables in Panel C.

In Panel B, we analyze the relationship between applicant gender and our text-based measures of proposal quality. Specifications 1 through 3 demonstrate this result using the full set of 500 words, while specifications 4 and 5 explore the division between high-scoring and low-scoring words, respectively Finally, in specifications 6 and 7, we focus solely on broad and narrow words, respectively. In all cases, we find strong evidence that female applicants use words that diminish their chances of receiving high scores from reviewers. In addition, we find that female applicants

are at a greater disadvantage in their use of narrow words, compared to their disadvantage in the use of broad words.

Having established the baseline result that female applicants' word choices put them at a disadvantage, we now turn to the impact of those word choices on reviewer scores in Panel C. Our analysis begins in specification 2, where we control for a wide range of text-based metrics such as the count of unique frequent words, grammatical composition, and the grade level[14] of proposal text. While a number of these measures have a significant association with reviewer scores, they do not significantly reduce the observed gender gap in our sample. However, when we add the text-based score predictions starting in specification 3, we see a large drop in the coefficient on female applicants. The inclusion of the score prediction based on narrow words again seems to be a major driver of the gender gap, as it renders the applicant gender effect insignificant, while the score prediction based on broad words fails to do so. When controlling for both measures, we continue to find no significant effect of applicant gender on reviewer scores, even as we add controls for topic areas and applicant publications. Notably, these measures do not eliminate the interaction between female applicants and female reviewers, which remains positive and significant.[15] Table 5 therefore highlights word choice as a crucial driver of gender disparities, even under blinded review.

TABLES 5A, 5B, & 5C HERE

Having covered a range of mechanisms individually, we now synthesize our prior analyses of reviewer scores to examine the relative impact of each dimension in contributing to gender -based

[14] Our measure of grade level is a composite rating based on the arithmetic average of the Flesch-Kincaid, Gunning-Fog, and SMOG measures. See Hengel (2019) for details. We also test for the positive and negative words used in Vinkers et al (2015), but find no gender differences in our sample.

[15] Indeed, in this specification, the sum of the baseline effect and the interaction is positive and significant, indicating that female reviewers have an affirmative preference for female applicants, after controlling for text-based score predictions derived from male applicants' proposals.

disparities. Table 6 presents our results, beginning with the baseline analysis in column 1, which controls only for the round of the program, reviewer fixed effects, and reviewer-cross-round characteristics. Adding gender-based interactions between applicants and reviewers in column 2 increases the magnitude of the applicant gender effect, which now reflects the scoring patterns of male reviewers. Moving to column 3, we add topic-area fixed effects, which reduce the magnitude of the gender gap by approximately ten percent. In column 4, we add publication characteristics, which lead to only a marginal 3% further decline in the gender gap. In column 5, we add the mechanism of repeat applications, which also explains only 3% of the gender score-gap. In column 6, we introduce a range of text-based controls such as word count, grammatical composition, the rate of scientific words, and the grade level of proposal text; these measures offer only a 2% decrease in the gender gap. Finally, in column 7, we control for word choice in the form of our text-based score prediction; this eliminates over 50% of the remaining gender score gap, and renders the overall disparity insignificant. Importantly, these controls do not reduce the interaction effect that we find between female applicants and reviewers, suggesting that female reviewers are not influenced by word choice in the same way as male reviewers. Thus, the main conclusion of Table 6 is that the gender score-gap is driven in large part by the choice of words female applicants use to describe their proposals; after controlling for this dimension, the disparity falls by more than 50% and is no longer statistically significant.

TABLE 6 HERE

D. The Impact of Gender on Innovative Outcomes

So far, our results have focused on explaining the disparities in reviewer scores received by female applicants. While this is inherently valuable as a means of identifying the underlying causes of the lack of demographic diversity in innovative fields, it does not address questions related to efficiency: does the lack of gender diversity in high-scoring applicants reflect an inefficient

disparity and unreasonable barriers, or is it a reflection of the fact that female applicants face challenges that are likely to interfere with their performance, even if they were selected by reviewers? To address these questions, we analyze *ex-post* outcomes for our sample of applicants, looking specifically at publications, NIH grants, and Phase-2 outcomes within the GCE program. Our results begin in Table 7, which focuses specifically on *funded* applicants: conditional on funding, how did different demographic groups perform? This "treatment-on-treated" analysis is meant to capture the organizational perspective: did the evaluation process select the strongest applicants, regardless of demographic characteristics? We consider a wide range of outcomes, and show that in most cases, the review process did indeed select a pool where there were no significant differences between demographic groups. In particular, in columns 1 through 7, after controlling for proposal scores and pre-period outcomes, we find no ex-post disparities in publication-related outcomes across applicant gender. By contrast, in columns 8 and 9, we find (weak) evidence of gender's impact on innovative outcomes. Focusing on the outcome of NIH grants in column 8, we find that female applicants are slightly more likely to obtain such grants in the post-proposal period; we also find a large but insignificant point-estimate for the high-value R01 grants that cover multiple years of research in column 9. Importantly, all specifications also control for word choice, which has a different impact here relative to our earlier tables. Previously, we highlighted that male applicants seemed to benefit in terms of reviewer scores by using general words more often, while female applicants tended to use topic-specific words, which tended to lead to lower reviewer scores. By contrast, Table 7 suggests that while using broad words may help obtain high scores from reviewers, such proposals do not tend to perform as well in terms of ex-post outcomes. In effect, reviewers may well be overly credulous to the broad claims of such proposals, which tend to under-perform across multiple measures if selected for GCE funding.

TABLE 7 HERE

The focus on the sub-sample of funded applicants is valuable, but it is not always the correct perspective for evaluating the relationship between demography and innovation. While Table 7 captures the effectiveness of the *selection* process, it does not offer insight into the *impact* of funding on the applicants in our sample. It may well be the case that some of the applicants receiving funding would have done just as well without it, perhaps because of access to other sources of money or support. To evaluate the causal impact of funding, we therefore need to compare funded applicants against those who (just barely) did not receive funding. In Table 8, we establish a difference-in-differences estimator by focusing on the sub-sample of proposals, which received positive reviews from at least one reviewer. Within this "high-scoring" sample, we examine not only the baseline impact of funding, but also its differential impact across genders.

We begin our analysis in columns 1 and 2, focusing on all published articles and on articles published in top journals,[16] respectively. We find that while funding has only a weakly-positive and insignificant impact on publications, there is a significant positive interaction between funding and female applicants, particularly for top-journal publications. This suggests that Gates foundation funding has a significantly greater impact when it is allocated to female applicants, relative to male applicants. The reason for this difference seems to be driven in part by the negative baseline effect for female applicants: without funding, they under-perform male applicants, and this effect is once again stronger for top-journal publications. A similar pattern appears in columns 3 through 5, which track novelty and exploration by tracking new journals, coauthors, and Medical Subject Heading (MeSH) terms derived from applicants' publications. In all three outcome measures, we find that the baseline effect of being a female applicant is significantly negative, indicating that unfunded women are likely to perform worse than unfunded men. However, we

[16] We define top journals as those in the top decile of impact factor. Our results are robust to redefining top journals as either top-5% or top-25% by impact factor.

also find positive interaction effects between funding and female applicants, leading to no significant difference across gender between funded applicants. Once again, our difference-in-differences results come not from funded women outperforming men, but from unfunded women significantly under-performing unfunded men. In effect, securing funding from the GCE program can "level the playing field" for female applicants, and generates a larger impact than the funding devoted to male applicants.

In line with our earlier results on the importance of funding, we turn to the results focusing on NIH grants in columns 6 and 7. The overall pattern of results is similar in both cases, but the strongest effects can be seen in column 7, which focuses on the coveted multi-year R01 grants. For this outcome, we see negative but insignificant baseline effects for female applicants. More importantly, we find strong positive interactions between female applicants and GCE funding, with the positive interaction effects more-than-compensating for the negative baseline impacts of demographics. We can therefore conclude that female applicants' careers are more responsive to funding in this outcome dimension: the GCE program's impact is most beneficial to these groups when it leads them to obtain additional funding from the NIH, especially through its R01 program. This implies that GCE funding can lead to a "multiplier effect," where it allows successful applicants to be more effective at raising external funding as their careers progress.

Finally, revisiting the impact of proposal text, we once again find that the use of broad and narrow words, do not predict an increase in any of our ex-post outcomes. By contrast, the metric of proposal text grade level, which did not predict reviewer scores in Tables 5C and 6, is now a significant positive predictor of virtually all measures of follow-on innovation. This suggests that while communication style does offer valuable information on the quality of applicants' ideas, reviewers are focusing on the wrong metrics when evaluating the innovative proposals in our sample.

TABLE 8 HERE

V. Discussion and Conclusions

In this paper, we addressed two primary research questions: first, how can innovative organizations increase gender inclusion, and second, what are the potential impacts on innovation if they succeed in doing so? By taking advantage of our unique empirical setting and its blinded-review process, we were able to eliminate the direct effects of bias and discrimination, and focus exclusively on the indirect mechanisms that contribute to the gender disparities in our sample. Our main contributions are the identification of word choice, particularly along the narrow-broad spectrum, as an important driver of negative outcomes for female innovators, and the finding that women may offer a greater return on an organization's resources, in terms of future innovative outcomes. Our findings stand in contrast to those of Goldin and Rouse (2000), who show that anonymous applications are sufficient to significantly reduce disparities faced by female musicians in symphony orchestras. In our setting, significant differences in outcomes persist even under an anonymous evaluation process, suggesting the need for further analysis of the indirect drivers of gender disparity in the pursuit of innovation.

Beyond our primary contributions, a number of related topics would benefit from future research. First, our analysis of innovative outcomes only offers a limited duration of ex-post data following GCE funding decisions, particularly for applicants in later rounds. In this timeframe, we saw evidence that female applicants were able to obtain significantly more additional NIH funding as a result of a successful application. While obtaining funding from the NIH or similar organizations is not an end-goal by itself, this additional funding can be reasonably expected to serve as an input to future research and innovation. Allocating more resources to female innovators may well lead to improved career trajectories and greater innovative output, especially over longer horizons. Indeed, if the findings in our sample apply to the broader scientific and innovative communities,

it is likely that female innovators are systematically under-funded relative to the quality of their ideas. Future work could explore these patterns across longer time horizons and in other stages of the innovation ecosystem.

Second, our text-based analysis focuses on the relatively straightforward measure of the presence or absence of words that appear frequently within our sample. A more sophisticated analysis of a larger sample, potentially in concert with the use of an external corpus of scientific writing, would offer the opportunity to identify detailed patterns in the types of words that either help or harm evaluations, particularly for female innovators.

Finally, we were unable to effectively explore dimensions of diversity and inclusion beyond that of gender in our sample, due to low number of applicants in other under-represented categories (e.g. racial minorities). More work is needed to determine whether our results are generalizable to other dimensions of organizational inclusion. Thus, we would encourage future research to look to other dimensions of diversity, such as ethnicity, national origin, and socio-economic status, as key drivers of the innovative process.

References

Ahlqvist, S., London, B., & Rosenthal, L. (2013). Unstable identity compatibility: How gender rejection sensitivity undermines the success of women in science, technology, engineering, and mathematics fields. *Psychological Science*, *24*(9), 1644-1652.

