Towards recovery: Scientists with better ratings of their institution's response to the COVID-19 pandemic have more optimistic forecasts about their future research

Kyle R. Myers*[1,2], Karim R. Lakhani[1,2,3], Dashun Wang[4,5,6,7]

[1]Harvard Business School, Harvard University, Boston, MA, USA
[2]Laboratory for Innovation Science at Harvard, Harvard University, Boston, MA, USA
[3]Institute for Quantitative Social Science, Harvard University, Boston, MA, USA
[4]Center for Science of Science and Innovation, Northwestern University, Evanston, IL, USA
[5]Northwestern Institute on Complex Systems, Northwestern University, Evanston, IL, USA
[6]McCormick School of Engineering, Northwestern University, Evanston, IL, USA
[7]Kellogg School of Management, Northwestern University, Evanston, IL, USA
*Corresponding author, email: kmyers@hbs.edu

Abstract

Using a survey with randomized elements and truth-telling incentives, we document scientists' beliefs about current and future disruptions to their research and how these beliefs depend on the eventual length of the COVID-19 pandemic. Overall, scientists with more favorable ratings of their institution's response to the pandemic have more optimistic forecasts about their research, even when controlling for their current level of disruptions. This relationship holds amongst groups with the most pessimistic forecasts, who should be prioritized as science policy responses continue to develop.

Introduction

Many of the most pressing questions in science policy today concern the support and recovery from the dramatic disruptions caused by the COVID-19 pandemic. Research institutions around the world have implemented, and continue to develop, policies and initiatives in hope of alleviating the setbacks caused by the pandemic. An understanding how well institutions' efforts may be able to mitigate the pandemic's effects is not only immediately useful to ongoing policy debates, but it may also prove important for the long-term vitality of the scientific enterprise.

Initial work has used data on article pre-prints, grant applications, and time-use surveys to investigate the immediate effects of this pandemic, finding very heterogeneous experiences across different groups of scientists [1-3]. But it remains unclear whether these early data points are good predictors of long-term outcomes, or whether institutions have the capabilities to mitigate these disruptions given the sweeping scale of the pandemic. For example, research universities and colleges have granted tenure clock extensions, funding agencies have loosened regulations surrounding no-cost extensions on grants, and national academies have convened panels to collect and disseminate new knowledge [4-5]. But are these efforts helping? And if so, whom are they helping?

Here, we present some of the first systematic evidence that relates scientists' beliefs about their institutions' pandemic responses to forecasts of their research productivity in the coming years. Our main goal is not to estimate any causal effects of institutions' policies that have been put in place in response to the pandemic. This early in the pandemic, we lack large-scale data on policy specifics not to mention exogenous variation in these policies. Rather, the main null hypothesis we test is that the sweeping scale of the pandemic has been (and continues to be) so dramatic that we could not estimate any significant correlation – statistically or economically speaking – between scientists' research forecasts and their ratings of their institution's handling of the pandemic. Importantly, our survey uses randomization techniques and truth-telling incentives in an effort to ensure that the forecast-rating correlation we estimate is not driven by differences in beliefs about the duration of the pandemic or dishonest responses.

With our data, we expect that such a correlation could exist even if it were truly the case that research institutions' pandemic policies have been ineffective for the following reason: if a scientist has not faced major disruptions due to the pandemic, then they will likely demand less support from their institution; therefore, even if the institution's pandemic policies did not truly have an effect on the scientists' future research, a correlation could arise because of the lack of disruptions leading to both an optimistic forecast and a better rating of the institution. For this reason, we also use the large number of other covariates collected in our survey in an attempt to condition this correlation on the degree to which scientists have been disrupted by the pandemic (i.e., changes in time allocations, field of study, age, reported project interruptions, etc.). In short, we find that the forecast-rating correlation is robust and largely unchanged even when conditioning on these measures – we cannot "kill" the association. At the end of the paper, we briefly discuss the implications for policymakers as well as researchers who will come to study the effects of this pandemic on the scientific enterprise.

Methodology

Survey and measurement methods

To provide timely data on how the pandemic has and may continue to affect scientists, we conducted a survey of active academic researchers from June to August 2020. In this survey, we asked scientists about recent or ongoing disruptions to their work (e.g., changes in work time; a permanent disruption to an ongoing research project; a lost promotion), as well as their forecasts of how much their research might ultimately be disrupted in the future. Using five questions (see Table 1), we asked scientists to forecast changes in their research funding, publication output and impact, as well as the amount of time and funding it would take to return to a "normal" pace of work once the pandemic was over.

Table 1. Forecasted disruption questions.

