4. L. Y. Yeung, J. L. Ash, E. D. Young, *J. Geophys. Res.* **119**, 10 (2014).
5. D. A. Stolper *et al.*, *Science* **344**, 1500–1503 (2014).
6. H. P. Affek, *Am. J. Sci.* **313**, 309–325 (2013).
7. B. H. Passey, G. A. Henkes, *Earth Planet. Sci. Lett.* **351-352**, 223–236 (2012).
8. D. A. Stolper *et al.*, *Geochim. Cosmochim. Acta* **126**, 169–191 (2014).
9. W. Guo, J. L. Mosenfelder, W. A. Goddard III, J. M. Eiler, *Geochim. Cosmochim. Acta* **73**, 7203–7225 (2009).
10. J. Tang, M. Dietzel, A. Fernandez, A. K. Tripati, B. E. Rosenheim, *Geochim. Cosmochim. Acta* **134**, 120–136 (2014).
11. H. P. Affek, S. Zaarur, *Geochim. Cosmochim. Acta* **143**, 319–330 (2014).
12. S. Ono *et al.*, *Anal. Chem.* **86**, 6487–6494 (2014).
13. R. D. Guy, M. L. Fogel, J. A. Berry, *Plant Physiol.* **101**, 37–47 (1993).
14. C. L. R. Stevens, D. Schultz, C. Van Baalen, P. L. Parker, *Plant Physiol.* **56**, 126–129 (1975).
15. Y. Helman, E. Barkan, D. Eisenstadt, B. Luz, A. Kaplan, *Plant Physiol.* **138**, 2292–2298 (2005).
16. H. C. Urey, L. J. Grieff, *J. Am. Chem. Soc.* **57**, 321–327 (1935).
17. Materials and methods are available as supplementary materials on *Science* Online.
18. W. Hillier, T. Wydrzynski, *Coord. Chem. Rev.* **252**, 306–317 (2008).
19. L. Rapatskiy *et al.*, *J. Am. Chem. Soc.* **134**, 16619–16634 (2012).
20. A. M. Angeles-Boza *et al.*, *Chem. Sci.* **5**, 1141 (2014).
21. A. M. Angeles-Boza, J. P. Roth, *Inorg. Chem.* **51**, 4722–4729 (2012).
22. B. Luz, E. Barkan, M. L. Bender, M. H. Thiemens, K. A. Boering, *Nature* **400**, 547–550 (1999).
23. A. Angert, S. Rachmilevitch, E. Barkan, B. Luz, *Global Biogeochem. Cycles* **17**, 1030 (2003).
24. M. Knox, P. D. Quay, D. Wilbur, *J. Geophys. Res.* **97** (C12), 20335–20343 (1992).
25. B. Luz, E. Barkan, *Geochim. Cosmochim. Acta* **69**, 1099–1110 (2005).
26. M. H. Cheah *et al.*, *Anal. Chem.* **86**, 5171–5178 (2014).
27. K. E. Tempest, S. Emerson, *Mar. Chem.* **153**, 39–47 (2013).
28. R. S. Thurston, K. W. Mandernack, W. C. Shanks III, *Chem. Geol.* **269**, 252–261 (2010).
29. L. W. Juranek, P. D. Quay, *Annu. Rev. Mar. Sci.* **5**, 503–524 (2013).
30. L. Y. Yeung, E. D. Young, E. A. Schauble, *J. Geophys. Res.* **117**, D18306 (2012).
31. B. M. Hoffman, D. Lukoyanov, D. R. Dean, L. C. Seefeldt, *Acc. Chem. Res.* **46**, 587–595 (2013).
32. B. Kok, B. Forbush, M. McGloin, *Photochem. Photobiol.* **11**, 457–475 (1970).
33. T. Noguchi, *Phil. Trans. R. Soc. B.* **363**, 1189–1195 (2008).
34. N. Cox *et al.*, *Science* **345**, 804–808 (2014).

ACKNOWLEDGMENTS

We thank H. Hu and N. Levin for performing oxygen triple-isotope analyses of the terrarium water at Johns Hopkins University, and E. Schauble for helpful discussions during the course of this work. This research was supported in part by the National Science Foundation (EAR-1049655 and DGE-1144087), the National Aeronautics and Space Administration Cosmochemistry program, and the Deep Carbon Observatory. The data and model parameters used in this study are available in the supplementary materials (tables S1 to S3).

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/348/6233/431/suppl/DC1
Materials and Methods
Supplementary Text
Figs. S1 to S5
Tables S1 to S3
References (*35*–*50*)

7 January 2015; accepted 13 March 2015
10.1126/science.aaa6284

RESEARCH FUNDING

Big names or big ideas: Do peer-review panels select the best science proposals?

Danielle Li[1]*† and Leila Agha[2,3]*†

This paper examines the success of peer-review panels in predicting the future quality of proposed research. We construct new data to track publication, citation, and patenting outcomes associated with more than 130,000 research project (R01) grants funded by the U.S. National Institutes of Health from 1980 to 2008. We find that better peer-review scores are consistently associated with better research outcomes and that this relationship persists even when we include detailed controls for an investigator's publication history, grant history, institutional affiliations, career stage, and degree types. A one–standard deviation worse peer-review score among awarded grants is associated with 15% fewer citations, 7% fewer publications, 19% fewer high-impact publications, and 14% fewer follow-on patents.

In 2014, the combined budgets of the U.S. National Institutes of Health (NIH), the U.S. National Science Foundation, and the European Research Council totaled almost $40 billion. The majority of these funds were allocated to external researchers whose applications were vetted by committees of expert reviewers. But as funding has become more competitive and application award probabilities have fallen, some observers have posited that "the system now favors those who can guarantee results rather than those with potentially path-breaking ideas that, by definition, cannot promise success" (*1*). Despite its importance for guiding research investments, there have been few attempts to assess the efficacy of peer review.

Peer-review committees are unique in their ability to assess research proposals based on deep expertise but may be undermined by biases, insufficient effort, dysfunctional committee dynamics, or limited subject knowledge (*2*, *3*). Disagreement about what constitutes important research may introduce randomness into the process (*4*). Existing research in this area has focused on understanding whether there is a correlation between good peer-review scores and successful research outcomes and yields mixed results (*5*–*7*). Yet raw correlations do not reveal whether reviewers are generating insight about the scientific merit of proposals. For example, if applicants from elite institutions generally produce more highly cited research, then a system that rewarded institutional rankings without even reading applications may appear effective at identifying promising research.

In this paper, we investigate whether peer review generates new insights about the scientific quality of grant applications. We call this ability peer review's "value-added." The value-added of NIH peer review is conceptually distinct from the value of NIH funding itself. For example, even if reviewers did a poor job of identifying the best applications, receiving a grant may still improve a researcher's productivity by allowing her to maintain a laboratory and support students. Whereas previous work has studied the impact of receiving NIH funds on the productivity of awardees (*8*, *9*), our paper asks whether NIH selects the most promising projects to support. Because NIH cannot possibly fund every application it receives, the ability to distinguish potential among applications is important for its success.

We say that peer review has high value-added if differences in grants' scores are predictive of differences in their subsequent research output, after controlling for previous accomplishments of the applicants. This may be the case if reviewers generate additional insights about an application's potential, but peer review may also have zero or even negative value-added if reviewers are biased, mistaken, or focused on different goals (*10*).

Because research outcomes are often skewed, with many low-quality or incremental contributions and relatively few ground-breaking discoveries (*2*, *11*), we assess the value-added of peer review for identifying research that is highly influential or shows commercial promise. We also test the effectiveness of peer review in screening out applications that result in unsuccessful research (see the supplementary materials for full details on data and methods).

NIH is the world's largest funder of biomedical research (*12*). With an annual budget of approximately $30 billion, it supports more than 300,000 research personnel at more than 2500 institutions (*12*, *13*). A funding application is assigned by topic to one of approximately 200 peer-review committees (known as study sections).

Our main explanatory variable is the "percentile score," ranging from 0 to 100, which reflects an application's ranking among all other applications reviewed by a study section in a given fiscal year; lower scores correspond to higher-quality applications. In general, applications are funded in order of their percentile score until the budget of their assigned NIH institute is exhausted. The average score in our sample is 14.2, with a standard deviation (SD) of 10.2; only about 1% of funded grants in our sample had a score worse than 50. Funding has become more competitive in recent years; only 14% of applications were funded in 2013.

[1]Harvard University, Cambridge, MA 02138, USA. [2]Boston University, Boston, MA 02215, USA. [3]National Bureau of Economic Research, Cambridge, MA 02138, USA.
***Corresponding author. E-mail: dli@hbs.edu (D.L.); lagha@bu.edu (L.A.)** †Both authors contributed equally to this work.

Our sample consists of 137,215 research project (R01) grants funded from 1980 through 2008. R01s are project-based renewable grants that are NIH's primary grant mechanism, accounting for about half of its extramural grant spending. Of the grants in our sample, 56% are for new projects; the remaining successfully competed for renewal. We focus on funded grants because funding is likely to have direct effect on research productivity, making it difficult to infer the success of peer review by comparing funded and unfunded grants. Because our sample grants have the same funding status, we can attribute any remaining relationship between scores and outcomes to peer review, rather than funding. Because grants are almost always funded in order of their score, there is relatively little scope for selection on unobservables to introduce bias.

Our primary outcome variables are (i) the total number of publications that acknowledge grant support within 5 years of grant approval (via PubMed); (ii) the total number of citations that those publications receive through 2013 (via Web of Science); and (iii) patents that either directly cite NIH grant support or cite publications acknowledging grant support [via the *U.S. Patent and Trademark Office* (USPTO)]. These publication, citation, and patent outcomes are designed to reflect NIH's stated goals of rewarding research with high scientific and technical merit.

We also measure applicant-level characteristics: an investigator's publication and grant history, educational background, and institutional affiliation. We match investigators with publications using their full last name and their first and middle initials (*14*). We track the number of articles an applicant published in the 5 years before submitting her application, as well as the impact of those publications as measured by the citations they have received by the time the application is evaluated. We identify "high-impact" publications as being among the top 0.1%, 1%, and 5% most cited, compared with articles published in the same year. To more precisely assess the quality of an applicant's ideas, we repeat this exercise for articles in which the applicant is a first or last author only. Our regression results include separate controls for each type of publication: any authorship position, and first or last author publications. By counting only citations received up to the date of grant review, we ensure that our measures contain only information available to reviewers at the time they evaluate the application.

Table 1. Do peer-review scores predict future citations and publications? Each reported figure is the coefficient on scores from a single Poisson regression of grant outcomes on NIH peer-review scores; standard errors are reported in parentheses. The actual sample size used per regression depends on the number of nonzero observations for the dependent variable. The independent variable is the percentile score. "Future citations" refers to the total number of citations, to 2013, that accrue to all publications that acknowledge funding from a given grant. "Future publications" refers to the total number of such publications. Subject-year controls refer to study section by fiscal year fixed effects, as well as NIH institute fixed effects. PI publication history includes controls for number of past publications, number of past citations, and number of past hit publications. PI career characteristics include controls for degrees and experience (time since highest degree). PI grant history controls for number of previous R01s and non-R01 NIH funding. PI institution and demographics control for the rank of the PI's institution, as well as gender and some ethnicity controls. Standard errors are clustered at the study section year level. *, statistical significance at the 10% level; **, 5% level; ***, 1% level.