Azoulay, P., Graff Zivin, J. S., & Manso, G. (2011). Incentives and creativity: evidence from the academic life sciences. *The RAND Journal of Economics*, *42*(3), 527-554.

Becker, G. (1971). *The Economics of Discrimination*. University of Chicago Press.

Bell, A., Chetty, R., Jaravel, X., Petkova, N., & Van Reenen, J. (2018). Who becomes an inventor in America? The importance of exposure to innovation. *The Quarterly Journal of Economics*, *134*(2), 647-713.

Bilimoria, D., Joy, S., & Liang, X. (2008). Breaking barriers and creating inclusiveness: Lessons of organizational transformation to advance women faculty in academic science and engineering. *Human Resource Management*, *47*(3), 423-441.

Bodenhausen, G. V., & Wyer, R. S. (1985). Effects of stereotypes in decision making and information-processing strategies. *Journal of personality and social psychology*, *48*(2), 267.

Bohnet, I., Van Geen, A., & Bazerman, M. (2015). When performance trumps gender bias: Joint vs. separate evaluation. *Management Science*, *62*(5), 1225-1234.
Bornmann, L., Mutz, R., & Daniel, H. D. (2007). Gender differences in grant peer review: A meta-analysis. *Journal of Informetrics*, *1*(3), 226-238.
Boudreau, K. J., Guinan, E. C., Lakhani, K. R., & Riedl, C. (2016). Looking across and looking beyond the knowledge frontier: Intellectual distance, novelty, and resource allocation in science. *Management Science*, *62*(10), 2765-2783.
Brands, R. A., & Fernandez-Mateo, I. (2017). Leaning out: How negative recruitment experiences shape women's decisions to compete for executive roles. *Administrative Science Quarterly*, *62*(3), 405-442.
Brooks, A. W., Huang, L., Kearney, S. W., & Murray, F. E. (2014). Investors prefer entrepreneurial ventures pitched by attractive men. *Proceedings of the National Academy of Sciences*, *111*(12), 4427-4431.
Buser, T. (2016). The impact of losing in a competition on the willingness to seek further challenges. *Management Science*, *62*(12), 3439-3449.
Buser, T., Niederle, M., & Oosterbeek, H. (2014). Gender, competitiveness, and career choices. *The Quarterly Journal of Economics*, *129*(3), 1409-1447.
Castilla, E. J. (2008). Gender, race, and meritocracy in organizational careers. *American Journal of Sociology*, *113*(6), 1479-1526.
Castilla, E. J. (2015). Accounting for the gap: A firm study manipulating organizational accountability and transparency in pay decisions. *Organization Science*, *26*(2), 311-333.
Castilla, E. J., & Benard, S. (2010). The paradox of meritocracy in organizations. *Administrative Science Quarterly*, *55*(4), 543-676.
Ceci, S. J., Ginther, D. K., Kahn, S., & Williams, W. M. (2014). Women in academic science: A changing landscape. *Psychological Science in the Public Interest*, *15*(3), 75-141.
Cook, L. D., & Kongcharoen, C. (2010). *The idea gap in pink and black* (No. w16331). National Bureau of Economic Research.
Correll, S. J. (2001). Gender and the career choice process: The role of biased self-assessments. *American journal of Sociology*, *106*(6), 1691-1730.
Dezsö, C. L., & Ross, D. G. (2012). Does female representation in top management improve firm performance? A panel data investigation. *Strategic Management Journal*, *33*(9), 1072-1089.
Ding, W. W., Murray, F., & Stuart, T. E. (2013). From bench to board: Gender differences in university scientists' participation in corporate scientific advisory boards. *Academy of Management Journal*, *56*(5), 1443-1464.
Earl-Novell, S. (2001). "Gendered" styles of writing and the "inequality in assessment" hypothesis: an explanation for gender differentiation in first class academic achievement at university. *International journal of sociology and social policy*, *21*(1/2), 160-172.
Fernandez, R. M., & Campero, S. (2014). Does Competition Drive Out Discrimination? In *New Haven, CT: Presentation at Economy and Society@ Yale Conference*.
Fernandez, R. M., & Campero, S. (2017). Gender sorting and the glass ceiling in high-tech firms. *ILR Review*, *70*(1), 73-104.
Fernandez, R. M., & Fernandez-Mateo, I. (2006). Networks, race, and hiring. *American sociological review*, *71*(1), 42-71.
Freeman, R. B., & Huang, W. (2015). Collaborating with people like me: Ethnic coauthorship within the United States. *Journal of Labor Economics*, *33*(S1), S289-S318.
Gigerenzer, G., & Gaissmaier, W. (2011). Heuristic decision making. *Annual review of psychology*, *62*, 451-482.

Gill, D., & Prowse, V. (2014). Gender differences and dynamics in competition: The role of luck. *Quantitative Economics*, *5*(2), 351-376.
Ginther, D. K., Kahn, S., & Schaffer, W. T. (2016). Gender, race/ethnicity, and National Institutes of Health R01 research awards: is there evidence of a double bind for women of color? *Academic medicine: journal of the Association of American Medical Colleges*, *91*(8), 1098.
Goldin, C. (2014). A grand gender convergence: Its last chapter. *American Economic Review*, *104*(4), 1091-1119.
Goldin, C., & Rouse, C. (2000). Orchestrating impartiality: The impact of" blind" auditions on female musicians. *American economic review*, *90*(4), 715-741.
Gompers, P. A., & Wang, S. Q. (2017). *Diversity in innovation* (No. w23082). National Bureau of Economic Research.
Hengel, E. (2019). Publishing while Female. Are women held to higher standards? Evidence from peer review. *Working paper.*
Hoffman, M., Kahn, L. B., & Li, D. (2017). Discretion in hiring. *The Quarterly Journal of Economics*, *133*(2), 765-800.
Ibarra, H., & Obodaru, O. (2009). Women and the vision thing. *Harvard business review*, *87*(1), 62-70.
Kamas, L., & Preston, A. (2012). The importance of being confident; gender, career choice, and willingness to compete. *Journal of Economic Behavior & Organization*, *83*(1), 82-97.
Kanze, D., Huang, L., Conley, M. A., & Higgins, E. T. (2018). We ask men to win and women not to lose: Closing the gender gap in startup funding. *Academy of Management Journal*, *61*(2), 586-614.
Kelly, E. L., Ammons, S. K., Chermack, K., & Moen, P. (2010). Gendered challenge, gendered response: Confronting the ideal worker norm in a white-collar organization. *Gender & Society*, *24*(3), 281-303.
Kunze, A., & Miller, A. R. (2017). Women helping women? Evidence from private sector data on workplace hierarchies. *Review of Economics and Statistics*, *99*(5), 769-775.
Lerchenmueller, M. J., & Sorenson, O. (2018). The gender gap in early career transitions in the life sciences. *Research Policy*, *47*(6), 1007-1017.
Lincoln, A. E., Pincus, S., Koster, J. B., & Leboy, P. S. (2012). The Matilda Effect in science: Awards and prizes in the US, 1990s and 2000s. *Social studies of science*, *42*(2), 307-320.
London, B., Downey, G., Romero-Canyas, R., Rattan, A., & Tyson, D. (2012). Gender-based rejection sensitivity and academic self-silencing in women. *Journal of personality and social psychology*, *102*(5), 961.
Magua, W., Zhu, X., Bhattacharya, A., Filut, A., Potvien, A., Leatherberry, R., ... & Kaatz, A. (2017). Are female applicants disadvantaged in National Institutes of Health peer review? Combining algorithmic text mining and qualitative methods to detect evaluative differences in R01 reviewers' critiques. *Journal of Women's Health*, *26*(5), 560-570.
Maitland, E., & Sammartino, A. (2015). Decision making and uncertainty: The role of heuristics and experience in assessing a politically hazardous environment. *Strategic Management Journal*, *36*(10), 1554-1578.
Marschke, G., Nunez, A., Weinberg, B. A., & Yu, H. (2018). Last Place? The Intersection between Ethnicity, Gender, and Race in Biomedical Authorship.
McIntyre, S., Moberg, D. J., & Posner, B. Z. (1980). Preferential treatment in preselection decisions according to sex and race. *Academy of Management Journal*, *23*(4), 738-749.

McKown, C., & Weinstein, R. S. (2002). Modeling the Role of Child Ethnicity and Gender in Children's Differential Response to Teacher Expectations 1. *Journal of Applied Social Psychology*, *32*(1), 159-184.

Monroe, K., Ozyurt, S., Wrigley, T., & Alexander, A. (2008). Gender equality in academia: Bad news from the trenches, and some possible solutions. *Perspectives on politics*, *6*(2), 215-233.

Mousavi, S., & Gigerenzer, G. (2014). Risk, uncertainty, and heuristics. *Journal of Business Research*, *67*(8), 1671-1678.

Murray, F., & Graham, L. (2007). Buying science and selling science: gender differences in the market for commercial science. *Industrial and Corporate Change*, *16*(4), 657-689.

Niederle, M., & Vesterlund, L. (2007). Do women shy away from competition? Do men compete too much?. *The quarterly journal of economics*, *122*(3), 1067-1101.

O'Neill, J. (2003). The gender gap in wages, circa 2000. *American Economic Review*, *93*(2), 309-314.

Østergaard, C. R., Timmermans, B., & Kristinsson, K. (2011). Does a different view create something new? The effect of employee diversity on innovation. *Research Policy*, *40*(3), 500-509.

Poppenhaeger, K. (2017). Unconscious Gender Bias in Academia: from PhD Students to Professors. *arXiv preprint arXiv:1711.00344*.

Reagans, R., & Zuckerman, E. W. (2001). Networks, diversity, and productivity: The social capital of corporate R&D teams. *Organization science*, *12*(4), 502-517.

Reuben, E., Sapienza, P., & Zingales, L. (2014). How stereotypes impair women's careers in science. *Proceedings of the National Academy of Sciences*, 201314788.

Robinson, G., & Dechant, K. (1997). Building a business case for diversity. *Academy of Management Perspectives*, *11*(3), 21-31.

Schmader, T., Whitehead, J., & Wysocki, V. H. (2007). A linguistic comparison of letters of recommendation for male and female chemistry and biochemistry job applicants. *Sex roles*, *57*(7-8), 509-514.

Shen, H. (2013). Mind the gender gap. *Nature*, *495*(7439), 22.

Sood, G., & Laohaprapanon, S. (2018). Predicting Race and Ethnicity from the Sequence of Characters in a Name. *arXiv preprint arXiv:1805.02109*.

Tasheva, S. N., & Hillman, A. (2018). Integrating diversity at different levels: multi-level human capital, social capital, and demographic diversity and their implications for team effectiveness. *Academy of Management Review*, (ja).

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *science*, *185*(4157), 1124-1131.

Vinkers, C. H., Tijdink, J. K., & Otte, W. M. (2015). Use of positive and negative words in scientific PubMed abstracts between 1974 and 2014: retrospective analysis. *Bmj*, *351*, h6467.

Witteman, H. O., Hendricks, M., Straus, S., & Tannenbaum, C. (2017). Female grant applicants are equally successful when peer reviewers assess the science, but not when they assess the scientist. *bioRxiv*, 232868.

Wold, A., & Wenneras, C. (2010). Nepotism and sexism in peer-review. In *Women, science, and technology* (pp. 64-70). Routledge.