Pandemic horizon preface		Assuming the pandemic ends in [*randomized horizon*] months from now…
Fundraising question	**Q1.**	…how much research funding do you expect in 2020-2021 compared to 2018-2019?
Publishing questions	**Q2.**	…how do you expect your publication quantity in 2020-2021 to compare to 2018-2019?
	Q3.	…how do you expect your publication impact in 2020-2021 to compare to 2018-2019?
Startup costs questions	**Q4.**	…how much time do you expect it to take to resume your normal pace of work?
	Q5.	…how much funding do you expect it will take to resume your normal pace of work?

Caption: The "[*randomized horizon*]" inserted into the preface of each of the five questions was randomized to equal either 1, 2, 3, 4, 6, or 8 months, with "now" implying the date at which the respondent was taking the survey.

A major challenge to forecasting the long-term effects of the pandemic is that it will clearly depend on how long the pandemic lasts, but the ultimate length of the pandemic is very uncertain. Thus, if a scientist reported a relatively optimistic forecast, it might be because they were less affected by the pandemic compared to others, or it might be that they expected the pandemic to end much sooner than others. To tackle this challenge, we incorporated a randomized design into our survey. When posing the five forecast questions, we asked the respondent to use a specific assumption about the eventual duration of the pandemic. We randomly presented each respondent with just one of these assumptions about the pandemic's horizon in the preamble of the five forecast questions – offering horizons between one and eight months into the future (e.g., "Assuming the pandemic ends in eight months," see the first row of Table 1). This random variation allows us to attribute differences in scientists' responses to how much the pandemic is affecting their work.

Using a simple standardized composite of their responses to the five forecast questions, we created a single variable we term the Forecasted Disruption Index (FDI) where larger values indicated the scientist forecasted relatively more disruptions compared to others. The FDI is based on standardizations of responses, so that a difference of one unit indicates that the respondent forecasts disruptions that are one standard deviation larger than the sample average.

In addition to the current and forecasted research disruption questions, we asked scientists to rate the overall quality of their own institution's response to the pandemic. The question was phrased as a five-point Likert scale: excellent, good, average, poor, terrible. For inclusion in the regressions we estimate, we collapse the poor and terrible ratings into a single rating (due to small numbers) and then convert into a simple continuous measure with one integer between each point on the scale ranging from one (poor/terrible) to four (excellent).

We also included a series of professional and socio-demographic questions as well as a number of questions related to ways in which the pandemic had already disrupted the scientists' research. Below, we report the results from the responses from 3,204 respondents who self-identified as either research faculty or principal investigators. The Supporting information contains further details on the survey design, sampling, response rates, and how we construct the FDI.

Soliciting beliefs: Approach and limitations

An important limitation of this approach is that when soliciting forecasts, some of the variation may reflect behavioral biases, an unwillingness to use the assumption we provided, or other incentives that lead respondents to not put effort into their response or to not report honestly. To address this concern, we encouraged respondents to report honestly by using incentives that stem from the Bayesian Truth Serum mechanism [6]. In short, each respondent is rewarded in a way that (theoretically) induces all respondents to report honestly; see the Supporting information for more.

Obviously, we would have preferred to be using standardized data on the specific policies put in place at each institution so we might better identify policy features that were more or less related to scientists' forecasts. But at this stage in the pandemic, that data does not exist. Thus, as we mentioned in the introduction, our approach provides a useful second-best data point that can begin to shed some light on the role of research institutions. In the discussion, we reference some very preliminary results based on publicly available crowd-sourced data on a few policy dimensions. But digging deeper into policy specifics will be an important future endeavor.

Empirical models

In our empirical analyses, we begin with a simple model to estimate the forecast for scientist i as follows:

$$ForecastedDisruptionIndex_i = \alpha + InstitutionRating_i \times \beta + PandemicHorizon_i \times \gamma + f(X_i) + \epsilon_i\ , \quad (1)$$

where the dependent variable is the FDI, the "$PandemicHorizon$" is the randomized assumption posed to the respondent about the months to the end of the pandemic, and X is a (possibly empty) vector of covariates. Instead of relying on manual specification of which of the many covariates (also solicited in the survey) we should include in each regression, we use standard Lasso-based model selection algorithms to select from a large set of possible covariates and transformations thereof. Depending on the specification, this set of possible covariates includes at least a set of indicator variables for the calendar week when the respondent completed the survey, and may also include all of the demographic, professional, and pandemic-related variables we solicit in the survey (see the Supporting information for the full list of possible covariates and the transformations).