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
			Dependent variable: Future citations			
Independent variable: NIH percentile score	−0.0203***	−0.0215***	−0.0162***	−0.0164***	−0.0162***	−0.0158***
	(0.0006)	(0.0008)	(0.0007)	(0.0007)	(0.0007)	(0.0007)
N	137,215	136,076	136,076	128,547	128,547	128,547
			Dependent variable: Future publications			
Independent variable: NIH percentile score	−0.0155***	−0.0091***	−0.0076***	−0.0077***	−0.0076***	−0.0075***
	(0.0003)	(0.0003)	(0.0003)	(0.0003)	(0.0003)	(0.0003)
N	137,215	136,111	136,111	128,580	128,580	128,580
Controls						
Subject-year		X	X	X	X	X
PI publication history			X	X	X	X
PI career characteristics				X	X	X
PI grant history					X	X
PI institution/demographics						X

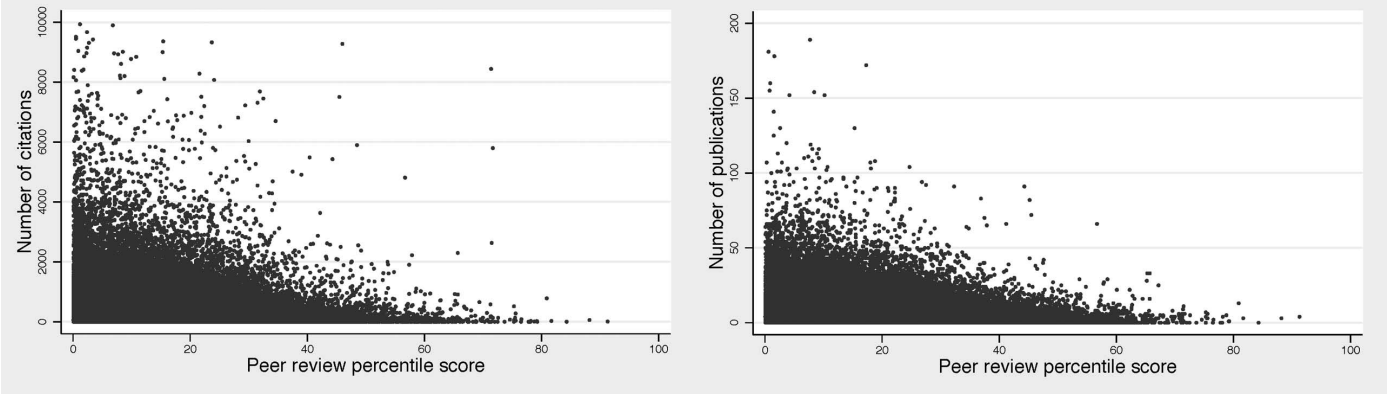


Fig. 1. Scatterplot of percentile scores and grant outcomes. The left panel plots the relationship between percentile scores and citations associated with a grant. Each dot represents a single grant. The right panel does the same for total publications. Extreme outliers with more than 10,000 citations or 200 publications are not displayed here.

We observe whether an applicant has an M.D., Ph.D., or both, as well as the year in which she received her final doctoral degree. We are missing degree and experience information for 0.45% and 7.16% of our sample, respectively; we include two separate indicators for missing these data. We measure whether this applicant previously received an R01 grant and whether the applicant has received any previous NIH funding. Using the name of the principal investigator (PI), we employ a probabilistic algorithm developed by Kerr to determine applicant gender and ethnicity (Hispanic or Asian) (*15*, *16*, *17*). We rank applicants' institutions by the number of NIH grants received over our study period and measure whether each applicant is from a top 5-, 10-, 20-, or 50-ranked institution. We are unable to determine the institutional affiliation of 14% of investigators; we include an indicator variable for missing institution information.

Consistent with previous work, there is substantial dispersion in research output even among the relatively well-developed projects that receive NIH R01 funding (*5*). The median grant in our sample received 116 citations to publications acknowledging the grant; the mean is more than twice as high, 291, with an SD of 574. This variation in citations underscores the potential gains from being able to accurately screen grant applications on the basis of their research potential.

Our first set of results describes peer review's value-added for identifying research likely to result in many publications or citations. Table 1 reports results from Poisson regressions of future outcomes on peer-review scores, with different controls for an applicant's previous performance. The supplementary materials describe many additional robustness checks.

Model 1 of Table 1 reports, without any control variables, the percentage change in the number of citations and publications associated with a grant, given a one point increase in its percentile score. We find that NIH evaluations are statistically related to grant quality; our estimated coefficients indicate that a one percentile point worse peer-review score is associated with 1.6% fewer publications and 2% fewer citations. To consider the magnitude of these findings more clearly, we will describe our results by reporting how predicted outcomes change with a 1-SD (10.17 point) worse percentile score; in Model 1, a 1-SD worse score is associated with a 14.6% decrease in grant-supported research publications and a 18.6% decrease in citations to those publications ($P < 0.001$). This calculation is based on the overall SD in percentile score among funded grants, unconditional on PI characteristics (*18*). Figure 1 illustrates the raw relationship between scores and citations and publications in a scatterplot; the plot suggests a negative sloping relationship (recall that higher percentile scores indicate less favorably reviewed research).

There are potential concerns with interpreting the unadjusted relationship between scores and outcomes as a measure of peer review's value. Some grants may be expected to produce more citations or publications and thus appear higher quality, independent of their true quality. Older grants have more time to produce publications that in turn have more time to accrue citations. A publication with 100 citations may be average in one field but exceptional in another.

Model 2 of Table 1 addresses these concerns by including detailed fixed effects for study sections by year cells and NIH institutes. The inclusion of these fixed effects means that our estimates are based only on comparisons of scores and outcomes for grants evaluated in both the same fiscal year (to account for cohort effects) and in the same study section (to account for field effects). We also include NIH institute-level fixed effects to control for differences in citation and publication rates by fields, as defined by a grant's area of medical application. Controlling for cohort and field effects does not attenuate our main finding. For a 1-SD (10.17 point) worse score, we expect an 8.8% decrease in publications and a 19.6% decrease in citations (both $P < 0.001$). This suggests that scores for grants evaluated by the same study

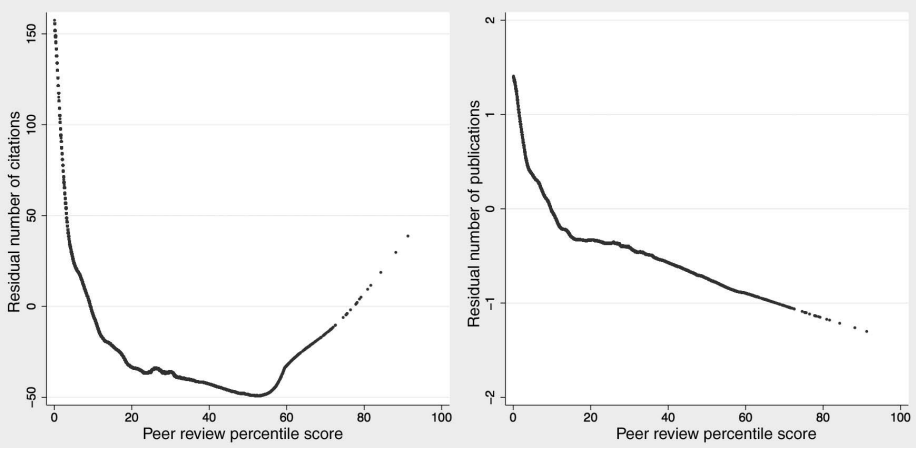


Fig. 2. Smoothed scatterplots of percentile scores and residual grant outcomes. These figures display smoothed scatterplots of the nonparametric relationship between unexplained variation in grant outcomes and percentile score, after accounting for differences in field of research, year, and applicant qualifications. The left panel plots the relationship between percentile scores and residual citations associated with a grant. The right panel does the same for residual publications.

Table 2. Do peer-review scores predict hit publications and follow-on patents? Each reported figure is the coefficient on scores from a single Poisson regression of grant outcomes on NIH peer-review scores; standard errors are in parentheses. High-impact publication is given by the count of publications acknowledging the grant that receive more citations than all but 0.1%, 1%, or 5% of publications from the same year. Direct patents are those that acknowledge funding from a grant; indirect patents are those that cite publications that acknowledge funding from a grant. We control for the same variables as described in Model 6 of Table 1.

	Dependent variable: High-impact publications			Dependent variable: Patents	
	Top 0.1% (1)	Top 1% (2)	Top 5% (3)	Direct (4)	Indirect (5)
Independent variable: NIH percentile score	−0.0246***	−0.0209***	−0.0172***	−0.0153***	−0.0149***
	(0.0025)	(0.0014)	(0.0009)	(0.0015)	(0.0022)
N	88,795	118,245	125,021	122,850	92,893
Controls					
Subject-year	X	X	X	X	X
PI publication history	X	X	X	X	X
PI career characteristics	X	X	X	X	X
PI grant history	X	X	X	X	X
PI institution/demographics	X	X	X	X	X

section in the same year and assigned to the same NIH institute are better than randomly allocated.

We may observe this pattern, however, if reviewers simply give good scores to applicants with strong research credentials, and applicants with strong credentials generally tend to produce better research. Model 3 of Table 1 adds controls describing a PI's publication history in order to ask whether study section scores contain information about the quality of an application that could not be predicted by simply examining a PI's curriculum vita.

Specifically, we include the following additional control variables: (i) the number of articles published in the past 5 years; (ii) the total number of citations those articles have received up to the year of grant review; (iii) three variables describing the number of top 0.1%, 1%, and 5% articles that the PI has published in the previous 5 years; and (iv) alternate versions of these variables constructed only with the subset of publications for which the applicant was a first or last author. Controlling for publication history attenuates but does not eliminate the relationship: a 1-SD (10.17 point) worse score is associated with a 7.4% decrease in future publications and a 15.2% decrease in future citations (both $P < 0.001$).

The association between better scores and better outcomes could also be explained by the Matthew effect, a sociological phenomenon wherein credit and citations accrue to established investigators simply because they are established, regardless of the true quality of their work (*19*, *20*). Were this the case, more connected applicants may receive better scores and more citations regardless of the true quality of their work. Our approach may thus credit peer review for responding to prestige, rather than the underlying quality of an applicant's ideas.

Model 4 controls for the PI's experience by adding indicators for whether the applicant has an M.D., Ph.D., or both, as well as a series of indicator variables capturing how many years have elapsed since receiving her terminal degree. If reviewers were simply giving better scores to candidates with more experience or skill writing grant proposals and publishing papers, then we would expect scores to become less predictive of future research output once we control for M.D./Ph.D. status and time since degree. Instead, our estimated relationship between peer-review scores and outcomes remains unchanged.

Model 5 considers the possibility that peer reviewers may be rewarding an applicant's grant proposal writing skills rather than the underlying quality of her work. Specifically, we include variables controlling for whether the PI received NIH funding in the past, including four indicators for having previously received one R01 grant, two or more R01 grants, one NIH grant other than an R01, and two or more other NIH grants. To the extent that reviewers may be responding to an applicant's experience and skill with proposal writing, we would expect the inclusion of these variables reflecting previous NIH funding to attenuate our estimates of value-added. We find, however, that including these variables does not substantively affect our findings.

Finally, in Model 6, we also control for institutional quality, gender, and ethnicity, to capture other potentially unobserved aspects of prestige, connectedness, or access to resources that may influence review scores and subsequent research productivity. Our estimates again remain stable: comparing applicants with statistically identical backgrounds, the grant with a 1-SD worse score is predicted to have 7.3% fewer future publications and 14.8% fewer future citations (both $P < 0.001$).

Across Models 3 to 6, the estimated relationship between peer-review scores and outcomes remains remarkably stable, even as we add more covariates that describe an applicant's past accomplishments, prestige, proposal-writing skill, and professional connections. Although these variables certainly cannot capture every potential source of omitted variables bias, the stability of our results suggests that political connections and prestige are not a primary driver of peer review's value-added.

Next, we explore whether reviewers' expertise enables them to identify the strongest applications or to more efficiently screen out weaker applications. We use a local linear regression model to nonparametrically identify the relationship between peer-review score and research quality. This flexibility will allow the predictive power of peer-review scores to differ at each point along the score spectrum. We implement this approach in two steps, which are described in detail in the supplementary materials. First, we construct the residuals from a linear regression of research outcomes on all of the explanatory variables in Model 6, excluding the study section percentile score itself. These residuals represent the portions of grants' citations or publications that cannot be explained by applicants' previous qualifications or by application year or subject area (as detailed above). We then produce a locally weighted, linearly smoothed scatterplot relating peer-review scores to these residual citations and publications.