Figure 1: Distribution of Proposals, Applicants, and Scores by Topic
Observations: 6794 Total Proposals, 1042 Silver Ratings, 230 Gold Ratings

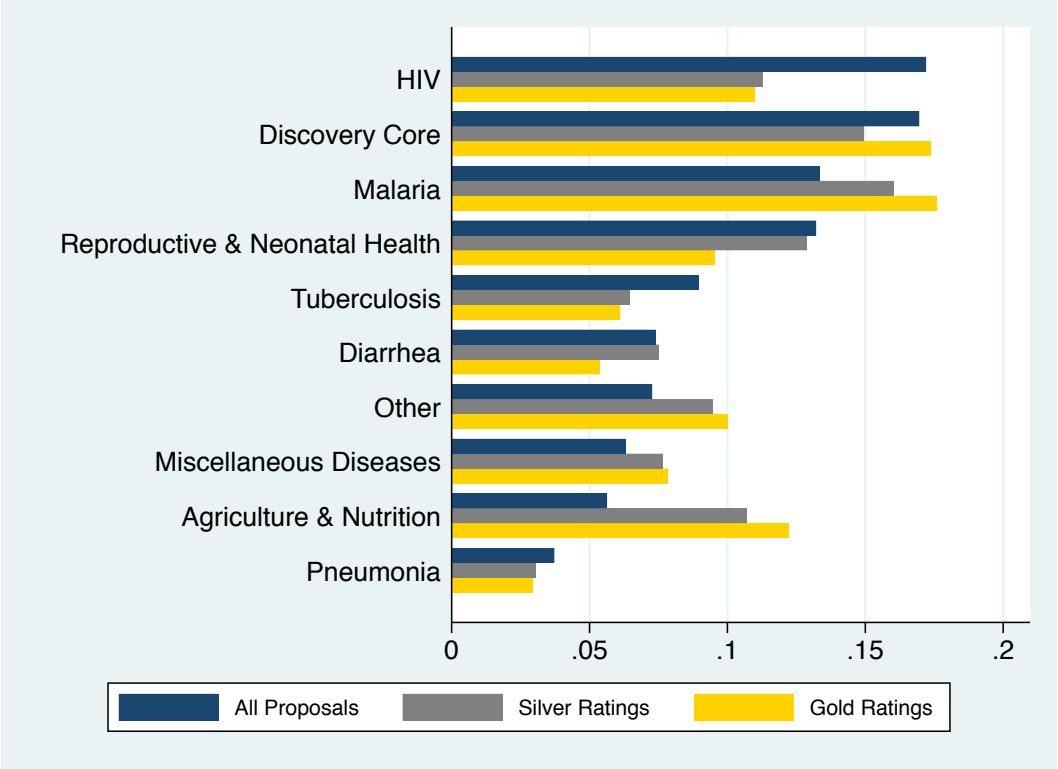

Figure 2: Rates of Female Applicants and Average Scores by Topic
Note: Circle areas represent the number of proposals in each topic. Total Proposals: 6794

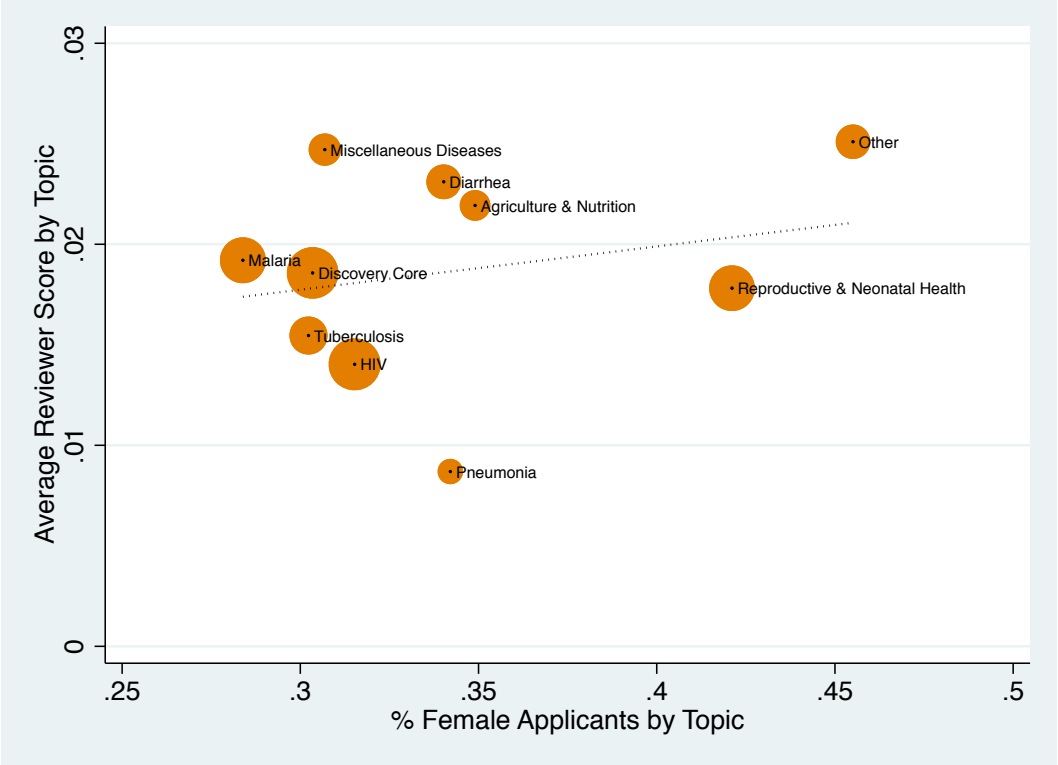

Figure 3: Male vs. Female Word Use Rates for All Frequent Words

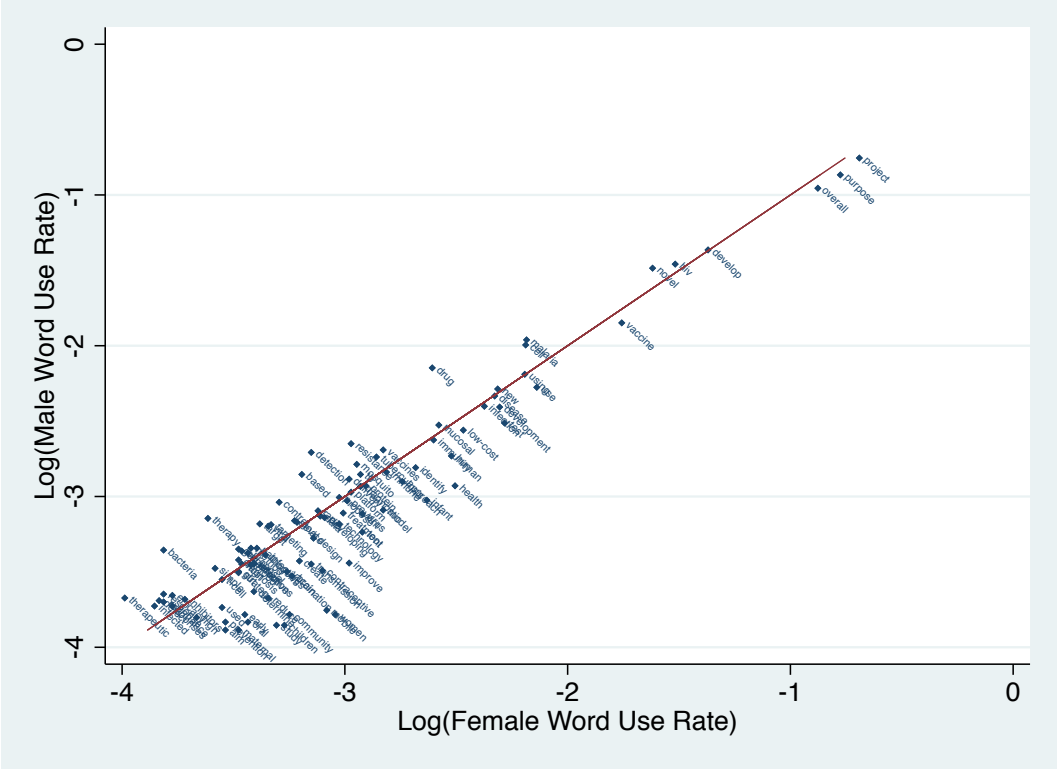

Figure 4: Male vs. Female Word Use Rates for Broad vs. Narrow Words

Note: Words selected based on high use frequency and significant correlation with proposal score. Narrow (broad) words are defined based on a high (low) variance in use rate across proposal topics.

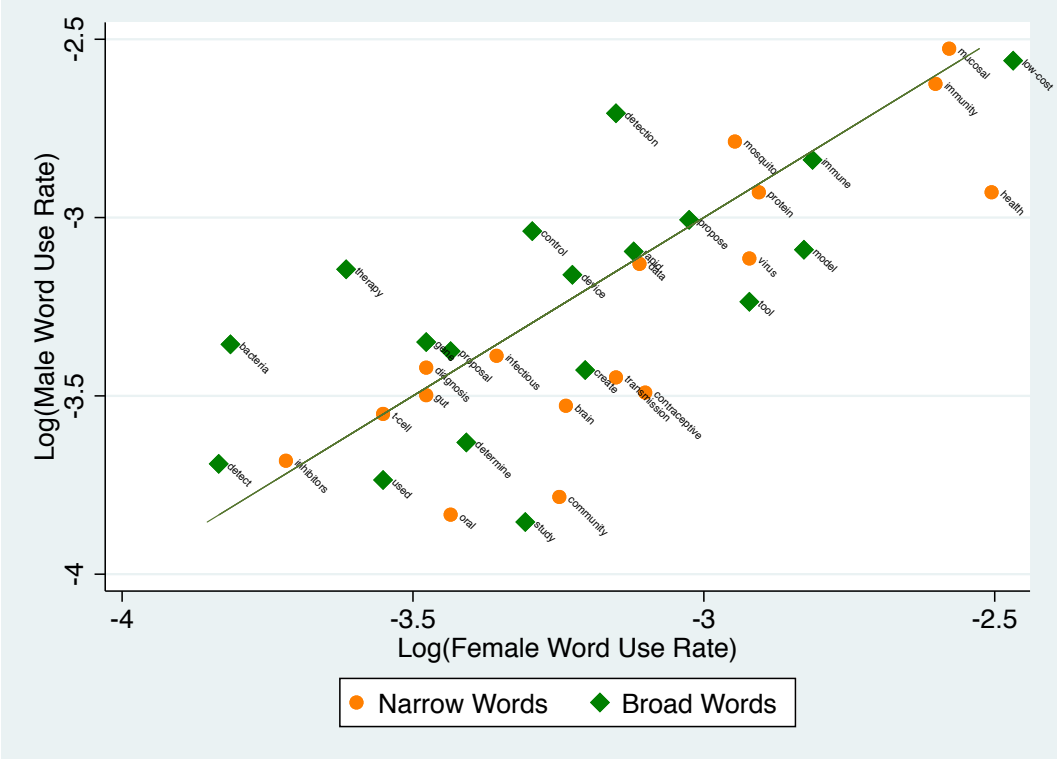

Figure 5: High-Scoring vs. Low-Scoring Word Use Rates for Gendered Frequent Words
Note: Words selected based on high use frequency and significant correlation with applicant gender.