In some specifications, we group scientists into groups g according to dimensions such as their institutional rating, field of study, or gender identity and estimate a specification of the forms:

$$ForecastedDisruptionIndex_i = \alpha_{g(i)} + PandemicHorizon_i \times \gamma_{g(i)} + \epsilon_i\ , \quad (2)$$

$$ForecastedDisruptionIndex_i = \alpha + InstitutionRating_i \times \beta_{g(i)} + PandemicHorizon_i \times \gamma + f(X_i) + \epsilon_i\ , \quad (3)$$

where Equation 2 will allow us to identify what groups of scientists whose forecasts appear most sensitive to the pandemic and its duration (per $\gamma_{g(i)}$ and $\alpha_{g(i)}$, the latter representing group-specific intercepts), and Equation 3 will allow to explore whether the forecast-ratings association varies across groups (per $\beta_{g(i)}$). Again, heterogenous forecast-ratings associations across different groups would not necessarily imply that policies are differentially helping certain scientists. But these initial exploratory analyses can still motivate future work as to how institutions' policies may have (purposefully or not) had unequal effects on scientists.

Results

Summary statistics

Table 2 reports the summary statistics for the five FDI questions, as well as a select set of the other covariates collected in the survey. The left-hand panel of the table reports the research disruptions variables. The summary statistics of the FDI are a mean of zero and a standard deviation of one by construction. More importantly, a one standard deviation change in the FDI is roughly equivalent to either a 20 p.p. decline in research funding or a 16-22 p.p. decline in publication output or impact in the coming two years, or needing two months longer or 35 p.p. more funding to resume a normal pace of work once the pandemic is over. Compared to the sample averages, these changes are clearly meaningful, where a 1 standard deviation increase in the FDI, if concentrated within a single component, is associated with a 100-200% (more negative) change relative to average. The right-hand side reports the summary statistics for a set of socio-demographic and professional variables.

Table 2. Summary statistics.

Disruptions, forecasted & experienced	**mean (s.d.)**	**Professional & socio-demographics**	**mean (s.d.)**
Forecasted change in funding (p.p.)	-18.1 (36.6)	Rating of institution's response {1-4}	2.67 (0.91)
Forecasted change in pub. quantity (p.p.)	-11.2 (31.3)	Age (years)	45.4 (13.1)
Forecasted change in pub. impact (p.p.)	-6.3 (24.1)	Gender identity=female {0,1}	0.37 (0.48)
Forecasted time to resume normal pace (days)	80.9 (94.4)	Located in U.S. {0,1}	0.71 (0.46)
Forecasted funding to resume normal pace (p.p.)	37.8 (65.1)	Field=Math, Stat., Econ., Business {0,1}	0.15
Experienced major project disruption {0,1}	0.28 (0.45)	Field=Health, Medicine {0,1}	0.18
Experienced lost opportunity to disseminate {0,1}	0.58 (0.49)	Field=Biology, Chemistry {0,1}	0.21
Experienced concern over income {0,1}	0.28 (0.45)	Field=Social sci., Psychology {0,1}	0.23
Experienced concern over job security {0,1}	0.22 (0.41)	Field=Physics, Eng., Enviro. {0,1}	0.23
Experienced concern over promotion {0,1}	0.06 (0.25)	Has 0-5-year-old dependent {0,1}	0.18 (0.38)

Experienced change in research time (p.p.)	-5.84 (21.3)	At higher education institution {0,1}	0.77 (0.43)
Experienced change in fundraising time (p.p.)	0.07 (10.6)	Tenured, if applicable {0,1}	0.68 (0.47)
		Soft money {0,1}	0.38 (0.48)

Caption: Based on 3,204 scientist-level observations (except for the Tenured variable, which is based on the 2,385 respondents who reported tenure as being applicable). (p.p.) indicates the variable is a percentage point, with the baseline being set to pre-pandemic levels. {0,1} indicates binary zero-one indicators. {1-4} indicates the variable takes on integer values between 1 and 4. "Soft money" refers to the researcher being required to fundraise a non-zero portion of their salary via grants. Non-U.S. based respondents are almost entirely from Canada, Europe, or a few countries in western Asia. Respondents not at higher education institutions are at other non-profit research institutions; researchers at for-profit organizations are excluded from this sample.

Do scientists think institutions are helping?

Table 3, columns (1-3) report the results from estimating versions of Equation 1 – relating scientists' FDI to their ratings of their institution's response to the pandemic – using increasingly saturated control-variable sets. As we successively introduce more control variables, we do depress the coefficient on "$InstitutionRating$" towards zero, as one would expect given the likelihood of endogeneity we highlighted earlier in the paper. But even in the most saturated model (column 3), we continue to estimate a relatively precise and meaningful forecast-rating association. Based on the estimated coefficient that corresponds to the pandemic horizon (which we also estimate with relative precision), this specification indicates that the difference in the FDI between two scientists who rated their institutions one point different is roughly equivalent to the difference in the FDI that was reported between two scientists who had a four months difference in their assumed pandemic horizon (0.115 / 0.029 = 3.97). To reiterate, our approach does not necessarily imply that this relationship is due to institutions' policies, per se. Still, this suggests that, despite the grand scale of the pandemic, there may be a role for institutional-level policies to help researchers cope with their research disruptions.