Figure 2 shows that peer reviewers add value by identifying the strongest research proposals. For all percentile scores less than 50 (the vast majority of awarded grants), worse scores are associated with lower expected residual citations and publications. The relationship is particularly steep at very low percentile scores, suggesting that study sections are particularly effective at discriminating quality among very well-reviewed applications.

One notable exception occurs for very poorly scored applications—those with percentile scores

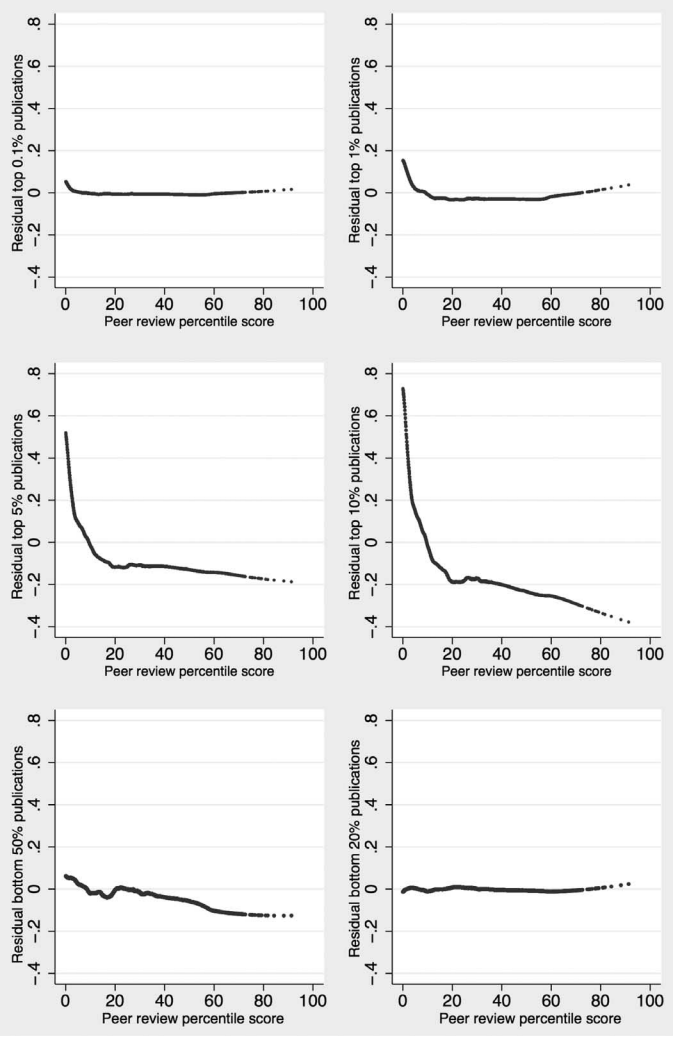


Fig. 3. Smoothed scatterplots of percentile scores and residual high- and low-citation publications. These figures display smoothed scatterplots of the nonparametric relationship between unexplained variation in grant outcomes and percentile score, after accounting for variation in field of research, year, and applicant qualifications. Each panel reports results on the number of residual publications in the indicated performance bin.

over 50—that were nonetheless funded. In this range, worse review scores are associated with higher citation counts. These applications constitute about 1% of funded applications and are highly unlikely to have met the standard award threshold but were instead funded "out of order." We find higher average quality for this set of selected grants, suggesting that when program officers make rare exceptions to peer-review decisions, they are identifying a small fraction of applications that end up performing better than their initial scores would suggest.

Our final analysis explores whether peer reviewers' value-added comes from being able to identify transformative science, science with considerable applied potential, or from being able to screen out very low-quality research. We define a "hit" publication as among the top 0.1%, 1%, or 5% most cited publications in its cohort, using all citations a publication receives through 2013. To explore whether reviewers have value-added in terms of identifying research with practical applications, we track the number of patents that explicitly acknowledge NIH funding. The majority of NIH grants, however, do not directly result in patents. Thus, we also count the number of patents that cite research funded by a grant (indirect patenting).We construct this variable by linking grants to publications using grant acknowledgment data and then applying a fuzzy matching algorithm that identifies publications cited by USPTO patents (*21*). This allows us to identify patents that cite publications that in turn acknowledge a grant. Importantly, this process (described further in the supplementary materials), allows us to identify patents regardless of whether those patents are assigned to the same investigator funded by the NIH grant. Indeed, most often these patents are held by private firms (*22*).

As reported in Table 2, peer-review scores have value-added identifying hit publications and research with commercial potential. A 1-SD (10.17 points) worse score is associated with a 22.1%, 19.1%, and 16.0% reduction in the number of top 0.1%, 1%, and 5% publications, respectively. These estimates are larger in magnitude than our estimates of value-added for overall citations, especially as we consider the very best publications. The large value-added for predicting tail outcomes suggests that peer reviewers are more likely to reward projects with the potential for a very high-impact publication and have considerable ability to discriminate among strong applications.

A 1-SD worse percentile score predicts a 14% decrease in both direct and indirect patenting. Because of the heterogeneous and potentially long lags between grants and patents, many grants in our sample may one day prove to be commercially relevant even if they currently have no linked patents. This time-series truncation makes it more difficult to identify value-added with respect to commercialization of research and means that our estimates are likely downward biased.

Finally, we investigate the nonparametric relationship between percentile scores and publication outcomes, testing which score ranges are associated with the highest numbers of "hit" publications, ranking at the top of the citation distribution, and which score ranges are associated with the highest numbers of "miss" publications, ranking near the bottom of the distribution. We follow the same local linear regression smoothing procedure outlined above and described in more detail in the supplementary materials.

Figure 3 shows that low percentile scores are consistently associated with higher residual numbers of hit publications, variation unexplained by the applicant's background or field of study. The relationship between scores and residual research outcomes is steepest among the most well-reviewed applications. For example, funded grants with percentile scores near 0 are predicted to produce 0.05 more publications in the top 0.1% of the citation distribution, compared with applications scored near the 10th percentile (holding constant applicant qualifications and field).

Although this may seem like a modest increase, there is a small number of such hit publications, so a 0.05 increase in their number corresponds to a doubling of the mean number of top 0.1% publications arising from a grant. This relationship between scores and hit publications becomes weaker among applications with less competitive scores; a 10-percentile point difference in scores in the range of 20 to 30 would predict only a 0.0004 difference in the number of top 0.1% publications. This finding runs counter to the hypothesis that, in light of shrinking budgets and lower application success rates, peer reviewers fail to reward those risky projects that are most likely to be highly influential in their field (*1*, *2*).

We don't find evidence that the peer-review system adds value beyond previous publications and qualifications in terms of screening out low-citation papers. Better percentile scores are associated with slightly more publications in the bottom 50% of the citation distribution. There is no discernible relationship between residual publications in the bottom 20% and peer-review score among the funded grants in our sample, suggesting that while these less influential anticipated publications are not rewarded by the peer-review system, they are also not specifically penalized.

Our findings demonstrate that peer review generates information about the quality of applications that may not be available otherwise. This does not mean that the current NIH review system would necessarily outperform other allocation mechanisms that do not rely on expert peer evaluations. Our analysis focuses on the relationship between scores and outcomes among funded grants; for that reason, we cannot directly assess whether the NIH systematically rejects high-potential applications. Our results, however, suggest that this is unlikely to be the case, because we observe a positive relationship between better scores and higher-impact research among the set of funded applications.

Although our findings show that NIH grants are not awarded purely for previous work or elite affiliations and that reviewers contribute valuable insights about the quality of applications, mistakes and biases may still detract from the quality of funding decisions. We have not included an accounting of the costs of peer review, most notably the time investment of the reviewers. These bibliometric outcomes may not perfectly capture NIH objectives or be the only measures relevant for evaluating social welfare; ideally, we would like to link grants with health and survival outcomes, but constructing those measures is difficult and beyond the scope of this paper. Future research may focus on whether the composition of peer-review committees is important to determining their success, including evaluator seniority and the breadth and depth of committee expertise.

REFERENCES AND NOTES

1. B. Alberts, M. W. Kirschner, S. Tilghman, H. Varmus, *Proc. Natl. Acad. Sci. U.S.A.* **111**, 5773–5777 (2014).
2. D. F. Horrobin, *JAMA* **263**, 1438–1441 (1990).
3. J. M. Campanario, *Sci. Commun.* **19**, 181–211 (1998).
4. S. Cole, J. R. Cole, G. A. Simon, *Science* **214**, 881–886 (1981).
5. J. Berg, Productivity metrics and peer review scores: NIGMS feedback loop blog (2011); https://loop.nigms.nih.gov/2011/06/productivity-metrics-and-peer-review-scores/.
6. J. Berg, Productivity metrics and peer review scores, continued: NIGMS feedback loop blog (2011); https://loop.nigms.nih.gov/2011/06/productivity-metrics-and-peer-review-scores-continued.
7. N. Danthi, C. O. Wu, P. Shi, M. Lauer, *Circ. Res.* **114**, 600–606 (2014).
8. B. A. Jacob, L. Lefgren, *Res. Policy* **40**, 864–874 (2011).
9. B. A. Jacob, L. Lefgren, *J. Public Econ.* **95**, 1168–1177 (2011).
10. J. H. Tanne, *BMJ* **319**, 336 (1999).
11. K. Arrow, The rate and direction of inventive activity: Economic and social factors (National Bureau of Economic Research, Cambridge, MA, 1962), pp. 609–626.
12. About NIH Web site (2014); http://www.nih.gov/about/.
13. E. R. Dorsey *et al.*, *JAMA* **303**, 137–143 (2010).
14. There is no further disambiguation, but we show that our results do not change when we restrict to investigators with rare names. See table S5 of the supplementary materials.
15. W. R. Kerr, The ethnic composition of US inventors, Working Paper 08-006, Harvard Business School (2008); http://www.people.hbs.edu/wkerr/Kerr%20WP08_EthMatch.pdf.
16. W. R. Kerr, *Rev. Econ. Stat.* **90**, 518 (2008).
17. Due to the limitations of the name-based matching algorithm, we cannot reliably distinguish African-American investigators.
18. For example, to calculate the 14.6% figure, we take the exponential of our estimated coefficient times the SD in scores, minus 1: exp(−0.0155 × 10.17) − 1.
19. R. K. Merton, *Science* **159**, 56–63 (1968).
20. P. Azoulay, T. Stuart, Y. Wang, *Manage. Sci.* **60**, 92–109 (2013).
21. P. Azoulay, J. S. G. Zivin, B. N. Sampat, The diffusion of scientific knowledge across time and space: Evidence from professional transitions for the superstars of medicine, Tech. Rep., National Bureau of Economic Research (NBER, Cambridge, MA, 2011).
22. P. Azoulay, J. Graff-Zivin, D. Li, B. Sampat, Public R&D investments and private sector patenting: Evidence from NIH funding rules, NBER working paper 20889 (2013); http://irps.ucsd.edu/assets/001/506033.pdf.

ACKNOWLEDGMENTS

We are grateful to P. Azoulay, M. Lauer, Z. Obermeyer, and B. Sampat for helpful comments, suggestions, and assistance with data. We also acknowledge assistance from M.-C. Chen, P. Kennedy, A. Manning, and especially R. Nakamura from the NIH Center for Scientific Review. This paper makes use of restricted-access data available from the National Institutes of Health. Those wishing to replicate its results may apply for access following the procedures outlined in the NIH Data Access Policy document available at http://report.nih.gov/pdf/DataAccessPolicy.pdf.

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/348/6233/434/suppl/DC1
Materials and Methods
Fig. S1
Tables S1 to S8
References (*23*–*29*)

8 October 2014; accepted 18 March 2015
10.1126/science.aaa0185

Science

Big names or big ideas: Do peer-review panels select the best science proposals?