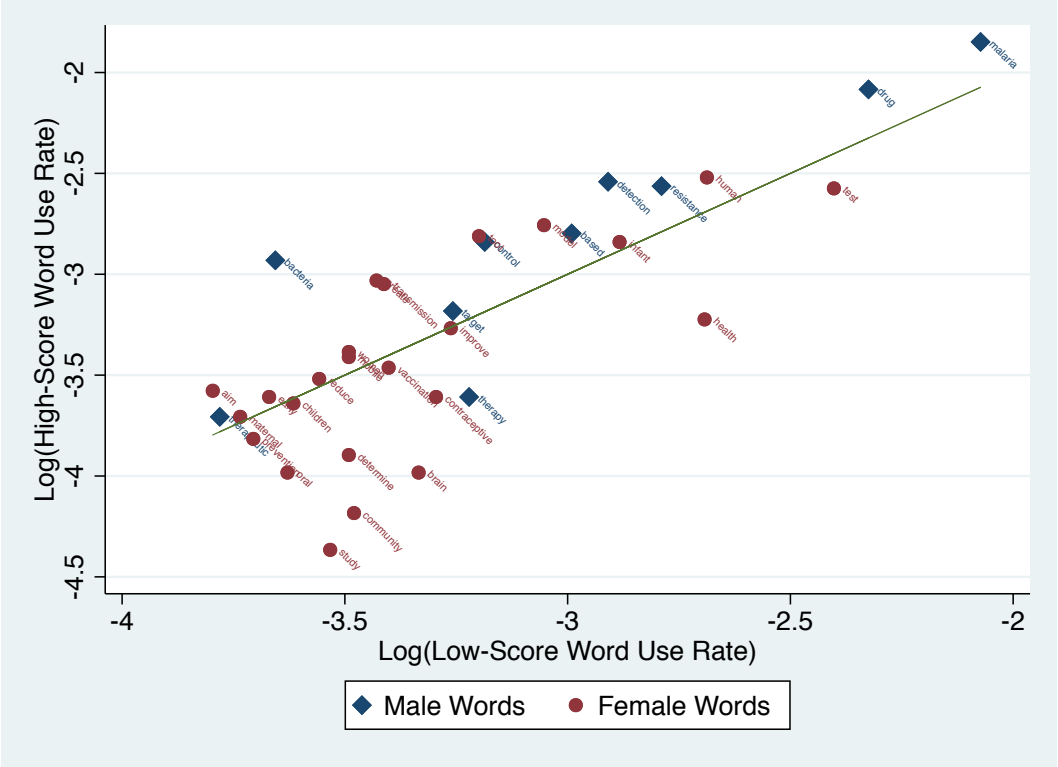

Figure 6: Frequent Words with Disparities in Reviewer Scores and Gender-Based Use

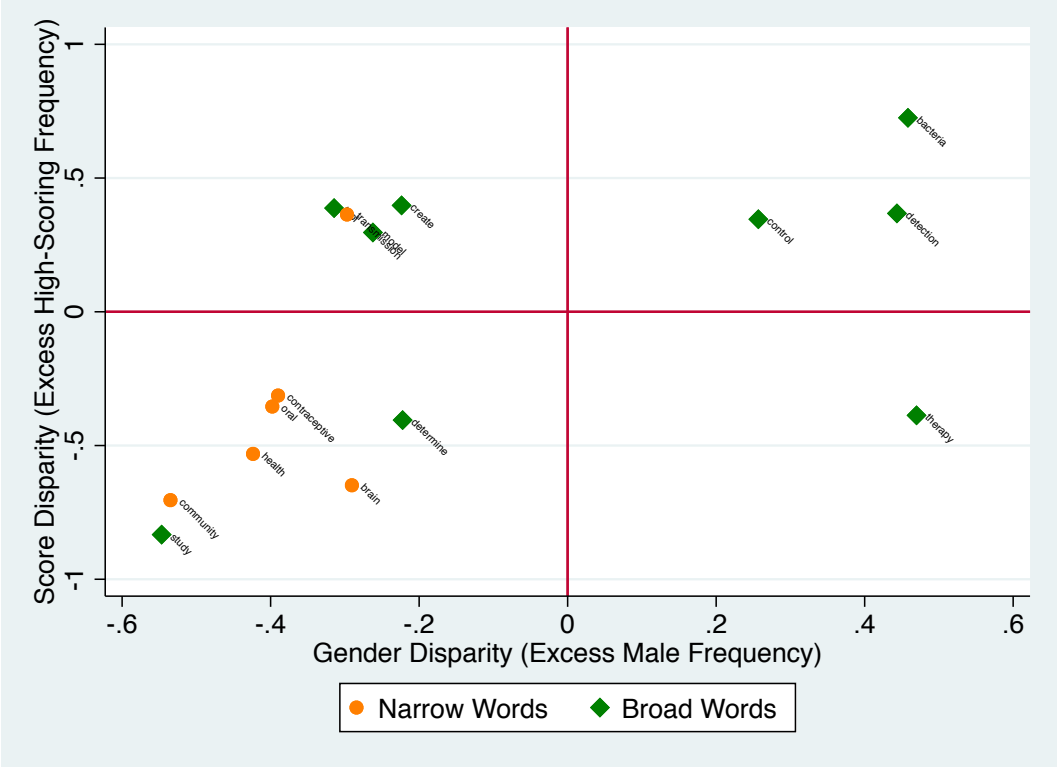

Table 1: Summary Statistics

Variable	N. of Obs.	Mean	Std. Dev.	Min	Max
APPLICANT CHARACTERISTICS					
Female Applicant Probability	5,058	0.343	0.475	0	1
Applicant Publication History Indicator	5,058	0.785	0.411	0	1
PROPOSAL CHARACTERISTICS					
GCE Round	6,794	6.083	5.198	1	19
Funding Indicator	6,794	0.093	0.291	0	1
High-Score Indicator	6,794	0.174	0.379	0	1
Reviewer Count	6,794	3.158	1.324	1	7
Repeat-Applicant-After-Failure Indicator	6,794	0.204	0.403	0	1
Proposal Text Characteristics:					
Unique Frequent Word Count	6,794	11.032	4.530	0	75
Noun Share	6,794	0.370	0.096	0.00	1.00
Adjective Share	6,794	0.163	0.079	0.00	0.75
Verb Share	6,794	0.121	0.062	0	0.5
Proposal Text Grade Level	6,794	15.954	3.035	1.78	30.65
Normalized Text-Based Score: Broad Words	6,794	0.000	1.000	-9.01	1.22
Normalized Text-Based Score: Narrow Words	6,794	0.000	1.000	-15.26	3.88
Conditional on Identifying Publication History:					
Pre-Period Publications	5,448	9.980	10.915	0	117
Post-Period Publications	5,448	16.122	20.513	0	287
Pre-Period NIH Grants	5,448	0.571	1.199	0	20
Post-Period NIH Grants	5,448	0.857	1.672	0	17
REVIEWER CHARACTERISTICS					
Female Reviewer Probability	132	0.161	0.355	0	1
Avg. Proposals per Round (All-GCE)	132	117.312	43.185	31	206
Avg. Proposals per Round (US Academics)	132	47.950	24.964	5.5	99
REVIEWER X ROUND CHARACTERISTICS					
Proposals Under Review	429	116.9	50.5	31	210
Reviewer Round Sequence	429	3.3	2.5	1	13
REVIEWER X PROPOSAL CHARACTERISTICS					
Proposal Sequence	21,453	71.6	48.9	1	208
Silver Rating	21,453	0.049	0.215	0	1
Gold Rating	21,453	0.011	0.103	0	1
Reviewer Score	21,453	0.020	0.111	0	1

Table 2: Impact of Applicant and Reviewer Gender on Reviewer Scores

VARIABLES	(1)	(2)	(3)	(4)	(5)
	DV = Reviewer Score				
Female Applicant	-0.165***	-0.163***	-0.163***	-0.141**	-0.194***
	(0.062)	(0.062)	(0.063)	(0.064)	(0.071)
Female Reviewer		0.052			
		(0.053)			
Female Applicant X Female Reviewer					0.364***
					(0.138)
Reviewer X Round Characteristics:					
Log(Proposals Under Review)	-0.812***	-0.840***	-0.658***	-0.759***	-0.764***
	(0.104)	(0.107)	(0.137)	(0.149)	(0.149)
Log(Reviewer Round Sequence)	0.008	-0.006	0.188**	0.145*	0.141*
	(0.038)	(0.038)	(0.086)	(0.086)	(0.085)
Log(Proposal Sequence)	-0.088**	-0.088**	-0.091**	-0.090**	-0.090**
	(0.036)	(0.036)	(0.036)	(0.037)	(0.037)
Round FEs	Y	Y	Y	Y	Y
Reviewer FEs	N	N	Y	Y	Y
Topic Area FEs	N	N	N	Y	Y
Observations	21,453	21,453	21,453	21,453	21,453
Pseudo R-squared	0.0315	0.0317	0.0392	0.0432	0.0435

Ordered logit specification; Robust standard errors clustered by reviewer in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 3A: Impact of Applicant Gender on Pre-Period Publications

VARIABLES	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	DV = Log(Applicant Career Length)		DV = Log(Pre-Period Publications)		DV = Share of Top-Journal Pre-Period Publications		DV = Share of Last-Author Pre-Period Publications	
Female Applicant	-0.377***	-0.385***	-0.310***	-0.112***	-0.023***	-0.011	-0.071***	-0.022**
	(0.034)	(0.034)	(0.035)	(0.031)	(0.008)	(0.008)	(0.010)	(0.009)
Log(Applicant Career Length)				0.512***		0.028***		0.133***
				(0.014)		(0.004)		(0.003)
Round FEs	Y	Y	Y	Y	Y	Y	Y	Y
Topic Area FEs	N	Y	N	Y	N	Y	N	Y
Observations	4,005	4,005	4,005	4,005	4,005	4,005	4,005	4,005
Pseudo R-squared	0.043	0.054	0.025	0.249	0.012	0.030	0.037	0.220

OLS specification; Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 3B: Impact of Applicant Gender and Publication History on Reviewer Scores

VARIABLES	(1)	(2)	(3)	(4)	(5)
	DV = Reviewer Score				
Female Applicant	-0.163***	-0.164**	-0.159**	-0.134**	-0.187***
	(0.063)	(0.065)	(0.065)	(0.066)	(0.070)
Female Applicant X Female Reviewer					0.359***
					(0.139)
Publication History Not Available		-0.040	-0.015	0.018	0.017
		(0.131)	(0.129)	(0.129)	(0.129)
Log(Applicant Career Length)		-0.067*	-0.079*	-0.067	-0.067
		(0.040)	(0.042)	(0.042)	(0.042)
Log(Pre-Period Publications)		0.088**	0.067	0.063	0.063
		(0.043)	(0.046)	(0.046)	(0.046)
Share of Top-Journal Pubs			0.349**	0.346**	0.344**
			(0.138)	(0.139)	(0.139)
Share of Last-Author Pubs			0.102	0.092	0.091
			(0.111)	(0.115)	(0.115)
Round FEs	Y	Y	Y	Y	Y
Reviewer FEs	Y	Y	Y	Y	Y
Topic Area FEs	N	N	N	Y	Y
Reviewer X Round Controls	Y	Y	Y	Y	Y
Observations	21,453	21,453	21,453	21,453	21,453
Pseudo R-squared	0.0391	0.0396	0.0402	0.0441	0.0444