Table 3. Forecast-rating regression results

	Dependent variable: $ForecastedDisruptionIndex$					
	Average			Group-specific		
	(1)	(2)	(3)	(4)	(5)	(6)
$InstitutionRating$	-0.255*** (0.0199)	-0.136*** (0.0191)	-0.115*** (0.0182)			
$PandemicHorizon$	0.025*** (0.0074)	0.022*** (0.0067)	0.029*** (0.0065)	0.031*** (0.0067)	0.029*** (0.0067)	0.028*** (0.0065)
$InstitutionRating$ $\times BenchScience$				-0.063*** (0.0239)		
$\times OtherFields$				-0.163*** (0.0243)		
$InstitutionRating$ $\times HasYoungDep.$					-0.163*** (0.0411)	

$\times NoYoungDep.$					-0.106***	
					(0.0191)	
$InstitutionRating$						
$\times Female$						-0.121***
						(0.0241)
$\times Male$						-0.112***
						(0.0207)
Control inclusion		+Baseline	+Full	+Full	+Full	+Full
Control selection	Manual	Manual	Lasso	Lasso	Lasso	Lasso

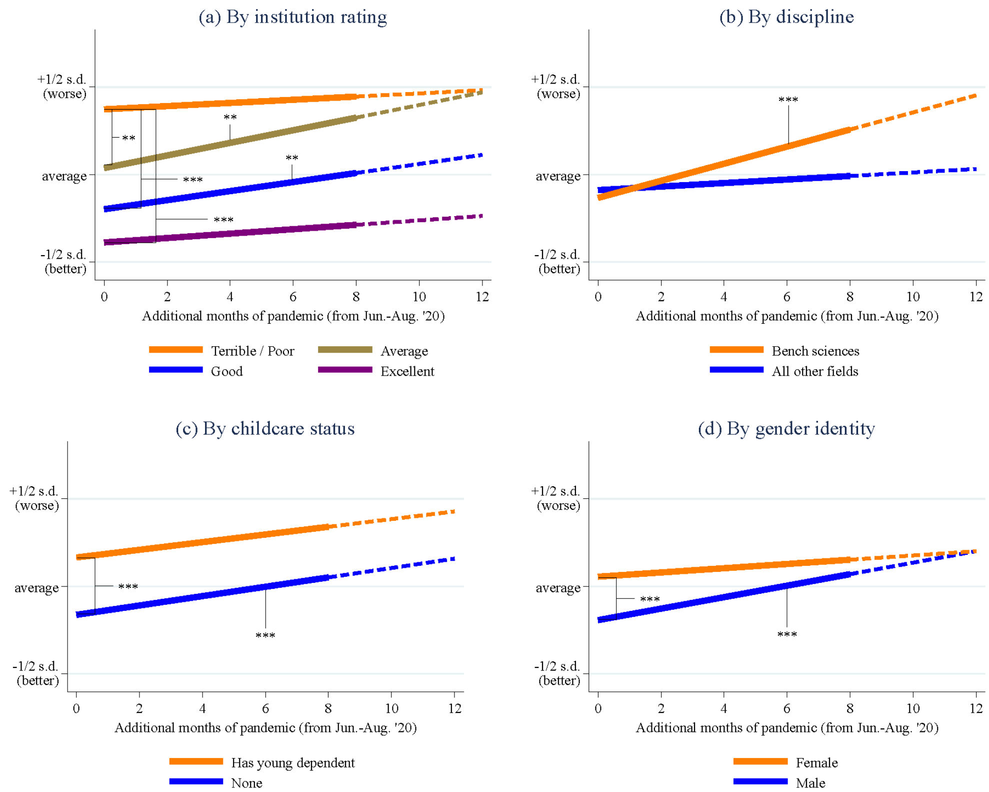
Caption: Reports results from estimating versions of Eq. 1 or Eq. 2. Based on 3,204 scientist-level observations; heteroskedasticity robust standard errors in parentheses, *p<0.1, **p<0.05, ***p<0.01. "$BenchScience$" disciplines include biology, chemistry, physics, engineering, and related fields. "+Baseline" indicates that all of the non-forecast variables in Table 2 are included as covariates (using indicator variables as needed) in addition to the horizon control and likewise, "+Full" indicates that all possible covariates, including transformations thereof (outlined in the Supporting information) are included as covariates. In the group-specific specifications, all controls are also interacted with group dummies. "Manual" control selection indicates that all possible covariates are in the regression model, while "Lasso" selection relies on the output of the Lasso routine and performs a post-Lasso regression using only the selected covariates.

