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

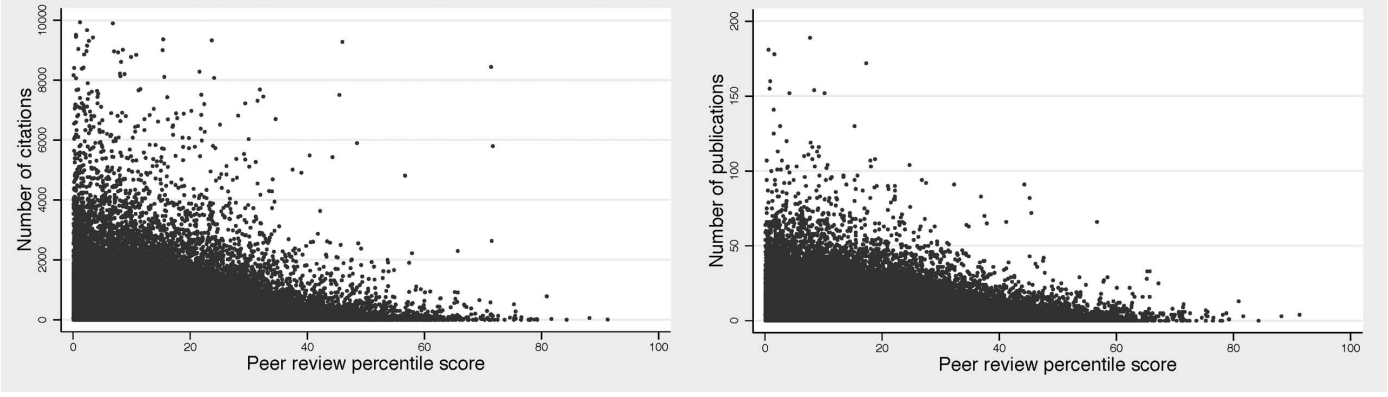

Fig. 1. Scatterplot of percentile scores and grant outcomes. The left panel plots the relationship between percentile scores and citations associated with a grant. Each dot represents a single grant. The right panel does the same for total publications. Extreme outliers with more than 10,000 citations 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

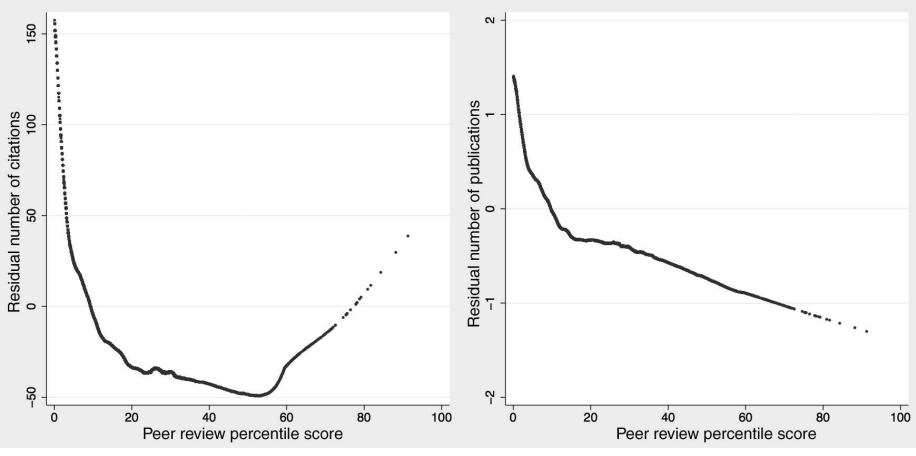

Fig. 2. Smoothed scatterplots of percentile scores and residual grant outcomes. These figures display smoothed scatterplots of the nonparametric relationship between unexplained variation in grant outcomes and percentile score, after accounting for differences in field of research, year, and applicant 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

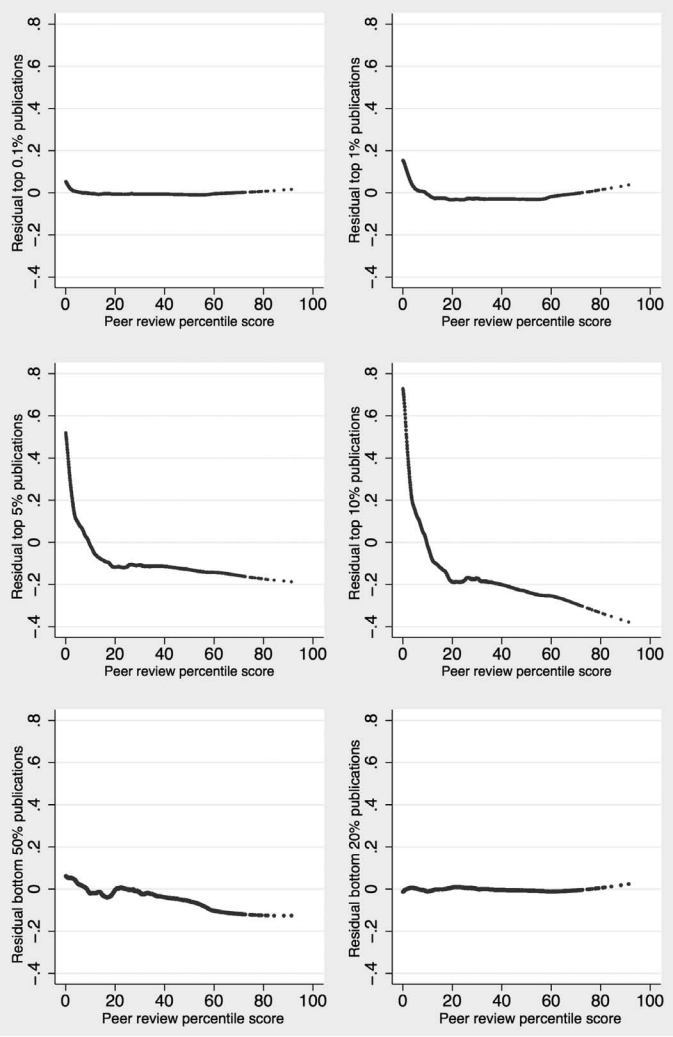

Fig. 3. Smoothed scatterplots of percentile scores and residual high- and low-citation publications. These figures display smoothed scatterplots of the nonparametric relationship between unexplained variation in grant outcomes and percentile score, after accounting for variation in field of research, 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
AAAS

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

Danielle Li and Leila Agha (April 23, 2015)

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

Editor's Summary

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 projects 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 more 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

Science (print ISSN 0036-8075; online ISSN 1095-9203) is published weekly, except the last week in December, by the American Association for the Advancement of Science, 1200 New York Avenue NW, Washington, DC 20005.