Ordered logit specification; Robust standard errors clustered by reviewer in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 4A: Applicant Propensity to Re-Apply After Rejection

VARIABLES	(1)	(2)	(3)	(4)
	DV = Repeat Applicant After Rejection			
Female Applicant	-0.186**	-0.177**	-0.098	-0.100
	(0.084)	(0.084)	(0.086)	(0.086)
Log(Applicant Career Length)			0.256***	0.262***
			(0.049)	(0.057)
Log(Pre-Period Publications)				0.059
				(0.047)
Share of Top-Journal Pubs				-0.389**
				(0.195)
Round FEs	Y	Y	Y	Y
Topic Area FEs	N	Y	Y	Y
Additional Publication Characteristics	N	N	N	Y
Observations	4,496	4,462	4,462	4,462
Pseudo R-squared	0.0813	0.0860	0.0940	0.0952

Logit specification; Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 4B: Impact of Repeat Applicants on Reviewer Scores

VARIABLES	(1)	(2)	(3)	(4)	(5)	(6)
	DV = Reviewer Score					
Female Applicant	-0.163***	-0.151**	-0.130**	-0.127*	-0.121*	-0.174**
	(0.063)	(0.063)	(0.064)	(0.065)	(0.070)	(0.078)
Female Applicant X Repeat Applicant					-0.022	-0.020
					(0.182)	(0.182)
Female Applicant X Female Reviewer						0.358**
						(0.139)
Repeat Applicant After Rejection		0.256***	0.252***	0.256***	0.262***	0.262***
		(0.061)	(0.064)	(0.064)	(0.080)	(0.080)
Log(Applicant Career Length)				-0.079*	-0.079*	-0.078*
				(0.042)	(0.042)	(0.042)
Log(Pre-Period Publications)				0.062	0.062	0.062
				(0.046)	(0.046)	(0.046)
Share of Top-Journal Pubs				0.349**	0.349**	0.347**
				(0.139)	(0.139)	(0.139)
Share of Last-Author Pubs				0.092	0.092	0.091
				(0.115)	(0.114)	(0.114)
Round FEs	Y	Y	Y	Y	Y	Y
Reviewer FEs	Y	Y	Y	Y	Y	Y
Topic Area FEs	N	N	Y	Y	Y	Y
Additional Publication Characteristics	N	N	N	Y	Y	Y
Reviewer X Round Controls	Y	Y	Y	Y	Y	Y
Observations	21,453	21,453	21,453	21,453	21,453	21,453
Pseudo R-squared	0.0392	0.0406	0.0445	0.0455	0.0455	0.0458

Ordered logit specification; Robust standard errors clustered by reviewer in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 5A: Applicant Characteristics and Proposal Word Choice

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	DV = Proposal Text Unique Word Count						
VARIABLES	All Frequent Words	High-Scoring Words	Low-Scoring Words	Male High-Scoring Words	Male Low-Scoring Words	Broad Words	Narrow Words
Female Applicant	0.007	-0.026***	0.055***	-0.019***	0.040***	-0.014**	0.024**
	(0.007)	(0.007)	(0.014)	(0.007)	(0.014)	(0.007)	(0.011)
Log(Frequent Word Count)		1.123***	1.075***	1.119***	1.084***	1.139***	1.060***
		(0.014)	(0.025)	(0.013)	(0.025)	(0.013)	(0.024)
Noun Share	-0.027	-0.317***	0.584***	-0.362***	0.684***	-0.330***	0.481***
	(0.045)	(0.039)	(0.076)	(0.038)	(0.079)	(0.038)	(0.062)
Adjective Share	0.573***	0.484***	-0.950***	0.402***	-0.799***	0.303***	-0.443***
	(0.051)	(0.046)	(0.095)	(0.045)	(0.097)	(0.046)	(0.075)
Verb Share	0.554***	0.328***	-0.592***	0.264***	-0.448***	0.418***	-0.602***
	(0.063)	(0.057)	(0.115)	(0.056)	(0.116)	(0.057)	(0.093)
Proposal Text Grade Level	0.036***	-0.005***	0.014***	-0.001	0.006**	-0.003*	0.006***
	(0.002)	(0.001)	(0.003)	(0.001)	(0.003)	(0.001)	(0.002)
Log(Applicant Career Length)	-0.009*	-0.003	0.007	-0.005	0.009	0.003	-0.005
	(0.004)	(0.004)	(0.009)	(0.004)	(0.009)	(0.004)	(0.007)
Share of Top-Journal Pubs	0.005	0.051***	-0.116***	0.046***	-0.103***	0.011	-0.019
	(0.016)	(0.015)	(0.035)	(0.015)	(0.033)	(0.015)	(0.026)
Round FEs	Y	Y	Y	Y	Y	Y	Y
Topic Area FEs	Y	Y	Y	Y	Y	Y	Y
Additional Text-Based Controls	Y	Y	Y	Y	Y	Y	Y
Applicant Publication Characteristics	Y	Y	Y	Y	Y	Y	Y
Observations	6,794	6,794	6,794	6,794	6,794	6,794	6,794
R-squared	0.212	0.286	0.134	0.281	0.146	0.270	0.147

Poisson Specification; Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 5B: Applicant Characteristics and Text-Based Score Predictions

The dependent variable is the fitted probability of an ordered-logit regression predicting a proposal score from reviewers, based purely on the words contained in the proposal's title and description, and calibrated using only the proposals of male applicants. Specifically, the outcome variable predicts reviewer scores based on the presence or absence of the 500 frequent words with the greatest impact on reviewer scores, further subdivided across words with a positive impact ("high-scoring") and words with a negative impact ("low-scoring") on reviewer scores in specifications 4 and 5, and across broad and narrow words (based on the standard deviation of word use across topics) in specifications 6 and 7.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	DV = Text-Based Reviewer Score Prediction						
VARIABLES	Basis: All Frequent Words			Basis: High-Scoring Words	Basis: Low-Scoring Words	Basis: Broad Words	Basis: Narrow Words
Female Applicant	-0.139***	-0.116***	-0.106***	-0.095***	-0.084***	-0.060**	-0.155***
	(0.029)	(0.029)	(0.029)	(0.029)	(0.029)	(0.029)	(0.032)
Log(Frequent Word Count)		-0.335***	-0.337***	-0.359***	-0.419***	-0.076*	0.532***
		(0.044)	(0.044)	(0.045)	(0.044)	(0.044)	(0.047)
Noun Share		-0.277*	-0.266*	-0.221	-0.316**	-0.199	0.065
		(0.155)	(0.155)	(0.149)	(0.153)	(0.166)	(0.200)
Adjective Share		0.540***	0.533***	0.371**	0.456***	0.437**	0.472**
		(0.168)	(0.168)	(0.162)	(0.166)	(0.180)	(0.199)
Verb Share		0.166	0.166	0.250	0.181	-0.098	-0.199
		(0.199)	(0.199)	(0.191)	(0.197)	(0.216)	(0.254)
Proposal Text Grade Level		-0.007	-0.007	0.002	-0.009*	-0.016***	0.001
		(0.005)	(0.005)	(0.006)	(0.005)	(0.005)	(0.005)
Log(Applicant Career Length)			0.018	0.034*	0.022	-0.016	-0.027
			(0.019)	(0.018)	(0.019)	(0.020)	(0.017)
Share of Top-Journal Pubs			0.054	-0.043	0.049	0.163***	0.036
			(0.056)	(0.057)	(0.055)	(0.056)	(0.061)
Round FEs	Y	Y	Y	Y	Y	Y	Y
Topic Area FEs	N	Y	Y	Y	Y	Y	Y
Additional Text-Based Controls	N	Y	Y	Y	Y	Y	Y
Additional Publication Characteristics	N	N	Y	Y	Y	Y	Y
Observations	6,794	6,794	6,794	6,794	6,794	6,794	6,794
R-squared	0.069	0.094	0.095	0.121	0.104	0.018	0.086

OLS Specification; Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 5C: Impact of Proposal Text on Reviewer Scores

This table takes the outcome variables for "broad" and "narrow" words from Table 5A and uses them to predict reviewer scores alongside the effect of being a female applicant. Importantly, the text-based score predictions are calibrated based only on proposals from male applicants, and then calculated for all proposals based on the presence or absence of the 500 frequent words with the greatest impact on reviewer scores in each proposal's title and description, sub-divided into broad and narrow words based on the standard deviation of word use rates across topics.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
VARIABLES	DV = Reviewer Score							
Female Applicant	-0.163***	-0.156**	-0.116*	-0.075	-0.052	-0.033	-0.027	-0.085
	(0.063)	(0.064)	(0.065)	(0.064)	(0.064)	(0.064)	(0.066)	(0.069)
Female Applicant X Female Reviewer								0.400**
								(0.157)
Score Prediction from Broad Words			0.904**		0.403***	0.401***	0.397***	0.398***
			(0.378)		(0.110)	(0.109)	(0.108)	(0.109)
Score Prediction from Narrow Words				0.663***	0.584***	0.580***	0.578***	0.579***
				(0.060)	(0.063)	(0.065)	(0.066)	(0.065)
Log(Frequent Word Count)		-0.122	-0.182	-0.619***	-0.589***	-0.610***	-0.614***	-0.615***
		(0.136)	(0.142)	(0.143)	(0.145)	(0.148)	(0.148)	(0.148)
Noun Share		-0.795**	-0.616	-0.630*	-0.553	-0.490	-0.456	-0.454
		(0.383)	(0.383)	(0.380)	(0.381)	(0.384)	(0.383)	(0.382)
Adjective Share		-0.213	-0.435	-0.419	-0.544	-0.372	-0.376	-0.365
		(0.433)	(0.442)	(0.434)	(0.438)	(0.461)	(0.460)	(0.460)
Verb Share		0.259	0.293	0.349	0.346	0.369	0.372	0.366
		(0.610)	(0.613)	(0.609)	(0.616)	(0.627)	(0.629)	(0.629)
Proposal Text Grade Level		-0.005	0.003	-0.007	-0.001	-0.005	-0.006	-0.006
		(0.016)	(0.016)	(0.015)	(0.015)	(0.015)	(0.015)	(0.015)
Log(Applicant Career Length)							-0.045	-0.044
							(0.043)	(0.043)
Share of Top-Journal Pubs							0.263*	0.261*
							(0.147)	(0.147)
Round FEs	Y	Y	Y	Y	Y	Y	Y	Y
Reviewer FEs	Y	Y	Y	Y	Y	Y	Y	Y
Topic Area FEs	N	N	N	N	N	Y	Y	Y
Additional Text-Based Controls	N	Y	Y	Y	Y	Y	Y	Y
Additional Publication Characteristics	N	N	N	N	N	N	Y	Y
Reviewer X Round Controls	Y	Y	Y	Y	Y	Y	Y	Y
Observations	21,453	21,453	21,453	21,453	21,453	21,453	21,453	21,453
Pseudo R-squared	0.0392	0.0401	0.0538	0.0667	0.0726	0.0757	0.0763	0.0766

Ordered logit specification; Robust standard errors clustered by reviewer in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Table 6: Combined Effects of All Explanatory Variables on Reviewer Scores