Stabilizing versus deteriorating effects

Our next goal is to investigate whether this relationship persists even amongst those scientists most sensitive to the pandemic. While the controls in Equation 1 are helpful in attempting to hold fixed the amount of disruption each scientist has experienced, it does not directly test whether the forecast-rating relationship itself holds at lower and higher levels of experienced disruption. In order to identify which groups are most likely to be the most disrupted, we estimated versions of Equation 2, grouping scientists along dimensions our prior work identified as being correlated with the largest pandemic-induced disruptions: being a female researcher, being in a "bench science" (biology, chemistry, physics, engineering, etc.), or having a young (under five years old) dependent [2].

Figure 1 displays the results from these regressions, including a version of the model where we group scientists based on their institution ratings (Panel a) to further illustrate the forecast-rating correlation we documented in Table 2. Overall, our findings our consistent with our prior work [2]. When grouping based on discipline, we find that, although the baseline of disruption (the y-intercept) is very similar for all fields, the bench scientists are much more sensitive to the continued duration of the pandemic. We also find that female scientists and those with young dependents both report having more pessimistic forecasts at baseline. In both cases, we don't estimate a statistically significant horizon coefficient for these more-affected groups; however, again in both cases, we cannot reject a null hypothesis that the horizon coefficients are the same for each group (i.e., that the horizon coefficient is similar for both those with and without young dependents, and that the coefficient is the same for both male and female scientists).

Figure 1. Disruption forecasts per pandemic horizon, by groups.



Caption: Each panel of the figure plots the linear projections of the FDI (y-axis; larger values indicate larger forecasted disruptions) based on estimates of Equation 2, using four different groupings of scientists in four separate regressions. Only horizons of 1, 2, 4, 6, and 8 months were used in the survey; thus, the linear projections after 8 months are out-of-sample. $^{*}p<0.1$, $^{**}p<0.05$, $^{***}p<0.01$, where the bracketed indicators aligned on the y-axes indicate whether a pair of group-specific intercepts (α_g) are significantly different from each other, and the individual line indicators indicate whether the slope of group-specific horizon line (γ_g) are significantly different from zero.

Is there hope for those who expect the worst?

Motivated by the results in Figure 1, we estimate versions of Equation 3 to test if the forecast-rating association differs across these different groupings of scientists. Columns (4-6) in Table 2 reports the results of these regressions. When we divide scientists along gender or childcare status, we find the forecast-rating association is very similar within both groups, with statistical tests not rejecting a null of equality. In other words, whether the scientist identifies as male or female or whether they have young dependents or not, those with better ratings of their institutions have equally more optimism in their research forecasts. If it was the case that the policies underlying these ratings were clearly differentially benefiting one of these groups, we would have expected not to estimate this equality.

However, in the case where we compare bench scientists to other disciplines (Table 2 column 4), we do estimate different forecast-rating relationships, where this association is significantly weaker amongst bench scientists (*F statistic*=9.49, *p*=0.002). This indicates that equal improvements in ratings are associated with smaller (but still significant) improvements in optimism amongst bench scientists compared to other disciplines. This finding warrants more investigation by policymakers and researchers in the near future to assess whether or not bench scientists are in need of alternative policies, or if, perhaps it is the case that bench scientists have been affected by the pandemic in ways our regression model is unable to account for. Though we would note that these regression models do allow the controls to vary by groupings, so if the same reported change in a variable (e.g., time spent on research) was more important in the bench sciences than in others, our model would have accounted for that. But to be sure, there may be many other effects unobservable to us in this data.

An early look at extension policies

The only two policy-related components of the survey asked respondents whether they had received (1) an extension on deliverables (e.g., due to a funder), and (2) an extension on their tenure clock (conditional on being an untenured tenure-track researcher). These policies may drive some of the variation in institutional ratings across respondents. Although, it is important to note that these sort of extension policies might not have a clear, direct effect on the FDI, since these forecasts are over a fixed period of time (the next two years), and the entire purpose of these extension policies is to appreciate that some scientists will be less productive in the coming years, and so they will be granted additional time on their deadlines. Thus, for example, a tenure-clock extension policy could technically "work" without changing a scientist's publication output per year, simply by giving them more years in which to produce.

In the Supporting information we report a series of regression results that investigate correlations between these two variables and both the institutional ratings and FDI, always conditioning on our measures of and proxies for scientists' disruptions experienced to-date. Overall, we consistently estimate a positive association between scientists reporting deliverable extensions and both the institutional ratings and optimism (lower values) in the FDI. For the tenure extension policy, we obtain noisy and imprecise estimates that are often statistically indistinguishable from zero (see the Supporting information for more). This latter result is particularly interesting given the inequitable effects across genders that tenure-clock extension policies have been shown to have in non-pandemic times [10].