Danielle Li and Leila Agha

Science **348** (6233), 434-438.
DOI: 10.1126/science.aaa0185

Proof that peer review picks promising proposals

A key issue in the economics of science is finding effective mechanisms for innovation. A concern about research grants and other research and development subsidies is that the public sector may make poor decisions about which p rojects to fund. Despite its importance, especially for the advancement of basic and early-stage science, there is currently no large-scale empirical evidence on how successfully governments select research investments. Li and Agha analyze m ore than 130,000 grants funded by the U.S. National Institutes of Health during 1980–2008 and find clear benefits of peer evaluations, particularly for distinguishing high-impact potential among the most competitive applications.

Science, this issue p. 434

www.sciencemag.org/content/348/6233/434/suppl/DC1

Supplementary Materials for

Big names or big ideas: Do peer-review panels select the best science proposals?

Danielle Li* and Leila Agha*

*Corresponding author. E-mail: dli@hbs.edu (D.L.); lagha@bu.edu (L.A.)

Published 24 April 2015, *Science* **348**, 434 (2015)
DOI: 10.1126/science.aaa0185

This PDF file includes:

A NIH Peer Review Background

The National Institutes of Health (NIH) is the primary organization within the United States government with responsibilities for health-related research. The NIH is the single largest funder of biomedical research, with an annual budget of approximately $30 billion dollars. More than 80% of the total budget supports extramural research through competitive grants that are awarded to universities, medical schools, and other research institutions, primarily in the United States.

In this section, we describe NIH funding policies as they were during the majority of our sample period.[1] Requests for proposals identify priority areas, but investigators are also free to submit applications on unsolicited topics under the extramural research program. Prior to the peer review reform in 2006, applications were 25 pages long; they were shortened to 12 pages. All applications are assigned to a review committee comprised of scientific peers, generally known as a study section. Peer reviewers are asked to ignore budgetary issues, limiting their attention to scientific and technical merit on the basis of five criteria: (1) significance; (2) approach; (3) innovation; (4) investigator skill; and (5) conducive environment. These are still the criteria used today.

Within a study section, applications are assigned to 1 to 3 reviewers who read the application and provide initial scores. Based on these scores, approximately 40 to 50% of applications are "triaged," or rejected without further discussion. The remaining applications are discussed at the meeting of the full study section. At this meeting, initial reviewers are typically asked to discuss an application and present their scores. This is followed by an open discussion by all reviewers and a brief period for everyone to revise their initial scoring based on the group deliberations before anonymously submitting their final scores.

The overall priority score for the proposal is based on the average across all study section members. Prior to 2006, these scores were from 1.0 (best) to 5.0 (worst) in increments of 0.1. [2] Scores are then normalized within review groups through the assignment of percentile scores to facilitate funding decisions. This paper uses this final percentile score as our measure of the study section's evaluation of a grant application.

[1] In 2006, NIH updated its peer review process to require, among other things, shorter applications and a less granular scoring system. Since our data covers the period from 1980 to 2008, the great majority of the applications we observe pre-date these changes. The fundamental features of the NIH peer review process, including the number and type of reviewers and a numeric scoring system were not changed by these 2006 revisions; as a result we do not expect that the efficacy of the peer review process was substantially altered by these revisions, but we note differences between the old and new systems.

[2] After the 2006 reform, these scores were from 1 (best) to 9 (worst) in increments of 1.

After scores have been assigned, applications are then sorted to the individual institutes responsible for funding the application. (A single study section will often contain applications assigned to multiple institutes for final funding.) Individual institutes make funding decisions by rank ordering applications according to their percentile scores and funding the lowest (best) scores first then proceeding until funding is exhausted. There are a few exceptions to this funding process: special consideration is given to early investigators who have not previously received R01 funding; and the institute may occasionally choose to fund "out of order" from the percentile score ranking in response to an appeals process or exceptional circumstances not recognized by the initial committee.

B Data and Variable Construction

B.1 Sample Details

We make use of data from the following sources: (a) administrative data on NIH-funded grants from the NIH IMPAC database; (b) life science publication data from the National Library of Medicine's PubMed database and Thomson Reuter's Web of Science; and (c) USPTO data on patent applications. Using these datasets we construct outcome variables measuring the research products associated with a given grant and control variables for the prior performance and characteristics of the applicant.

Our analysis includes all new and competing renewal R01 grants that received a peer review score and were successfully funded by the NIH between FY1980 and FY2008. Table S1 describes the characteristics of these grants in more detail. We restrict our sample to only awarded grants to avoid potential bias from unfunded projects whose research outcomes may be diminished by the lack of funding award.

One may be concerned that examining only funded grants will lead to selection bias in our sample. Selection bias occurs when funded and unfunded grants that look similar on observables nonetheless differ on unobservables in a way that impacts their future performance. At the NIH, this type of selection is unlikely to be a large concern because, in most cases, the only variables that determine whether a grant application is funded is its Institute, its funding cycle cohort, and its percentile score. These are all variables that we observe and include in our regression.

We need only be concerned about selection bias due to cases when a grant's Institute, year, and score are not the sole determinants of its funding status. This is in practice quite rare.[3] If the NIH makes exceptions to the payline and funds applications that it thinks have particularly high potential despite weak peer review scores, then this type of classical sample selection will tend to attenuate our estimated relationship between scores and grant outcomes (*23*). We explore this potential bias in Section D.4 by using an "identification at infinity" approach (see (*24*)) and restricting the sample to only very highly scored grants with very high funding probabilities. The logic underlying this specification is that selection into being funded or not should not impact our estimates for grants whose scores are so low that they are nearly always funded.

B.2 Measuring the Performance of Funded Grants

Our outcome measures are constructed as follows. We first match publications to grants using the unique grant identifier provided in the IMPAC data and PubMed grant acknowledgment data. In IMPAC, a grant is identified by a full grant number comprised of different parts, for example "1R01CA000001-1" This tells us the type of grant (new "1" or competing renewal "2"), the grant mechanism (in this case, "R01"), the funding Institute ("CA" is the National Cancer Institute), the project number ("000001") and the year of support (the first year, "1"). We match a grant to the publications that acknowledge this funding source and are published within 5 years after application approval. Publications citing multiple NIH R01 grants are matched to both applications.

Our grant application data include only years in which a grant was subject to a competitive evaluation. The goal behind our matching strategy is to find all publications arising out of proposed research evaluated by a study section at a given point in time. As such, if a publication cites grant support from a year in which it was automatically renewed, we attribute that publication to the nearest previous competitive year. For example, suppose that a grant is approved in 1999 and lasts through 2002. If a publication cites support from that grant for the year 2000, we would match that publication to the 1999 application, the year in which the 2002 funding was actually adjudicated.

There are also cases in which publications cite partial grant numbers, usually the Institute and project number but not the support year. In this case, we attribute this publication to the nearest competitive cycle prior to its publication, but not more than 5 years prior to the last automatic renewal. For example, consider a grant initially approved in 1999 and renewed in 2002.

[3] We do not have data on which specific grants are funded out of order but the NIH currently funds fewer than $<3\%$ of cases out of order.

If a 2005 publication cites the sample project number but does not specify the funding year, we would attribute it to the 2002 review year. If a 2010 publication cites that project number, we would attribute it to neither cycle because 2010 is more than 5 years after the last renewal year (in this case, 2002).

After matching grants to publications using grant acknowledgements, we track the citations to those publications using data from Web of Science. We construct several additional outcome measures with this citation data. The first outcome is the total number of forward citations accruing to the linked set of publications, through 2013. In addition, we construct a set of variables tracking the number of "hit" publications associated to a grant. A publication is deemed a hit if it is cited in the 99.9th, 99th, or 95th percentile of life science articles published in the same year; we construct one variable for each threshold. Because we use grant acknowledgement data from PubMed and citation data from Web of Science, we construct a crosswalk between these two datasets. This means that our outcome variables, with the exception of our simple publication count, include only publications indexed in both PubMed and Web of Science.

We test the robustness of our main findings to two alternative rules for linking grant applications to publication outcomes. First, since there is the potential for long lags in completing and publishing research, we report alternative results in Section D.2 that attribute up to 10 years of publications acknowledging grant support to the initial application. Second, since some researchers may fail to acknowledge their funding source in published papers, we also test an alternative matching strategy. In Section D.2 below, we match each application to *all* publications authored by the PI within 5 years following grant approval and report analogous regression results for this outcome. These alternative publication measures are linked to citation counts using the same methodology outlined above.

The final set of outcome variables measure the patent output of grants. We match grants to patents in two ways. The first is to examine the set of patents that directly acknowledge financial support from the NIH. Beginning in 1981, the Bayh-Dole Act required that patents report all sources of federal support. The second is to identify patents that build on the knowledge produced by a particular grant. This outcome captures the broader commercial relevance of a grant application, which is a characteristic that reviewers and policymakers may care about. For each grant, we construct the number of patents that cite publications acknowledging grant support. This approach to linking grants and patents does not require a restrictive assumption that the commercial applications of a research project match the original domain of the research; it will capture the full

range of downstream applications, including potentially unanticipated spillovers to other areas of research.

Determining whether patents cite publications is more difficult than tracing patent citations: while the cited patents are unique seven-digit numbers, cited publications are free-form text (/textit25). Moreover, the USPTO does not require that applicants submit references to literature in a standard format. For example, Harold Varmus's 1988 *Science* article "Retroviruses" is cited in 29 distinct patents, but in numerous different formats, including: "Varmus. "Retroviruses" Science 240:1427-1435 (1988)" (in patent 6794141) and "Varmus et al., 1988, Science 240:1427-1439" (in patent 6805882). As this example illustrates, there can be errors in author lists and page numbers. Even more problematic, in some cases certain fields (e.g. author name) are included, in others they are not. Journal names may be abbreviated in some patents, but not in others.

To address these difficulties, we applied a matching algorithm developed by Azoulay, Graff Zivin and Sampat that compares each of several PubMed fields—first author, page numbers, volume, and the beginning of the title, publication year, or journal name—to all references in all biomedical and chemical patents issued by the USPTO since 1976 (*21*). The sample of biomedical and chemical patents were identified by using the patent class-field concordance developed by the National Bureau of Economic Research (*26*). We considered a dyad to be a match if four of the fields from PubMed were listed in a USPTO reference. Overall, the algorithm returned 1,058,893 distinct PMIDs cited in distinct 322,385 patents. Azoulay, Graff Zivin and Sampat discuss the performance of this algorithm against manual searching, and tradeoffs involved in calibrating the algorithm (*21*).

B.3 Measuring the Prior Qualifications of Applicants

We also construct variables describing an applicant's qualifications at the time of grant review. We match an applicant's name to his or her publication history using PubMed data. For each applicant, we construct the total number of publications she has published in the previous 5 years; and three variables describing the number of hit publications, those that fall in the 99.9th, 99th, or 95th percentile of citations among life science articles published in the same year. We repeat this exercise for publications in which the PI is a first or last author.

We calculate citation percentiles using citations received up to the time of grant review, as opposed to counting all citations received through 2013. In our main analysis, we restrict to matches based on first and middle initials as well as full last names. This still leaves room for ambiguity; Table S6 shows that our results are robust to restricting to applicants with rarer names. For all

our publication-based variables, we restrict to original research articlees; this excludes, for instance, reviews and letters to the editor.

In addition to publication history, we also construct an applicant's grant history using the NIH IMPAC database. We track whether an applicant has received an R01 in the past or whether she had received any non-R01 funding from the NIH. An applicant is counted as having received prior NIH funding only if she is listed as the primary recipient; this would not include applicants who have received fellowships through another investigator.

Our analysis also includes controls for a PI's degrees—an MD, Ph.D., or both—as well as for her institutional affiliation. The latter is available for the majority of the grant recipients in our sample (92%). We group these institutions into tiers, ranking them based on the total number of NIH grants the institution has received over our study period. We then assign the PI's institution to one of 5 categories: top 5, top 10, top 25, top 50, and lower than 50.