VARIABLES	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	DV = Reviewer Score						
Female Applicant	-0.163***	-0.214***	-0.194***	-0.186***	-0.180**	-0.173**	-0.076
	(0.063)	(0.069)	(0.071)	(0.070)	(0.070)	(0.070)	(0.069)
Female Applicant X Female Reviewer		0.348**	0.364***	0.357**	0.358**	0.358**	0.398**
		(0.143)	(0.138)	(0.139)	(0.139)	(0.140)	(0.157)
Repeat Applicant After Rejection					0.256***	0.257***	0.259***
					(0.064)	(0.064)	(0.063)
Log(Frequent Word Count)						-0.136	-0.618***
						(0.141)	(0.149)
Noun Share						-0.711*	-0.447
						(0.384)	(0.380)
Adjective Share						-0.006	-0.322
						(0.451)	(0.457)
Verb Share						0.322	0.434
						(0.613)	(0.624)
Proposal Text Grade Level						-0.012	-0.006
						(0.016)	(0.015)
Score Prediction from Broad Words							0.397***
							(0.108)
Score Prediction from Narrow Words							0.580***
							(0.066)
Log(Applicant Career Length)				-0.066	-0.078*	-0.078*	-0.055
				(0.042)	(0.042)	(0.042)	(0.043)
Log(Pre-Period Publications)				0.062	0.062	0.058	0.032
				(0.046)	(0.046)	(0.046)	(0.048)
Share of Top-Journal Pubs				0.342**	0.348**	0.347**	0.259*
				(0.139)	(0.139)	(0.139)	(0.147)
Share of Last-Author Pubs				0.092	0.091	0.096	0.127
				(0.115)	(0.115)	(0.115)	(0.116)
Round FEs	Y	Y	Y	Y	Y	Y	Y
Reviewer FEs	Y	Y	Y	Y	Y	Y	Y
Topic Area FEs	N	N	Y	Y	Y	Y	Y
Additional Text-Based Controls	N	N	N	N	N	Y	Y
Reviewer X Round Controls	Y	Y	Y	Y	Y	Y	Y
Observations	21,453	21,453	21,453	21,453	21,453	21,453	21,453
Pseudo R-squared	0.0392	0.0395	0.0435	0.0445	0.0458	0.0466	0.0779

Ordered logit specification; Robust standard errors clustered by reviewer in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Table 7: Impact of Applicant Gender on Post-Period Outcomes

VARIABLES	(1) Phase 2 Applications	(2) Phase 2 Successes	(3) Article Count	(4) Top-Journal Article Count	(5) New Journal Count	(6) New Coauthor Count	(7) New MeSH Count	(8) NIH Grant Count	(9) NIH R01 Count
Female Applicant	-0.022	-0.109	0.011	-0.009	0.001	0.010	-0.014	0.373*	0.339
	(0.104)	(0.511)	(0.082)	(0.145)	(0.116)	(0.112)	(0.108)	(0.218)	(0.303)
Average Proposal Score	0.010	1.294	-0.090	-0.210	-0.070	0.111	-0.034	0.749	-0.157
	(0.206)	(1.069)	(0.201)	(0.342)	(0.281)	(0.316)	(0.229)	(0.493)	(1.128)
Score Prediction from Broad Words	0.030	-0.214	0.015	-0.128**	0.050	-0.187**	0.016	-0.223***	-0.261**
	(0.049)	(0.234)	(0.038)	(0.063)	(0.047)	(0.080)	(0.041)	(0.075)	(0.118)
Score Prediction from Narrow Words	0.006	-0.026	0.021	0.038	-0.009	0.037*	0.012	-0.074***	0.046
	(0.020)	(0.122)	(0.014)	(0.030)	(0.018)	(0.021)	(0.022)	(0.027)	(0.058)
Proposal Text Grade Level	0.009	0.020	0.015*	0.030**	0.023**	0.015	0.018*	0.038	0.015
	(0.010)	(0.042)	(0.008)	(0.014)	(0.011)	(0.012)	(0.010)	(0.026)	(0.031)
Pre-Period Article Count	0.036	0.178	0.836***	0.383***	0.164*	0.229**	0.320***	0.367***	0.676***
	(0.045)	(0.172)	(0.043)	(0.102)	(0.099)	(0.093)	(0.092)	(0.132)	(0.162)
Pre-Period Focal Outcome				0.584***	0.458***	0.561***	0.331***	0.719***	0.742***
				(0.089)	(0.115)	(0.069)	(0.083)	(0.141)	(0.285)
Additional Controls:									
Round FEs	Y	Y	Y	Y	Y	Y	Y	Y	Y
Topic Area FEs	Y	Y	Y	Y	Y	Y	Y	Y	Y
Applicant Career Length FEs	Y	Y	Y	Y	Y	Y	Y	Y	Y
Sample of Applicants:	Funded & Active	Funded & Active	Funded & Active	Funded & Active	Funded & Active	Funded & Active	Funded & Active	Funded & Active	Funded & Active
Observations	500	500	500	500	500	500	500	500	500
Pseudo R-squared	0.0525	0.368	0.476	0.337	0.225	0.628	0.490	0.377	0.350

Poisson specification; Robust standard errors clustered by applicant in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Table 8: Effectiveness of Funding by Applicant Gender

VARIABLES	(1) Article Count	(2) Top-Journal Article Count	(3) New Journal Count	(4) New Coauthor Count	(5) New MeSH Count	(6) NIH Grant Count	(7) NIH R01 Count
Female Applicant	-0.086	-0.265**	-0.200**	-0.259**	-0.199**	-0.022	-0.208
	(0.067)	(0.113)	(0.092)	(0.102)	(0.087)	(0.182)	(0.252)
Female Applicant X Funding	0.195**	0.404**	0.378***	0.376***	0.302**	0.411*	0.868**
	(0.095)	(0.160)	(0.131)	(0.134)	(0.133)	(0.237)	(0.397)
Funding Indicator	0.053	0.139	-0.032	0.187**	0.067	0.113	-0.143
	(0.054)	(0.093)	(0.074)	(0.085)	(0.067)	(0.161)	(0.262)
Average Proposal Score	-0.113	-0.659	0.104	-0.223	-0.132	-0.508	-1.265
	(0.233)	(0.436)	(0.356)	(0.345)	(0.271)	(0.512)	(1.160)
Score Prediction from Broad Words	0.030	-0.028	0.033	-0.035	0.054	-0.171***	-0.143*
	(0.031)	(0.043)	(0.034)	(0.082)	(0.040)	(0.064)	(0.087)
Score Prediction from Narrow Words	0.007	0.034	0.001	0.018	0.004	-0.020	0.018
	(0.013)	(0.022)	(0.018)	(0.021)	(0.015)	(0.025)	(0.031)
Proposal Text Grade Level	0.015**	0.029***	0.020**	0.021**	0.022***	0.038**	0.024
	(0.006)	(0.010)	(0.008)	(0.010)	(0.008)	(0.019)	(0.026)
Pre-Period Article Count	0.841***	0.532***	0.217***	0.240***	0.309***	0.423***	0.475***
	(0.027)	(0.071)	(0.065)	(0.059)	(0.064)	(0.086)	(0.121)
Pre-Period Focal Outcome		0.435***	0.426***	0.480***	0.341***	0.878***	1.561***
		(0.069)	(0.084)	(0.042)	(0.064)	(0.102)	(0.218)
Additional Controls:							
Round FEs	Y	Y	Y	Y	Y	Y	Y
Topic Area FEs	Y	Y	Y	Y	Y	Y	Y
Applicant Career Length FEs	Y	Y	Y	Y	Y	Y	Y
Pre-Period Applicant Characteristics	Y	Y	Y	Y	Y	Y	Y
Sample of Applicants:	High-Scoring & Active	High-Scoring & Active	High-Scoring & Active	High-Scoring & Active	High-Scoring & Active	High-Scoring & Active	High-Scoring & Active
Observations	1,163	1,163	1,163	1,163	1,163	1,163	1,163
Pseudo R-squared	0.438	0.282	0.178	0.509	0.407	0.274	0.283

Poisson specification; Robust standard errors clustered by applicant in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Empirical Appendix