Investigating the longer-term causal effects of these sort of policies using program evaluation methods will be an important avenue for future research.

Discussion

While still far from conclusive, these patterns suggest that, despite the dramatic scale of the pandemic and its heterogeneous impacts, research institutions play an important in mitigating the effects of the pandemic and, eventually, rebuilding the scientific enterprise. We have found that there are some scientists who have stabilized and there are some that, without further support, may continue to deteriorate as the pandemic continues. Those in the latter group should be prioritized as responses continue to develop.

These findings will also eventually prove useful to those who will come to study the longer-term effects of the pandemic on the scientific enterprise. An optimistic view of the pandemic would note that it may provide some natural experiments that could be used to improve our understanding of scientific production. Our results indicate that as others undertake these sorts of studies, they should pay close attention to the unequal ways the pandemic has affected scientists, and the many different ways that institutions will have attempted to respond.

Acknowledgements

We are extremely thankful to Nina Cohodes, Alexandra Kesick, Wei Yang Tham, Jerry Thursby, Marie Thursby, Ruihan Wang, and Yian Yin for assistance with the creation, distribution, and management of the survey and responses. Peter Schiffer and Jay Walsh shared helpful thoughts on the research disruption questions. This work was financially supported by the Air Force Office of Scientific Research under award number FA9550-19-1-0354, National Science Foundation SBE 1829344, and the Alfred P. Sloan Foundation G-2019-12485 and G-2020-13873, and the Harvard Busines School.

References

1. Vincent-Lamarre P, Sugimoto CR, Larivière V. The decline of women's research production during the coronavirus pandemic. Nature Index. 2020 May 19.
2. Myers KR, Tham WY, Yin Y, Cohodes N, Thursby JG, Thursby MC, Schiffer P, Walsh JT, Lakhani KR, Wang D. Unequal effects of the COVID-19 pandemic on scientists. Nature Human Behaviour. 2020 Sep;4(9):880-3.
3. Lauer, M. An Early Look at Applications Submitted During the Pandemic. National Institutes of Health. 2020 July 28. https://nexus.od.nih.gov/all/2020/07/28/an-early-look-at-applications-submitted-during-the-pandemic/
4. Council On Government Relations. Institutional and Agency Responses to COVID-19 and Additional Resources. 2020. https://www.cogr.edu/institutional-and-agency-responses-covid-19-and-additional-resources/
5. The National Academies of Science, Engineering, and Medicine. COVID-19 Responses & Resources. 2020. https://www.nationalacademies.org/topics/covid-19-resources
6. Cvitanić J, Prelec D, Riley B, Tereick B. Honesty via choice-matching. American Economic Review: Insights. 2019 Sep;1(2):179-92.
7. Johns Hopkins University. Coronavirus Resource Center. 2020. https://coronavirus.jhu.edu
8. Tackle coronavirus in vulnerable communities. 2020. Nature, 581(7808), 239–240.
9. American Association of University Professors. Visualizing change: The annual report on the economic status of the profession, 2016-2017. 2017.
10. Antecol H, Bedard K, Stearns J. Equal but inequitable: Who benefits from gender-neutral tenure clock stopping policies? American Economic Review. 2018 Sep;108(9):2420-41.

Supplementary information for

Towards recovery: Scientists with better ratings of their institution's response to the COVID-19 pandemic have more optimistic forecasts about their future research

Kyle R. Myers*[1,2], Karim R. Lakhani[1,2,3], Dashun Wang[4,5,6,7]

[1]Harvard Business School, Harvard University, Boston, MA, USA
[2]Laboratory for Innovation Science at Harvard, Harvard University, Boston, MA, USA
[3]Institute for Quantitative Social Science, Harvard University, Boston, MA, USA
[4]Center for Science of Science and Innovation, Northwestern University, Evanston, IL, USA
[5]Northwestern Institute on Complex Systems, Northwestern University, Evanston, IL, USA
[6]McCormick School of Engineering, Northwestern University, Evanston, IL, USA
[7]Kellogg School of Management, Northwestern University, Evanston, IL, USA
*Corresponding author: kmyers@hbs.edu

Human research participants
The study protocol has been approved by the Institutional Review Board (IRB) from Harvard University and Northwestern University. Informed consent was obtained from all participants.

Author contributions
KRM, KRL, and DW conceived the project and designed the survey; KRM collected data, performed empirical analyses, and wrote the manuscript; KRM, KRL, and DW edited the manuscript.

Competing interest declaration
The authors declare no competing interests.

Data availability
Because of the sensitive nature of some of the variables collected, the IRB-approved protocol does not permit individual-level data to be made unrestricted and publicly available. Researchers interested in obtaining restricted, anonymized versions of this individual-level data should contact the authors to inquire about obtaining an IRB-approved institutional data sharing agreement.