B.4 Summary Statistics

Summary statistics are reported in Table S1 and Figure 1. Our sample includes 137,215 awarded R01 applications, funded by 21 NIH institutes, in 617 study sections over 29 years. 56% of applications are new applications, and the remainder of the sample are competing renewals. The average peer review score is 14.22 percentile points, and the average award amount is $1,220,731.

Figure 1 shows that, consistent with prior work, there is substantial dispersion in research output even among the relatively well-developed projects that receive NIH R01 funding (*5*). This is most readily apparent when examining the distribution of citations associated with a grant: the median grant in our sample has received 116 citations to publications acknowledging the funding award, but the mean is more than twice as high, 291, with a standard deviation of 574. This variation in citations underscores the potential gains from being able to accurately screen grant applications on the basis of their research potential.

As reported in Table S1, each grant in our sample is acknowledged by an average of 7.4 publications within 5 years; the median number of publications is 5. As of 2013, the average number of citations accruing to grant-acknowledging publications is 291, and the median number of citations is 116. While the median number of publications in the top 0.1%, 1%, and 5% of the distribution (ranked by citations) is 0, the averages suggest that highly successful publications are not an uncommon outcome of NIH funded work; for example, each grant is associated with

0.6 publications in the top 0.1% on average. Similarly, the median number of patents directly or indirectly acknowledging the NIH grant is 0, but the averages are 0.10 and 4.75, respectively.

Past performance variables indicate that PIs typically have a long and well-cited publication history. The median number of first or last authored publications within the past 5 years is 12, and median citations to those publications at the time of grant application is 62. 79% of applicants have a Ph.D. degree (including those who also have MD's), and 70% of applicants have received R01 funding in the past. 24% of applicants are associated with a top 10 research institution, when ranked by the number of NIH grants awarded to that institution.

C Econometric Models

We use poisson regression models to test whether peer review scores predict outcomes of grant-funded research, after conditioning on the principal investigator's past research productivity, institutional affiliation, and field of research. We call this relationship peer review's *value added* following the economics of education literature that measures the value of a teacher's contribution to improving student achievement, after taking into account that student's prior performance and demographic characteristics (*27-29*). The poisson regressions model the count structure of our outcome variables, including numbers of citations, publications, and patents associated with each funded grant. Standard errors are clustered at the study session-year level. The poisson regressions take the following form:

$$E(Outcome_i|Score_i, \mathbf{X_i}) = exp(\beta Score_i + \mathbf{X_i}\gamma) \quad (1)$$

$Outcome_i$ is a measure of the grant application i's eventual success, for example, the number of publications acknowledging the grant or the number of citations to those publications. The key independent variable of interest is $Score_i$ which measures the percentile score assigned by the peer review study section. $\mathbf{X_i}$ is a vector of control variables which vary by regression specification as outlined below. Each successive model adds further control variables.

1. Model 1 (first reported in Table 1 Column 1) is a parsimonious regression that simply tests whether better peer review scores predict better research outcomes, without the inclusion of any control variables. We successively add control variables to the model, testing whether the

observed relationship continues after conditioning on additional characteristics of the research field and investigator.

2. Model 2 (Table 1 Column 2) includes 3858 fixed effects for study section by year interactions as well as fixed effects for the 27 different funding institutes within the NIH. The inclusion of these fixed effects accounts for differences in publication and citation counts across academic disciplines.

3. Model 3 (Table 1 Column 3) adds control variables for the principal investigator's past research productivity over the past 5 years: the number of publications, the number of citations these publications have received at the time of grant application, and whether any of the publications fall into the top 0.1%, 1%, or 5% of papers published that year at the time of application. These variables are included separately for first and last authored publications and for all publications.

4. Model 4 (Table 1 Column 4) adds controls for the principal investigator's education: degree type and experience. These include whether the PI has an M.D., Ph.D., or both, and the number of years since completing his terminal degree.

5. Model 5 (Table 1 Column 5) adds controls for whether the principal investigator has received NIH funding in the past, including four indicators for having received 1 R01 grant, 2 or more R01 grants, 1 NIH grant other than an R01, and 2 or more other NIH grants.

6. Model 6 (Table 1 Column 6) is our final regression specification and it adds indicators for the PI's institutional quality, gender, and ethnicity. Our measures of institutional quality include indicators for whether his institution is in the top 5, top 10, top 20, or top 50 institutions ranked by total number of awarded grants in our sample. Demographic characteristics are probabilistically matched by investigator name and include indicators for female, asian, hispanic, or missing ethnicity.

Table 1 shows the coefficients on peer review scores for each of these models. Table S2 displays coefficients on all of the covariates from each poisson regression with citations as the outcome measure; Table S3 does the same with publications as the dependent variable.

The second type of analysis presented in the paper assesses the nonparametric relationship between grant outcomes and peer review scores, after accounting for differences in publication

records that are due to the applicant's field and academic qualifications. For this analysis, we begin by running linear fixed effects regressions which take the following form:

$$Outcome_i = \mathbf{X_i c} + e_i \tag{2}$$

The control variables included in $\mathbf{X_i}$ match the Model 6 specification described above, but the regression excludes percentile score as an explanatory variable. The control variables include study section-year fixed effects, institute fixed effects, gender and ethnicity indicators, past research productivity measures, and education, employment measures. We calculate residuals from this regression as the difference between each grant's realized outcome and the predicted outcome. We then use a locally weighted scatterplot smoothing technique to display the relationship between the variation in grant outcomes unexplained by the linear regression model and the peer review percentile scores.

In order to smooth the residual values, we run a series of linear regressions of the residualized outcomes on percentile score. We run a separate regression centered around each observation in our dataset; these regressions include up to 20% of the sample that is nearest to this observation's percentile score. Weights are applied so that points with more distant observations receive less weight according to the tricube formula.[4] The predictions of this set of local, weighted regressions are the smoothed values of the residual outcome variable, which we then use in a scatterplot graphed against percentile scores. Results of this analysis are reported in Figures 3 and 4 and discussed in the main body of the paper.

D Robustness checks and additional results

In this section, we first interpret in more detail the coefficients on the control variables in our main set of regression results. Then, we describe a series of robustness checks investigating whether our findings are sensitive to the control variables included in the analysis, the sample selected, or the choice of regression model.

[4] In particular, for a regression centered around the observation $(Score_i, Outcome_i)$ the tricube weight takes the following form:

$$w_j = \left\{1 - \left(\frac{|Score_j - Score_i|}{1.0001 \max(Score_{i_+} - Score_i, Score_i - Score_{i_-})}\right)^3\right\}^3 \tag{3}$$

where $Score_{i_-}$ is the smallest value of $Score$ included in the regression, and $Score_{i_+}$ is the largest.

D.1 Further discussion of main results

Results on the primary coefficient of interest, peer review percentile score, are discussed in the main body of the paper. Table S2 Column 6 displays results describing the relationship between PI's past accomplishments and citations to grant-acknowledging publications, after controlling for the percentile score. In particular, we see that competing renewals receive 49% more citations, which may be reflective of more citations accruing to more mature research agendas ($P<0.001$). Applicants with M.D. degrees amass more citations to their resulting publications ($P<0.001$), which may be a function of the types of journals they publish in, citation norms, and number of papers published in those fields. Applicants from research institutions with the most awarded NIH grants garner more citations ($P<0.001$), as do applicants who have previously received R01 grants ($P<0.001$). Lastly, researchers early in their career tend to produce more highly cited work than more mature researchers ($P<0.001$).

Table S3 Column 6 displays similar results, but with the outcome variable being total publications which acknowledge the grant, published within 5 years of the grant award. Patterns are similar to those noted above for total citations, with the exception that applicants from top-ranked research institutions do not seem to publish more prolifically (conditional on their peer review scores and past publication and citation performance), but rather receive more citations per publication.

D.2 Robustness to alternative outcome measures

In this section, we test three alternative methods of linking grant applications to future publications. Results are reported in Table S4. These regressions match the Model 6 specification and include a full set of controls for study section by year, institute, publication history, degree, experience, previous grant receipt, demographics, and institutional affiliations.

1) Table S4 Column 1 replicates results from Table 1 column 6 in the main body of the paper. Recall that grants are linked to all publications acknowledging grant support, published within 5 years of grant approval.

2) Table S4 Column 2 reports results with an alternate outcome measure that links grants to all publications acknowledging the grant and published within 10 years. This outcome extends the time frame during which we track grant publications. We find a very similar coefficient relating percentile score to future citation counts, just slightly lower at -0.0153 compared to -0.0158 in the original specification.

3) Table S4 Column 3 reports results that use an alternative procedure for linking grants to publications; rather than relying on grant acknowledgements, we instead attribute all publications authored by the principal investigator within 5 years to the grant itself. This broader match will eliminate any potential bias if investigators who are more diligent about appropriately acknowledging grant funding are also more likely to receive well-scored grants. However, the broader match will also lead us to attribute some publications unrelated to the grant proposal, and thus we expect the estimated coefficients to be attenuated towards zero since peer review committees were not evaluating the potential of these proposals. Consistent with this hypothesis, we find a one point improvement in peer review score is associated with a 1% more citations, which is smaller than the 1.6% in the main specification but still significant at the 0.1% level.

4) Table S4 Column 4 repeats the broader name-based matching of applications to future publications, but extends the match to include all publications authored by the PI within 10 years following grant approval. These results are very similar to the Column 3 findings; a 1 point improvement in peer review score is associated with 1% increase in citations.

D.3 Robustness to inclusion of additional control variables

In this section, we probe the robustness of our results to the inclusion of additional control variables. Table S5 reports results. Except as otherwise noted, these regressions begin with the Model 6 controls as the baseline covariates and augment them with additional measures as described below.

1) Table S5 Column 1 replicates results from Table 1 column 6 in the main body of the paper. Recall that this regression controls for study section by year, institute, publication history, degree, experience, previous grant receipt, demographics, and institutional affiliations.

2) Table S5 Column 2 tests whether the results reported in Table 1 are sensitive to including a control for the total dollars awarded to each grant. Since peer reviewers are instructed not to consider the funding amounts requested in the grant proposal, we do not expect funding levels to confound our regression findings. NIH funding allocations are made by funding the lowest-scored project and proceeding to higher-scored projects until the budget for that research subject is exhausted. However, one may be concerned that the applications with the best peer review scores also receive the most funding, and that it is through this funding channel that these grants have more productive research outcomes. We find that controlling for funding allocations does not substantially attenuate our main result. For the total citation outcome, the coefficient on percentile score remains almost

unchanged at -0.016 ($P<0.001$) in our preferred Model 6 specification and in the specification which adds a control for total dollars funded.

3) We investigate the robustness of our results to the inclusion of fixed effects for PI institutional affiliations, rather than controlling for institutional quality with tiered quality bins (as in Model 6). Results with institution fixed effects, along with all of the other Model 6 control variables, are reported in Table S5 Column 3. Again, the results are consistent, with the coefficient on percentile score going to -0.015 ($P<0.001$) after the inclusion of institution fixed effects for the total citation outcome (from -0.016 in Model 6).

4) We reports results from regressions where we include very detailed controls for past publications and relax the functional form assumptions about the relationship between an investigator's publication history and his future research outcomes. In particular, we consider the following covariates: number of publications in each decile of the citation distribution; the number of highly successful publications in the top 5%, 1%, and 0.1% of the citation distribution; total number of past publications; and the total number of past citations. For each of these variables, we include a 5-year and 10-year publication history, and separate measures for whether it was a first or last authored publication, or any other authorship position. Further, we cut these variables into 25 separate quantile bins and include create indicators for each bin. All of these indicators are included in the final regression results, greatly relaxing the log-linear functional form assumption for the relationship between, for example, number of past publications, and future research outcomes. These regressions continue to include controls for the investigator's degree type, institutional affiliation tier, career age, past grant recipiency, grant renewal status, NIH funding institution, and study section by year fixed effects.