Figure A1: Frequent Words with Significant Gender Disparities

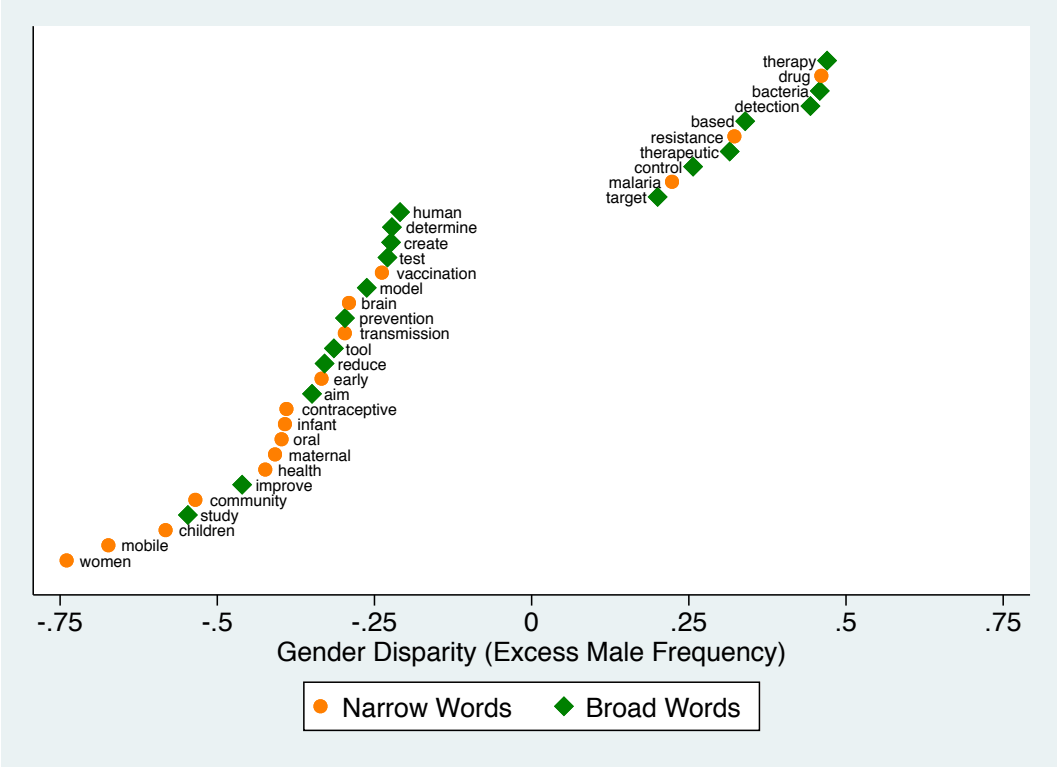

Figure A2: Frequent Words with Significant Score Disparities

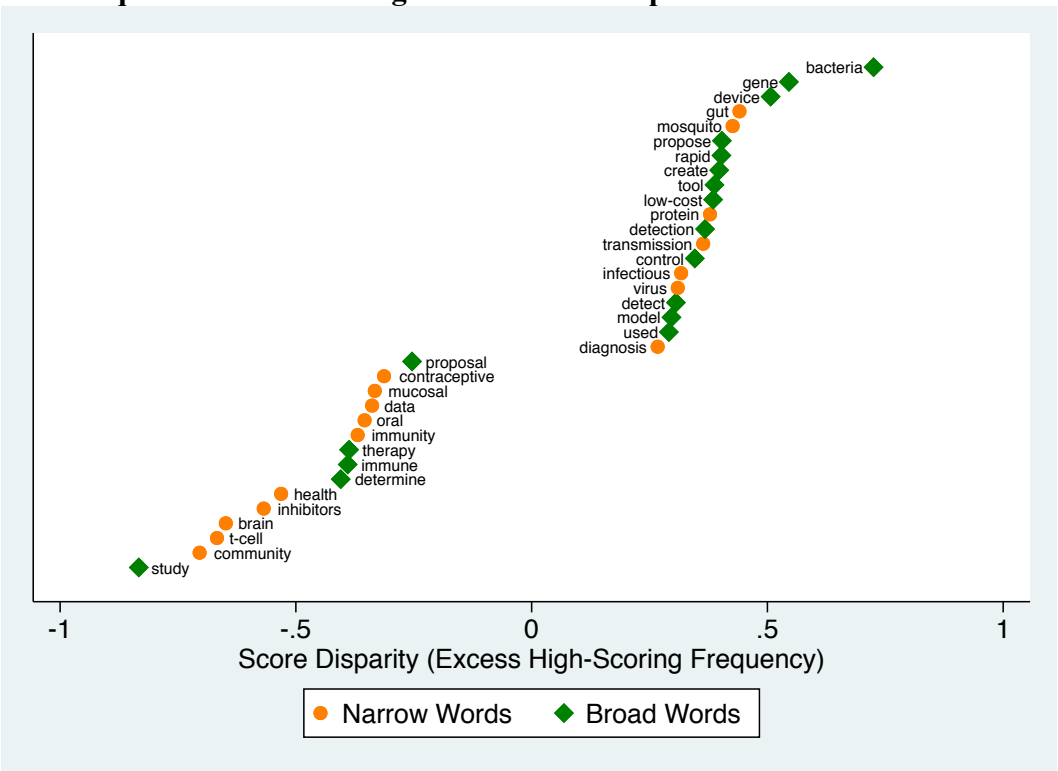

Figure A3: High-Scoring vs. Low-Scoring Word Use Rates for Broad and Narrow Words
Note: Words selected based on high use frequency.

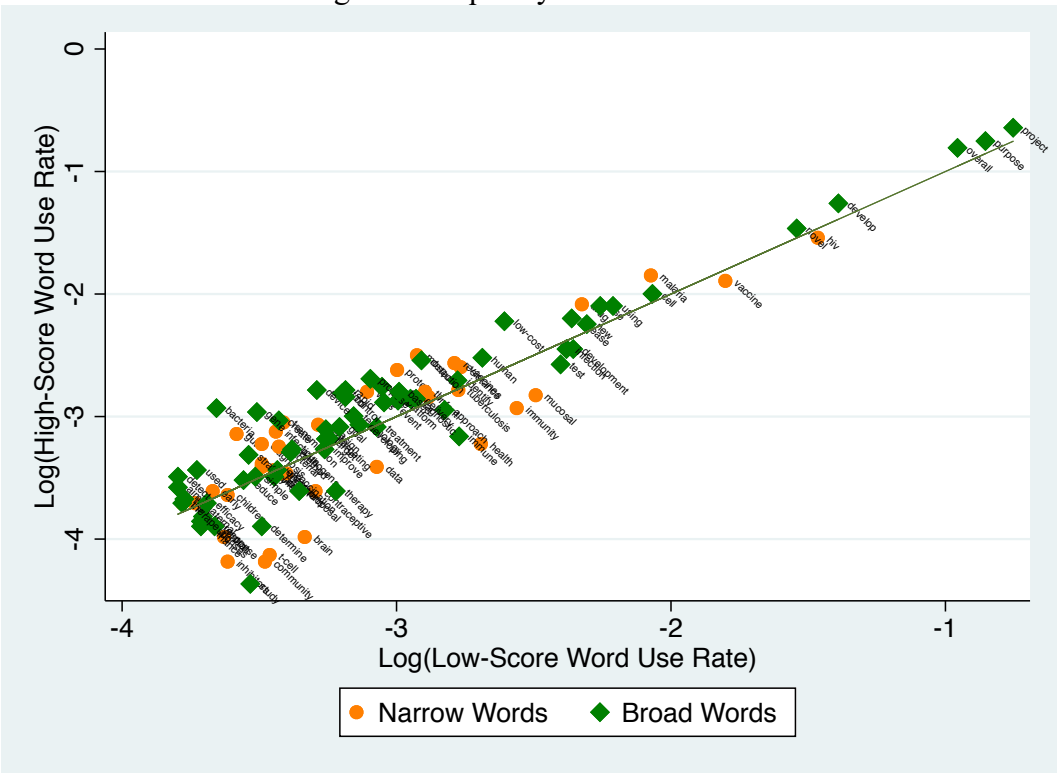

Figure A4: High-Scoring vs. Low-Scoring Word Use Rates for Selected Words
Note: Words selected based on high use rate and significant correlation with applicant gender.