Code availability
Code necessary to reproduce all plots and statistical analyses will be made freely available.

Table of Contents

S1 Survey sampling and recruitment

S1.1 Web of Science corresponding authors

To compile a large, plausibly random list of active scientists, we leverage the Web of Science (WoS) publication database. The WoS database is useful for two reasons: (1) it is one of the most authoritative citation corpuses available[i]; (2) among other large-scale publication datasets, WoS is the only one, to our knowledge, with systematic coverage of corresponding author email addresses.

By mid-April 2020, the cumulative number of deaths due to COVID-19 had reached approximately 115,000 with nearly 1,800 deaths per day in the U.S. and 3,000 deaths per day in Europe.[ii] Throughout the U.S. and Europe, schools and workplaces were typically required to be closed and restrictions on gatherings of more than 10 people were in place in most countries.[iii] For scientists, not only did this drastically change their daily lives, it severely limited the possibilities of using traditional workspaces.[iv]

Therefore, in this paper we are primarily interested in active scientists residing in the U.S. and Europe. We start from 21 million WoS papers published in the last decade (2010-2019). In an attempt to focus on scientists likely to still be active and in a more stable research position, we link the data to journal impact factor information (WoS Journal Citation Reports), and exclude papers published in journals in the bottom 25% of the impact factor distribution for its WoS-designated category. We use the journal impact factor calculated for the year of publication, and for papers published in 2019, we use the latest version (2018). We then extract all author email addresses associated with papers. For each email address in this list, we consider it as a potential participant if: (1) it is associated with at least two papers in the ten-year period, and (2) the most recent country of residence, defined by the first affiliation of the most recent paper, is in the U.S. or Europe.

We have approximately 2.5 million unique email addresses after filtering, with about 521,000 in the U.S. and 938,000 in Europe. We then randomly shuffled the two lists separately and sampled roughly 280,000 email addresses from the U.S. and 200,000 from Europe for a total of 480,000. We oversampled the U.S. as a part of a broader outreach strategy underlying this and other research projects.

S1.2 Participant recruitment

We recruited participants by sending them email invitations with the following text:

We need your help to shed light on how the coronavirus pandemic is affecting scientists like you. Please take a brief moment to complete this short five-minute survey as part of a research study. Your responses will help scientists and policymakers understand and respond to this rapidly evolving situation.

Follow this link to the Survey:
[link]
Or, copy and paste the URL below into your internet browser:
[link]

Upon completion, you can choose charities to receive a donation on your behalf, and you may have the chance of winning a $100 gift card. Please feel free to forward this survey to any other scientists you know (e.g., professors, post-doctoral researchers, graduate students). We need everyone's input to fully understand the breadth of how science is currently changing.

Thank you for your time,

Kyle Myers, Ph.D. & Karim Lakhani, Ph.D.
Laboratory for Innovation Science at Harvard

Dashun Wang, Ph.D.

S2 Survey instrument

S2.1 Survey questions

The survey includes questions on professional information (position type, institution type, fields of study, type of research, tenure status), demographic information (age, gender, cohabitation, dependents, citizenship), time allocation (time spent on different activities before and after the pandemic outbreak) and predicted changes in future publication and funding. Below are the excerpted questions underlying some of the main variables used in our analyses. For the full version of the survey instrument please contact the authors. We did not require respondents to answer any of the demographic questions regarding gender, age, dependents, cohabitation, or citizenship.

Q. Which of the following best describes your current position?

Faculty or Principal Investigator | Post-doctoral researcher | Graduate student in a doctoral program | Retired scientist no longer engaged in research | Other

Q. Which of the following best describes your field of study?

General: [list of 20 fields]

Q. Gender:

Male | Female | Other | Prefer not to say

Q. Number of dependents of any age you care for:

0 | 1 | 2 | 3 or more | Prefer not to say

Q. In what age group(s) are your dependents? Note. You may select multiple

0-2 years old | 3-5 years old | 6-11 years old | 12-18 years old | 18-65 years old | Over 65 years old

Q. How would you rate your institution's response to the pandemic?

Terrible | Poor | Average | Good | Excellent

Q. Previously in 2018 and 2019, approximately how much research funding did you oversee or manage per year? Note: Ignore overhead or indirect costs; focusing only on funds used directly for research regardless of source (e.g., home institution, federal grant).

Approximately:

<10,000 $/year | 10,000-20,000 $/year | 20,000-50,000 $/year | 50,000-100,000 $/year | 100,000-200,000 $/year | 200,000-500,000 $/year | 500,000-1,000,000 $/year | 1,000,000-2,000,000 $/year | >2,000,000 $/year

For the following, please consider your "research publications" as all of your publications that focus on a research question. (e.g., journal articles, conference proceedings, patents, books. Ignore commentary, editorials, etc.)