Our findings are robust to the inclusion of these very rich controls for the investigator's publication history as reported in Table S5 Column 4; a 1 point improvement in peer review percentile score is associated with a 1.3 percent increase in citations ($P<0.001$) in the new specification with rich publication controls, compared to 1.6 percent increase in our Model 6 specification. Similarly, a 1 point improvement in score is associated with a 0.7% increase in the number of publications ($P<0.001$) in the new specification, compared to a 0.8% increase in our Model 4 specification. Moving from the 25th to the 75th percentile of the score distribution (i.e. from a score of 21 to a score 6), is associated with a 20% increase in citations, and 10% fewer publications.

5) Table S5 Column 5 excludes all variables related to a PI's publication history, but continues to include controls for study section by year, institute, degrees, experience, quality of institutional

affiliation, gender, ethnicity, and previous grant receipt. We run this model to test our hypothesis that the explanatory variables driving most of the omitted variables bias in the bivariate Model 1 specification are those related to the PI's publication history. Indeed, including our full vector of other controls (excluding publication history) leads to a very similar estimated relationship between scores and outcomes of -0.0196 in Table S5 Column 5 compared to -0.0203 in the bivariate model reported in Table 1 Column 1. The models show remarkable coefficient stability after including publication variables as we add other indicators for applicant background, suggesting that omitted variables bias due to unobserved heterogeneity in applicant prestige is unlikely to play a strong role in explaining our regression results.

6) Table S5 Column 6 restricts the sample to only include PIs with rare names, defined as investigators with a first initial, middle initial, and last name combination that is unique among authors in PubMed. One limitation of the analyses reported in this paper is that we match applicant investigators to past publications using investigator names. Because this match may not be exact, we tend to over-count the publication history of investigators with common names, introducing measurement error in the independent variables. Results from the rare name sample are very similar to the original Model 6 specification: a one point improvement in score associated with 1.4% increase in citations ($P<0.001$)

7) Table S5 Column 7 presents our identification at infinity results to address concerns about potential selection on unobservables. We restrict the sample to only include applications that receive percentile scores of 15 or less; almost 93% of applications with scores less than 15 are funded. Using just this subsample, we find a stronger relationship between scores and outcomes: a one point improvement in score associated with 2.9% increase in citations ($P<0.001$). This stronger relationship may be in part due to the fact that any sample selection bias is likely to have attenuated our results; reducing the potential for selection should then increase our estimates of value added. It is also possible, as suggested by Figures 3 and 4 in the main body, that the true relationship between scores and outcomes is also stronger for this set of grants.

D.4 Alternative sample selection criteria

In this section, we test whether peer review committees can successfully discriminate application quality even amongst applicants who are less likely to have a long, proven academic track record. These specifications provide an even stronger test of peer review's ability to draw conclu-

sions about the potential impact of application's scientific proposal. In all cases, regressions match the Model 6 specification with a full set of control variables.

1-2) Previous work has suggested that the peer review scores are more predictive of research outcomes for renewal applications than for new grant applications due to the stronger signals of research quality available once the research agenda is further advanced (*6*). We investigate this hypothesis, separating our sample into new grant approvals and renewal applications, reported in Table S6 Columns 1 and 2, respectively. Using the Model 6 regression specification, we find that although peer review scores are somewhat stronger predictors of citations and publications for renewal applications, there remains a strong relationship between peer review scores and citations even for new applications. In particular, a one point improvement in percentile score is associated with 1.3% increase in citations for new grant applications ($P<0.001$), compared to a 1.8% increase in citations for renewal applications ($P<0.001$)

3-4) Next we investigate whether peer review committees can assess application quality amongst applicants who have never previously been a principal investigator on a NIH grant. Table S6 Columns 3 and 4 report results on the sample with and without prior NIH funding, respectively. The relationship between peer review scores and citation outcomes are very similar for these two samples, with a one point improvement in score associated with 1.43% more citations for applicants with no prior NIH grants, compared to 1.46% for applicants with prior grants.

5-6) Lastly, we split our sample according to the PI's experience. Table S6 Column 5 reports results for investigators who are within 10 years of completing their doctoral terminal degree, and column 6 excludes these junior investigators. Restricting to junior investigators should also limit the scope for omitted variables bias due to unobserved variation in applicant fame or prestige that is not captured by the included regressors, since junior investigators have not had as much time to develop professional reputations. We find that the relationship between peer review scores and citations is highly similar for less experienced investigators, suggesting further support for the hypothesis that peer review committees are successfully assessing application quality independent of signals of applicant prestige.

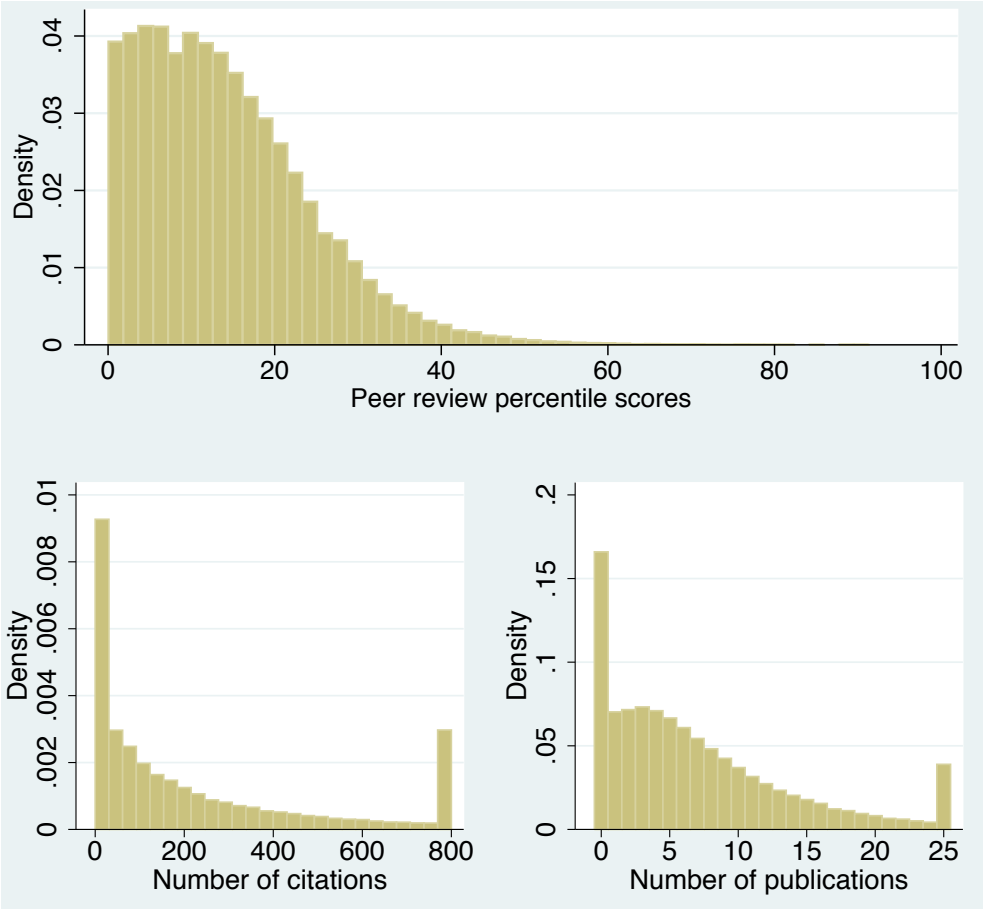
D.5 Linear regression model results

1) Table S7 replicates the findings from our main results in Table 1, using linear ordinary least squares regressions rather than poisson regressions. The included covariates in each regression parallel those included in Models 1-6 as described above. In column 1, before the inclusion of

any control variables, a 1 point improvement in percentile score is associated with 5.3 additional citations ($P<0.001$). This effect attenuates to 4.1 additional citations per 1 point score improvement in the Model 6 specification ($P<0.001$), which includes controls for publication history, institutional affiliation, gender, ethnicity, degrees earned, career age, past grant receipt, NIH funding institute, and study section by year fixed effects. Similarly, without control variables (Table S7 Column 1), each improvement in percentile score is associated with 0.11 more publications ($P<0.001$), which attenuates to 0.05 fewer publications in Model 6 (Table S7 Column 4, $P<0.001$).

2) Table S8 replicates the findings from Table 2 on the value added of peer review for identifying hit publications and research with commercial applications, using linear ordinary least squares regressions rather than poisson regressions. The included covariates in each regression match those in Model 6, as described above. We find that improved peer review scores are associated with more hit publications and follow-on patenting. In particular, a one standard deviation improvement in percentile score is associated with 0.008 more publications in the top 0.1%, 0.038 in the top 1%, and 0.133 in the top 5% ($P<.0001$), holding constant the investigator's publication history, institutional affiliation, gender, ethnicity, degrees earned, career age, past grant receipt, NIH funding institute, and study section by year fixed effects. Because these publication counts are overlapping and cumulative, it is not surprising that we find a greater coefficient on the top 5% outcome compared to the others. Turning to the patent outcomes, we find that a 1 standard deviation improvement in percentile scores is associated with 0.015 more patents acknowledging the grant directly and 0.78 more patents acknowledging the grant indirectly.

Figure 1: Distribution of Percentile Scores and Grant Outcomes



Notes: The top panel is the distribution of our independent variable, the percentile score received by a funded NIH R01 grant. A percentile score of X means that a grant received a better score than all but $X\%$ of grants evaluated by the same study section in the same year. The bottom panels plot the distribution of future research outcomes associated to a grant: the total number of publications as well as citations. For the purposes of these histograms, citations have been top-coded at 800 and publications have been top-coded at 25; the rest of the analysis does not impose any such top-coding.

Supplementary Table 1: Summary Statistics

	Mean	*Median*	*Standard Deviation*
Sample Coverage			
# Grants	137,215		
# Institutes/Centers	24		
# Study Sections	617		
# Years	29		
Grant Characteristics			
% New	56.43		
Peer review percentile score	14.22	12.60	10.17
Funding Amount	$1,220,731	$1,092,000	$889,956
Grant Outcome Variables			
# Citations to acknowledged publications	291.12	116	573.94
# Acknowledged publications	7.36	5	8.54
# Top 0.1% acknowledged publications	0.05	0	0.30
# Top 1% acknowledged publications	0.25	0	0.85
# Top 5% acknowledged publications	1.00	0	2.14
# Directly acknowledged patents	0.10	0	0.69
# Indirectly acknowledged patents	4.75	0	19.55
PI Characteristics and Prior Performance			
# Citations in past 5 years, first or last author	178	62	470
# Publications in past 5 years, first or last author	25	12	93
# Top 0.1% acknowledged publications in past 5 years, first or last author	1.77	0	4.57
# Top 1% acknowledged publications in past 5 years, first or last author	4.16	2	10.87
# Top 5% acknowledged publications in past 5 years, first or last author	7.46	4	20.38
# Citations in past 5 years	341	104	1260
# Publications in past 5 years	46	18	236
# Top 0.1% acknowledged publications in past 5 years	3.40	1	11.94
# Top 1% acknowledged publications in past 5 years	7.89	3	30.09
# Top 5% acknowledged publications in past 5 years	14.08	6	57.09
Years since doctorate	18.21	17	8.96
Has Ph.D	71.77		
Has MD	20.30		
Has MD/Ph.D.	7.16		
% Received prior R01 funding	70.42		
% Received prior funding, non R01	45.21		
Affiliated with Top 5 Research Institution	14.75		
Affiliated with Top 10 Research Institution	23.82		
Affiliated with Top 20 Research Institution	37.08		
Affiliated with Top 50 Research Institution	63.24		

Notes: The sample includes all NIH-funded R01 grants from 1980-2008. We restrict to new and competing renewal applications that received study section percentile scores. See Supporting Online Material for additional details about the definition of variables.