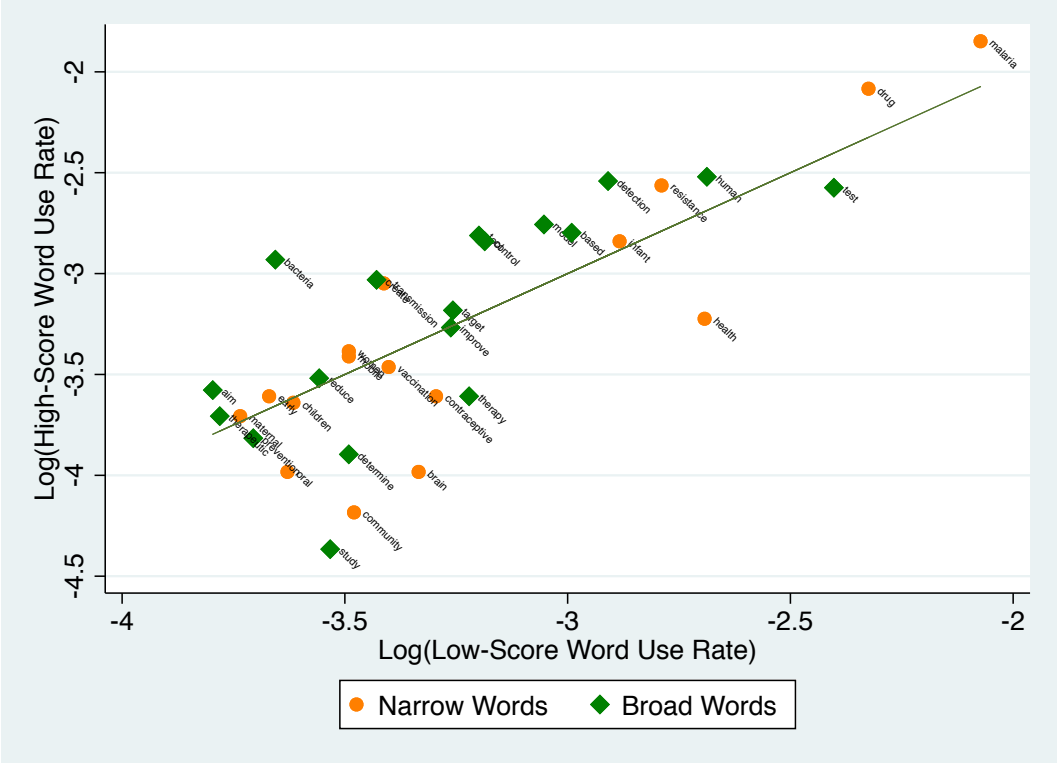

Figure A5: Male vs. Female Word Use Rates for High- and Low-Scoring Words

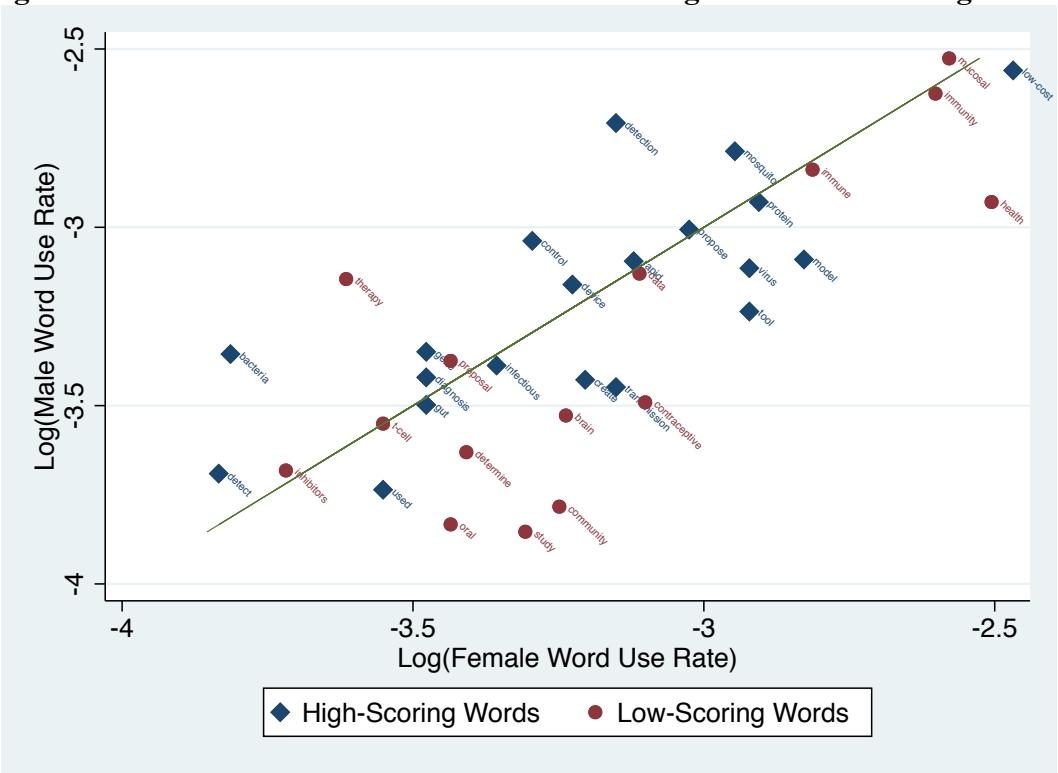

Figure A6: Comparing Word-Level Score Disparities for Male and Female Applicants

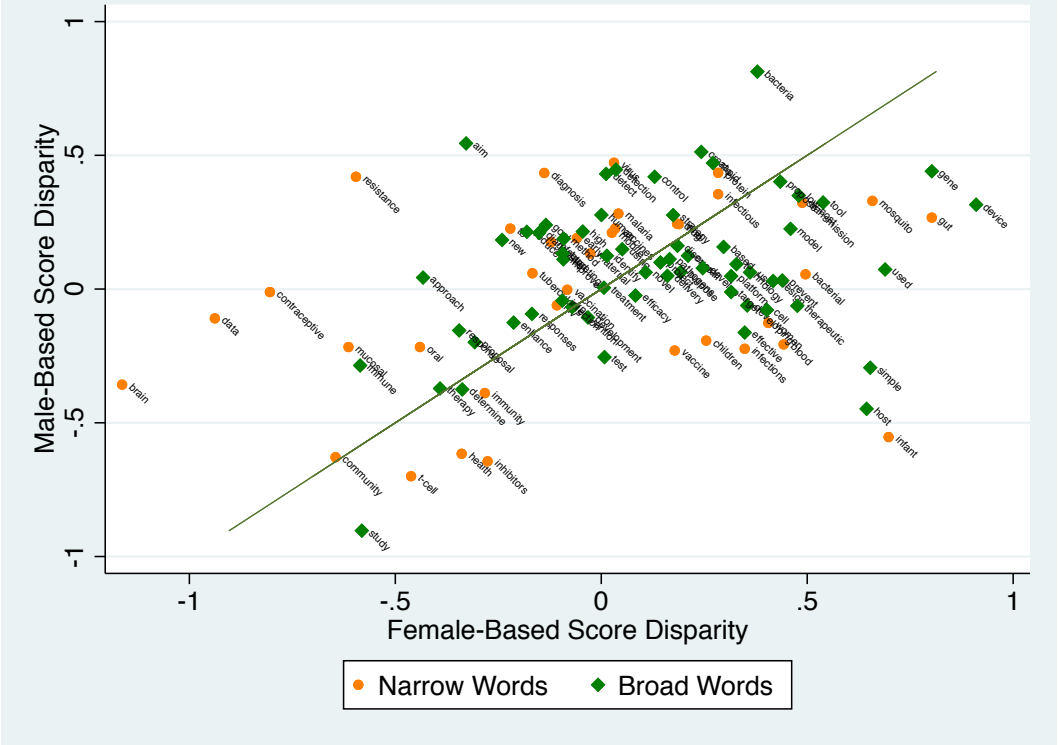

Figure A7: Male-Based Score Disparities and Gender-Based Use for Selected Words

Note: the y-axis tracks words that score well (or poorly) when used by male applicants

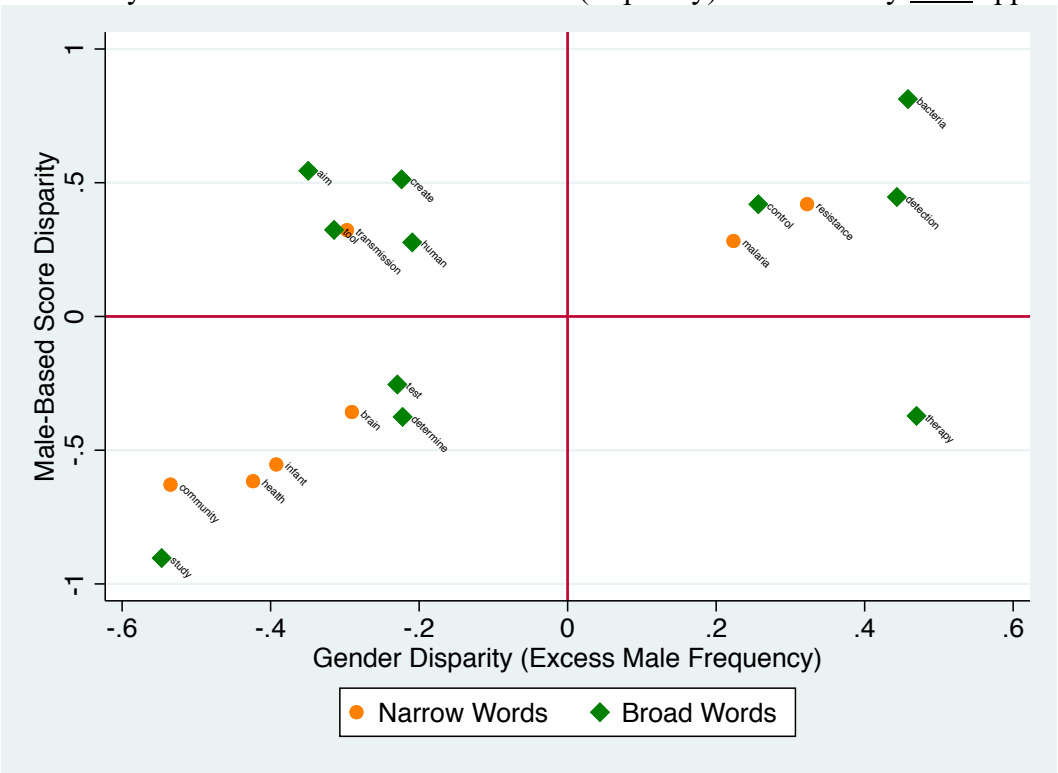

Figure A8: Female-Based Score Disparities and Gender-Based Use for Selected Words

Note: the y-axis tracks words that score well (or poorly) when used by female applicants

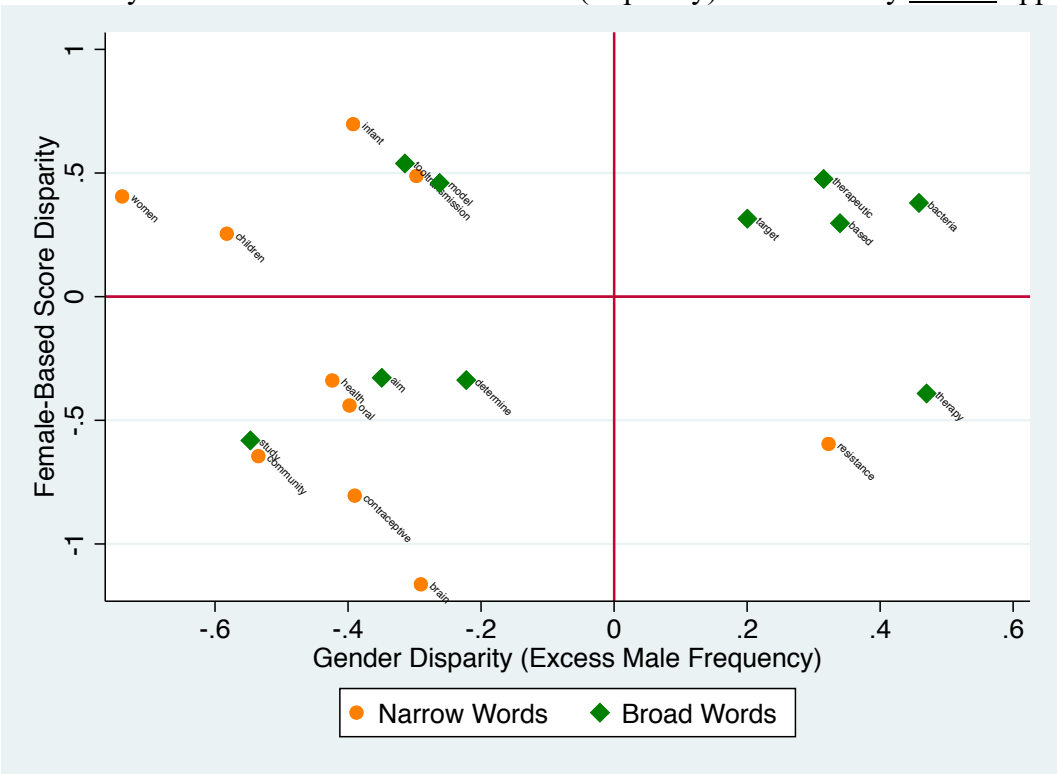

Table A1: Selected High- and Low-Scoring Words by Topic

This table lists common words with significant associations with reviewer scores within each of the topic areas in our sample. The top three high-scoring and low-scoring words are selected within each topic based ona combination of topic-specific score disparity and within-topic word frequency.

Topic Area	# of Proposals	High-Scoring Words			Low-Scoring Words		
Overall	6794	bacteria	engineering	device	health	fetal	study
HIV	1169	latent	eliminate	latently	microbicide	immune	targets
Discovery Core	1152	polio	sensor	devices	acid	reduce	nucleic
Malaria	906	sensors	blocking	acoustic	biomarkers	strategy	inhibitors
Reproductive & Neonatal Health	898	setting	device	pleasure	biomarkers	brain	fetal
Tuberculosis	608	funciton	urine	detect	cell	resistant	molecular
Diarrhea	502	dysfunction	bacteria	synthetic	rotavirus	asd	oral
Other	494	latrine	waste	new	wearable	disease	improve
Miscellaneous Diseases	429	snails	innovative	low-cost	blood	protective	onchocerciasis
Agriculture & Nutrition	383	plants	nematodes	block	vitamin	immune	feedback
Pneumonia	253	drug	mobile	resistance	children	infection	mucosal

Table A2: Relative Word Frequency Analysis: Score Disparity vs. Female Disparity

This table analyzes the top 1000 most frequent words used in the titles and descriptions of our sample of proposals. It focuses on the relationship between "score disparity," or the rate at which a given word appears disproportionately in high-scoring proposals relative to low-scoring, and "female_disparity," or the rate at which the word appears disproportionately in female-submitted proposals relative to those from male applicants.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	DV = Word-Level Score Disparity						
VARIABLES	Sample: All Words	Sample: All Words	Sample: All Words	Sample: Broad Words	Sample: Narrow Words	Sample: High-Scoring Words	Sample: Low-Scoring Words
Word-Level Female Disparity	-0.161***	-0.142***	-0.099**	-0.086	-0.104*	-0.059**	-0.017
	(0.035)	(0.039)	(0.040)	(0.063)	(0.053)	(0.029)	(0.041)
Round Controls	N	Y	Y	Y	Y	Y	Y
Topic Area Controls	N	N	Y	Y	Y	Y	Y
Observations	996	996	996	498	498	498	498
R-squared	0.022	0.126	0.148	0.163	0.195	0.260	0.159

OLS Specification; Robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Table A3: Applicant Characteristics and Proposal Text Grade Level

.

	(1)	(2)	(3)	(4)
	DV = Proposal Text Gradel Level			
VARIABLES	Flesch-Kincaid	Gunning Fog	SMOG	Average
Female Applicant	0.008	-0.050	-0.024	-0.022
	(0.061)	(0.086)	(0.046)	(0.059)
Log(Total Word Count)	3.452***	3.442***	4.239***	3.711***
	(0.189)	(0.269)	(0.160)	(0.189)
Log(Frequent Word Count)	1.045***	1.351***	1.050***	1.149***
	(0.142)	(0.188)	(0.098)	(0.131)
Noun Share	0.090	-0.191	-0.150	-0.084
	(0.364)	(0.519)	(0.269)	(0.351)
Adjective Share	-1.887***	-3.152***	-0.837**	-1.959***
	(0.512)	(0.668)	(0.326)	(0.448)
Verb Share	-2.389***	-2.033**	-1.041***	-1.821***
	(0.562)	(0.792)	(0.388)	(0.529)
Log(Applicant Career Length)	-0.000	-0.003	-0.022	-0.008
	(0.039)	(0.055)	(0.030)	(0.038)
Share of Top-Journal Pubs	0.030	0.256	0.053	0.113
	(0.150)	(0.205)	(0.107)	(0.142)
Round FEs	Y	Y	Y	Y
Topic Area FEs	Y	Y	Y	Y
Applicant Publication Characteristics	Y	Y	Y	Y
Additional Text-Based Controls	Y	Y	Y	Y
Observations	6,794	6,794	6,794	6,794
R-squared	0.599	0.464	0.514	0.539

OLS Specification; Robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1