Supplementary Table 2: Do Peer Review Scores Predict Future Citations? All Covariates

	Dependent Variable: *Future Citations*					
	(1)	(2)	(3)	(4)	(5)	(6)
Independent Variable: *NIH Percentile Score*	-0.0203*** *(0.0006)*	-0.0215*** *(0.0008)*	-0.0162*** *(0.0007)*	-0.0164*** *(0.0007)*	-0.0162*** *(0.0007)*	-0.0158*** *(0.0007)*
Competing Renewal		0.5211*** *(0.0122)*	0.4502*** *(0.0118)*	0.5152*** *(0.0122)*	0.4822*** *(0.0125)*	0.4885*** *(0.0126)*
# of Citations, past 5 years, first or last author			-0.0002*** *(0.0001)*	-0.0002** *(0.0001)*	-0.0002** *(0.0001)*	-0.0001** *(0.0001)*
# of Publications, past 5 years, first or last author			-0.0011 *(0.0010)*	-0.0002 *(0.0010)*	-0.0005 *(0.0010)*	-0.0001 *(0.0010)*
# Top 0.1% acknowledged publications in past 5 years, first or last author			0.0301*** *(0.0041)*	0.0301*** *(0.0043)*	0.0293*** *(0.0043)*	0.0271*** *(0.0042)*
# Top 1% acknowledged publications in past 5 years, first or last author			0.0116** *(0.0057)*	0.0110* *(0.0058)*	0.0111* *(0.0058)*	0.0112** *(0.0057)*
# Top 5% acknowledged publications in past 5 years, first or last author			0.0076 *(0.0072)*	0.0061 *(0.0074)*	0.0060 *(0.0074)*	0.0066 *(0.0074)*
# of Citations, past 5 years			-0.0000 *(0.0001)*	-0.0000 *(0.0001)*	-0.0000 *(0.0001)*	-0.0000 *(0.0001)*
# of Publications, past 5 years			-0.0028*** *(0.0006)*	-0.0032*** *(0.0006)*	-0.0029*** *(0.0006)*	-0.0031*** *(0.0006)*
# Top 0.1% acknowledged publications in past 5 years			0.0011 *(0.0030)*	0.0019 *(0.0031)*	0.0017 *(0.0031)*	0.0027 *(0.0030)*
# Top 1% acknowledged publications in past 5 years			-0.0095** *(0.0043)*	-0.0101** *(0.0044)*	-0.0099** *(0.0044)*	-0.0095** *(0.0043)*
# Top 5% acknowledged publications in past 5 years			0.0293*** *(0.0055)*	0.0279*** *(0.0056)*	0.0277*** *(0.0056)*	0.0246*** *(0.0055)*
Has MD				0.2630*** *(0.0149)*	0.2727*** *(0.0151)*	0.2227*** *(0.0152)*
Has MD/Ph.D.				0.2555*** *(0.0189)*	0.2520*** *(0.0189)*	0.1976*** *(0.0192)*
Has other (non Ph.D) degree				0.0598 *(0.0942)*	0.0599 *(0.0949)*	0.0250 *(0.0910)*
N	137,215	136,076	136,076	128,547	128,547	128,547
Controls:						
Subject-year		X	X	X	X	X
PI publication history			X	X	X	X
PI career characteristics				X	X	X
PI grant history					X	X
PI institution/demographics						X

Notes: See Table 1

Supplementary Table 2 Continued: Do Peer Review Scores Predict Future Citations?
All Covariates

	Dependent Variable: *Future Citations*					
	(1)	(2)	(3)	(4)	(5)	(6)
Within 5 years of highest degree				-0.0144	-0.0363	-0.0376
				(0.0457)	*(0.0457)*	*(0.0450)*
Within 10 years of highest degree				-0.1598***	-0.2128***	-0.2074***
				(0.0425)	*(0.0428)*	*(0.0425)*
Within 15 years of highest degree				-0.2578***	-0.3284***	-0.3144***
				(0.0429)	*(0.0439)*	*(0.0435)*
Within 20 years of highest degree				-0.2976***	-0.3751***	-0.3590***
				(0.0435)	*(0.0447)*	*(0.0443)*
Within 25 years of highest degree				-0.3237***	-0.4023***	-0.3893***
				(0.0452)	*(0.0465)*	*(0.0461)*
Within 30 or more years of highest degree				-0.3975***	-0.4746***	-0.4630***
				(0.0453)	*(0.0464)*	*(0.0460)*
No career age information				-0.0582	-0.0213	0.0132
				(0.0520)	*(0.0524)*	*(0.0515)*
1 previous R01					0.0857***	0.0771***
					(0.0173)	*(0.0172)*
2 or more previous R01s					0.1259***	0.1072***
					(0.0171)	*(0.0170)*
1 previous non R01					0.0251**	0.0284**
					(0.0121)	*(0.0121)*
2 or more previous non R01s					0.0253*	0.0277**
					(0.0131)	*(0.0132)*
Top 5 Research Institution						0.0884***
						(0.0195)
Top 10 Research Institution						0.0043
						(0.0201)
Top 20 Research Institution						0.0569***
						(0.0161)
Top 50 Research Institution						0.0480***
						(0.0136)
Unknown Institution						-0.1621***
						(0.0248)
Female						-0.1585***
						(0.0119)
Asian						0.1577***
						(0.0180)
Hispanic						0.0455*
						(0.0248)
No ethnicity information						0.1126***
						(0.0345)
N	137,215	136,076	136,076	128,547	128,547	128,547
Controls:						
Subject-year		X	X	X	X	X
PI publication history			X	X	X	X
PI career characteristics				X	X	X
PI grant history					X	X
PI institution/demographics						X

Notes: See Table 1

Supplementary Table 3: Do Peer Review Scores Predict Future Publications? All Covariates

	Dependent Variable: *Future Publications*					
	(1)	(2)	(3)	(4)	(5)	(6)
Independent Variable: *NIH Percentile Score*	-0.0155***	-0.0091***	-0.0076***	-0.0077***	-0.0076***	-0.0075***
	(0.0003)	*(0.0003)*	*(0.0003)*	*(0.0003)*	*(0.0003)*	*(0.0003)*
Competing Renewal		0.5731***	0.5335***	0.5560***	0.5440***	0.5483***
		(0.0069)	*(0.0066)*	*(0.0069)*	*(0.0072)*	*(0.0072)*
# of Citations, past 5 years, first or last author			0.0000	0.0000	0.0000	0.0000
			0.0000	*0.0000*	*0.0000*	*0.0000*
# of Publications, past 5 years, first or last author			0.0088***	0.0093***	0.0091***	0.0087***
			(0.0009)	*(0.0010)*	*(0.0010)*	*(0.0010)*
# Top 0.1% acknowledged publications in past 5 years, first or last author			0.0148***	0.0151***	0.0144***	0.0136***
			(0.0028)	*(0.0030)*	*(0.0029)*	*(0.0029)*
# Top 1% acknowledged publications in past 5 years, first or last author			-0.0102***	-0.0104***	-0.0102***	-0.0093**
			(0.0039)	*(0.0039)*	*(0.0038)*	*(0.0038)*
# Top 5% acknowledged publications in past 5 years, first or last author			-0.0032	-0.0053	-0.0054	-0.0053
			(0.0046)	*(0.0045)*	*(0.0045)*	*(0.0045)*
# of Citations, past 5 years			-0.0001***	-0.0001***	-0.0001***	-0.0001***
			0.0000	*0.0000*	*0.0000*	*0.0000*
# of Publications, past 5 years			-0.0056***	-0.0059***	-0.0058***	-0.0058***
			(0.0005)	*(0.0005)*	*(0.0005)*	*(0.0006)*
# Top 0.1% acknowledged publications in past 5 years			0.0086***	0.0093***	0.0092***	0.0100***
			(0.0022)	*(0.0023)*	*(0.0023)*	*(0.0023)*
# Top 1% acknowledged publications in past 5 years			-0.0036	-0.0042	-0.0041	-0.0045
			(0.0030)	*(0.0031)*	*(0.0030)*	*(0.0029)*
# Top 5% acknowledged publications in past 5 years			0.0044	0.0048	0.0047	0.0041
			(0.0032)	*(0.0031)*	*(0.0031)*	*(0.0030)*
Has MD				0.1017***	0.1033***	0.1001***
				(0.0082)	*(0.0084)*	*(0.0084)*
Has MD/Ph.D.				0.1666***	0.1641***	0.1508***
				(0.0110)	*(0.0110)*	*(0.0113)*
Has other (non Ph.D) degree				0.0967**	0.0947**	0.0964**
				(0.0480)	*(0.0481)*	*(0.0478)*
N	137,215	136,171	136,171	128,638	128,638	128,638
Controls:						
Subject-year		X	X	X	X	X
PI publication history			X	X	X	X
PI career characteristics				X	X	X
PI grant history					X	X
PI institution/demographics						X

Notes: See Table 1

Supplementary Table 3 Continued: Do Peer Review Scores Predict Future Publications?
All Covariates

	Dependent Variable: *Future Publications*					
	(1)	(2)	(3)	(4)	(5)	(6)
Within 5 years of highest degree				0.0189	0.0111	0.0058
				(0.0261)	*(0.0263)*	*(0.0260)*
Within 10 years of highest degree				-0.0376	-0.0561**	-0.0598**
				(0.0264)	*(0.0266)*	*(0.0264)*
Within 15 years of highest degree				-0.0674**	-0.0928***	-0.0938***
				(0.0266)	*(0.0271)*	*(0.0267)*
Within 20 years of highest degree				-0.0604**	-0.0892***	-0.0901***
				(0.0265)	*(0.0271)*	*(0.0269)*
Within 25 years of highest degree				-0.0645**	-0.0955***	-0.0981***
				(0.0278)	*(0.0283)*	*(0.0279)*
Within 30 or more years of highest degree				-0.0682**	-0.1003***	-0.1053***
				(0.0274)	*(0.0277)*	*(0.0274)*
No career age information				-0.0454*	-0.0235	-0.0192
				(0.0262)	*(0.0263)*	*(0.0257)*
1 previous R01					0.0201**	0.0196**
					(0.0092)	*(0.0091)*
2 or more previous R01s					0.0442***	0.0422***
					(0.0095)	*(0.0095)*
1 previous non R01					0.0263***	0.0300***
					(0.0064)	*(0.0064)*
2 or more previous non R01s					0.0423***	0.0506***
					(0.0081)	*(0.0081)*
Top 5 Research Institution						-0.0018
						(0.0114)
Top 10 Research Institution						-0.0234**
						(0.0113)
Top 20 Research Institution						0.0115
						(0.0089)
Top 50 Research Institution						-0.0269***
						(0.0077)
Unknown Institution						-0.1076***
						(0.0134)
Female						-0.1075***
						(0.0064)
Asian						0.1845***
						(0.0093)
Hispanic						0.0862***
						(0.0137)
No ethnicity information						0.1426***
						(0.0213)
N	137,215	136,111	136,111	128,580	128,580	128,580
Controls:						
Subject-year		X	X	X	X	X
PI publication history			X	X	X	X
PI career characteristics				X	X	X
PI grant history					X	X
PI institution/demographics						X

Notes: See Table 1

Supplementary Table 4: Do Peer Review Scores Predict Future Citations and Publications?
Alternative Outcome Variables

	Main Estimate (Acknowledged within 5 years)	*Acknowledged within 10 years*	*Name-matched within 5 years*	*Name-matched within 10 years*
	(1)	(2)	(3)	(4)
		Dependent Variable: *Future Citations*		
Independent Variable: *NIH Percentile Score*	-0.0158*** *(0.0007)*	-0.0153*** *(0.0007)*	-0.0100*** *(0.0007)*	-0.0104*** *(0.0007)*
N	128,547	128,607	129,283	129,288
		Dependent Variable: *Future Publications*		
Independent Variable: *NIH Percentile Score*	-0.0075*** *(0.0003)*	-0.0074*** *(0.0003)*	-0.0044*** *(0.0012)*	-0.0051*** *(0.0012)*
N	128,580	128,638	136,901	136,903

Notes: Each reported figure is the coefficient on scores from a single Poisson regression of grant outcomes on NIH peer review scores; standard errors are reported in parentheses. All columns include the full set of controls described in Column 6 of Table 1. Column 1 defines future citations and publications based on all publications published within 5 years of grant award that acknowledge a grant's main project number. Column 2 extends this window to 10 years. Column 3 includes all publications by the same applicant within a 5 year window, regardless of grant acknowledgement. Column 4 examines publications within 10 years by the same applicant, without need for an explicit acknowledgement. See notes to Table 1 and Supporting Online Material for more details.

Supplementary Table 5: Do Peer Review Scores Predict Future Citations and Publications?
Robustness to Alternative Specifications

	Main Estimate	*Funding Amount*	*Institution Fixed Effects*	*Additional Past Publication Details*	*No Past Publication Details*	*Rare Names Only*	*SCORE < 15*
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
				Dependent Variable: *Future Citations*			
Independent Variable:	-0.0158***	-0.0158***	-0.0152***	-0.0131***	-0.0196***	-0.0144***	-0.0290***
NIH Percentile Score	*(0.0007)*	*(0.0007)*	*(0.0007)*	*(0.0007)*	*(0.0008)*	*(0.0008)*	*(0.0019)*
N	128,547	128,547	128,547	128,547	128,547	109,592	76,056
				Dependent Variable: *Future Publications*			
Independent Variable:	-0.0074***	-0.0075***	-0.0076***	-0.0068***	-0.0085***	-0.0061***	-0.0118***
NIH Percentile Score	*(0.0003)*	*(0.0003)*	*(0.0003)*	*(0.0003)*	*(0.0003)*	*(0.0004)*	*(0.0009)*
N	128,638	128,580	128,580	128,580	128,580	109,619	76,097

Notes: Each reported figure is the coefficient on scores from a single Poisson regression of grant outcomes on NIH peer review scores; standard errors are in parentheses. Unless otherwise stated, all columns include the full set of controls described in Column 6 of Table 1. Column 1 reproduces our main estimate from Table 1. Column 2 includes controls for the amount of funding a grant receives. Column 3 includes fixed effects for an individual's insitutional affiliation at the time of grant review. Column 4 includes more detailed controls for an applicant's publication history as described in SOM D.3. Column 5 estimates our main specification but without an information about an applicant's past publications. Column 6 restrict the sample to authors whose first initial, middle initial, and last name appear only once in PubMed. Column 7 restricts the sample to applications that were scored 15 or lower. See the Supporting Online Material and notes to Table 1 for more details.

Supplementary Table 6: Do Peer Review Scores Predict Future Citations and Publications? Alternative Samples

	Type 1 (New Grants)	*Type 2 (Renewal Grants)*	*No prior NIH funding*	*Prior NIH funding*	*Experience <10 Years*	*Experience >10 Years*
	(1)	(2)	(3)	(4)	(5)	(6)
	Dependent Variable: *Future Citations*					
Independent Variable: *NIH Percentile Score*	-0.0129*** *(0.0010)*	-0.0179*** *(0.0009)*	-0.0143*** *(0.0018)*	-0.0146*** *(0.0008)*	-0.0155*** *(0.0016)*	-0.0141*** *(0.0008)*
N	71,185	56,365	19,639	107,525	24,519	97,740
	Dependent Variable: *Future Publications*					
Independent Variable: *NIH Percentile Score*	-0.0055*** *(0.0005)*	-0.0090*** *(0.0004)*	-0.0048*** *(0.0009)*	-0.0066*** *(0.0004)*	-0.0048*** *(0.0007)*	-0.0066*** *(0.0004)*
N	71,236	56,367	19,710	107,544	24,525	97,756

Notes: Each reported figure is the coefficient on scores from a single Poisson regression of grant outcomes on NIH peer review scores; standard errors are reported in parentheses. All columns include the full set of controls described in Column 6 of Table 1. Column 1 restricts to new R01 grants. Column 2 restricts to renewed R01s. Column 3 restricts to PIs who have not been the primary recipient of any prior NIH-funding (can be non R01); Column 4 focuses on PIs who have received prior NIH-funding. Columns 5 and 6 split the sample based on whether a PI is within 10 years of her highest degree. See notes to Table 1 and the Supporting Online Material for more details.

Supplementary Table 7: Do Peer Review Scores Predict Future Citations and Publications? Linear Regression

	(1)	(2)	(3)	(4)	(5)	(6)
	Dependent Variable: *Future Citations*					
Independent Variable:	-5.3365***	-5.6592***	-4.1540***	-4.2745***	-4.2413***	-4.1420***
NIH Percentile Score	*(0.1608)*	*(0.2329)*	*(0.1976)*	*(0.2074)*	*(0.2086)*	*(0.2066)*
N	137,215	137,215	137,215	129,615	129,615	129,615
	Dependent Variable: *Future Publications*					
Independent Variable:	-0.1056***	-0.0590***	-0.0475***	-0.0480***	-0.0476***	-0.0475***
NIH Percentile Score	*(0.0021)*	*(0.0024)*	*(0.0022)*	*(0.0023)*	*(0.0023)*	*(0.0023)*
N	137,215	137,215	137,215	129,615	129,615	129,615
Controls:						
Subject-year		X	X	X	X	X
PI publication history			X	X	X	X
PI career characteristics				X	X	X
PI grant history					X	X
PI institution/demographics						X

Notes: Each reported figure is the coefficient on scores from a single OLS regression of grant outcomes on NIH peer review scores; standard errors clustered at the study section year are reported in parentheses. See notes to Table 1 and Supporting Online Materials for more details.

Supplementary Table 8: Do Peer Review Scores Predict Hit Publications and Follow-on Patents? Linear Regression

	Dependent Variable: *High Impact Publications*			Dependent Variable: *Patents*	
	Top 0.1%	Top 1%	Top 5%	Direct	Indirect
	(1)	(2)	(3)	(4)	(5)
Independent Variable:	-0.0008***	-0.0038***	-0.0133***	-0.0765***	-0.0015***
NIH Percentile Score	*(0.0001)*	*(0.0003)*	*(0.0007)*	*(0.0081)*	*(0.0002)*
N	129,615	129,615	129,615	129,615	129,615

Notes: Each reported figure is the coefficient on scores from a single OLS regression of grant outcomes on NIH peer review scores. All columns include the full set of controls described in Column 6 of Table 1.

References

1. B. Alberts, M. W. Kirschner, S. Tilghman, H. Varmus, Rescuing US biomedical research from its systemic flaws. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 5773–5777 (2014). Medline doi:10.1073/pnas.1404402111

2. D. F. Horrobin, The philosophical basis of peer review and the suppression of innovation. *JAMA* **263**, 1438–1441 (1990). Medline doi:10.1001/jama.1990.03440100162024

3. J. M. Campanario, Peer review for journals as it stands today–Part 1. *Sci. Commun.* **19**, 181–211 (1998). doi:10.1177/1075547098019003002

4. S. Cole, J. R. Cole, G. A. Simon, Chance and consensus in peer review. *Science* **214**, 881–886 (1981). Medline doi:10.1126/science.7302566

5. J. Berg, Productivity metrics and peer review scores: NIGMS feedback loop blog (2011); https://loop.nigms.nih.gov/2011/06/productivity-metrics-and-peer-review-scores/.

6. J. Berg, Productivity metrics and peer review scores, continued: NIGMS feedback loop blog (2011); https://loop.nigms.nih.gov/2011/06/productivity-metrics-and-peer-review-scores-continued.

7. N. Danthi, C. O. Wu, P. Shi, M. Lauer, Percentile ranking and citation impact of a large cohort of National Heart, Lung, and Blood Institute-funded cardiovascular R01 grants. *Circ. Res.* **114**, 600–606 (2014). Medline doi:10.1161/CIRCRESAHA.114.302656

8. B. A. Jacob, L. Lefgren, The impact of NIH postdoctoral training grants on scientific productivity. *Res. Policy* **40**, 864–874 (2011). Medline doi:10.1016/j.respol.2011.04.003

9. B. A. Jacob, L. Lefgren, The impact of research grant funding on scientific productivity. *J. Public Econ.* **95**, 1168–1177 (2011). Medline doi:10.1016/j.jpubeco.2011.05.005

10. J. H. Tanne, US National Institutes of Health updates peer review system. *BMJ* **319**, 336 (1999). Medline doi:10.1136/bmj.319.7206.336

11. K. Arrow, The rate and direction of inventive activity: Economic and social factors (National Bureau of Economic Research, Cambridge, MA, 1962), pp. 609–626.

12. About NIH Web site (2014); http://www.nih.gov/about/.

13. E. R. Dorsey, J. de Roulet, J. P. Thompson, J. I. Reminick, A. Thai, Z. White-Stellato, C. A. Beck, B. P. George, H. Moses 3rd, Funding of US biomedical research, 2003-2008. *JAMA* **303**, 137–143 (2010). Medline doi:10.1001/jama.2009.1987

14. There is no further disambiguation, but we show that our results do not change when we restrict to investigators with rare names. See table S5 of the supplementary materials.

15. W. R. Kerr, The ethnic composition of US inventors, Working Paper 08-006, Harvard Business School (2008); http://www.people.hbs.edu/wkerr/Kerr%20WP08_EthMatch.pdf.

16. W. R. Kerr, *Rev. Econ. Stat.* **90**, 518 (2008).

17. Due to the limitations of the name-based matching algorithm, we cannot reliably distinguish African-American investigators.

18. For example, to calculate the 14.6% figure, we take the exponential of our estimated coefficient times the SD in scores, minus 1: exp(–0.0155 × 10.17) – 1.

19. R. K. Merton, The Matthew effect in science: The reward and communication systems of science are considered. *Science* **159**, 56–63 (1968). doi:10.1126/science.159.3810.56

20. P. Azoulay, T. Stuart, Y. Wang, Matthew: Effect or fable? *Manage. Sci.* **60**, 92–109 (2013). doi:10.1287/mnsc.2013.1755

21. P. Azoulay, J. S. G. Zivin, B. N. Sampat, The diffusion of scientific knowledge across time and space: Evidence from professional transitions for the superstars of medicine, Tech. Rep., National Bureau of Economic Research (NBER, Cambridge, MA, 2011).

22. P. Azoulay, J. Graff-Zivin, D. Li, B. Sampat, Public R&D investments and private sector patenting: Evidence from NIH funding rules, NBER working paper 20889 (2013); http://irps.ucsd.edu/assets/001/506033.pdf.

23. J. J. Heckman, Sample selection bias as a specification error. *Econometrica* **47**, 153–161 (1979).

24. J. H. Abbring, J. J. Heckman, in *Handbook of Econometrics*, v. 6, pt. 2 (Elsevier Science, Amsterdam, 2007), chap. 72.

25. J. Callaert, B. Van Looy, A. Verbeek, K. Debackere, B. Thijs, Traces of prior art: An analysis of non-patent references found in patent documents. *Scientometrics* **69**, 3–20 (2006). doi:10.1007/s11192-006-0135-8

26. B. H. Hall, A. B. Jaffe, M. Trajtenberg, The NBER patent citation data file: Lessons, insights and methodological tools, Tech. Rep., National Bureau of Economic Research (NBER, Cambridge, MA, 2001).

27. T. J. Kane, D. O. Staiger, Estimating teacher impacts on student achievement: An experimental evaluation, Tech. Rep., National Bureau of Economic Research (NBER, Cambridge, MA, 2008).

28. J. E. Rockoff, The impact of individual teachers on student achievement: Evidence from panel data. *Am. Econ. Rev.* **94**, 247–252 (2004). doi:10.1257/0002828041302244

29. S. G. Rivkin, E. A. Hanushek, J. F. Kain, Teachers, schools, and academic achievement. *Econometrica* **73**, 417–458 (2005). doi:10.1111/j.1468-0262.2005.00584.x