Q. Assuming this pandemic lasts for another [1, 2, 3, 4, 6, 8] *months (until* [May, June, July, August, October, December]*), how do you think the quantity and impact of your research publications will change in 2020 and 2021 compared to 2018 and 2019? (i.e., what will be the percent change?)*

Quantity (i.e., number): [slider scale from -100% to +50% in 10% increments]

Impact (i.e., quality influence): [slider scale from -100% to +50% in 10% increments]

Q. Assuming this pandemic lasts for another [1, 2, 3, 4, 6, 8] *months (until* [May, June, July, August, October, December]*), how much time do you expect it to take to resume your normal pace of work?*

Approximately:

No time needed | About a week | 2 to 3 weeks | 3 to 5 weeks | 1 to 2 months | 3 to 5 months | 6 months or more

Q. Assuming this pandemic lasts for another [1, 2, 3, 4, 6, 8] *months (until* [May, June, July, August, October, December]*), how much funding do you expect it will take to resume your normal pace of work?*

Approximately:

No funds needed | Less than $1,000 | $1,000-$5,000 | $5,000-$10,000 | $10,000-$50,000 | $50,000-$100,000 | $100,000-$500,000 | More than $500,000

Q. Besides what you have just reported, have any of the following happened to you and your work because of the coronavirus pandemic, social distancing, or related events?

I lost important opportunities to disseminate research | A research project has been disrupted to the point that it will likely never be completed as intended | My income will be lower than expected | I am concerned about the security of my position

To fix quantities, scientists reporting values above the highest option were set along the following rules: > $2M became $3M, >6 mo. became 9 mo., and > $0.5M became $0.75M. These values are approximately what the next "largest" choice would have been if we had continued to use the same scale in response options and included an additional (bounded) choice.

While we allowed for twenty broad fields for respondents to choose from, the analyses that compare fields are based on aggregating those twenty fields into five broader categories (that contain roughly equivalent response numbers): Biology or Chemistry or Related, Health or Medicine or Related, Mathematics or Statistics or Economics or Related, Physics or Engineering or Environmental Sciences or Related, Sociology or Psychology or Other Social Sciences.

S2.2 Truth-telling incentive

A notorious challenge of collecting data via self-reported surveys is assessing the validity of respondents' responses. While secondary data sources can be used to validate responses in some settings, our solicitation of forecasts obviously prevents any ex-ante assessment of validity. However, a number of mechanisms have been developed that encourage respondents to report truthfully (or honestly, or to exert effort in their responses) by using incentive schemes where truthful reporting by all respondents is the equilibrium outcome that maximizes each respondent's payoff (e.g., probability they receive a prize). We choose to implement the Honesty Via Choice Matching (HVCM) mechanism, which stems from the "Bayesian Truth Serum" line of work.[v]

This was communicated to respondents with the following text in the consent script:

Compensation - This survey rewards your accuracy. After some questions, we ask you to predict the responses of scientists similar to you. If your predictions score in the top 10% of accuracy, you will be entered into a lottery to win a $100 gift card.

And it was also stated clearly in the introduction page of the survey (immediately succeeding the consent):

This survey rewards your accuracy. After some questions, we ask you to predict the responses of other scientists—specifically, other scientists who are in your field and are about your age. Your input will have a greater impact the more accurate it is. If your predictions score in the top 10% of accuracy, you will be entered into a lottery to win a $100 gift card.

Like many of these sorts of mechanisms, HVCM requires respondents to answer an auxiliary question after a focal question that, essentially, asks the respondent to predict how all other respondents will respond. The payment rule associated with HVCM (and many other mechanisms) incorporates respondents' accuracy in these predictions, rewarding those who are better able to predict others' responses. In the HVCM mechanism, respondents who answer the focal question in the same way actually influence each other's payoffs, which theoretically leads to (or, practically, at least nudges respondents closer to) an equilibrium where all respondents report truthfully.

We use this auxiliary question on all five of the research-disruption questions. Based on pilot tests of this mechanism, our use of the auxiliary predication question asks respondents to predict where they lie within the distribution only of individuals who are in the same broad age group and the same broad field of study. We make one slight modification of the standard HVCM mechanism to accommodate the fact that our five research disruption questions are all continuous in nature. Instead of asking respondents to predict the share of (other) respondents that will choose all of the possible answers to the focal question (which is large because we have discretized a continuous variable in each case), we only ask the respondent to predict where they lie in the distribution of all responses (i.e., "what share of your peers will answer with a number lower than the number you chose?"). Under the assumption that respondents with larger valued responses to the focal question will systematically predict that there is also a larger share of other respondents below them in the distribution, this approach preserves the one-to-one mapping of "types" in the sample and their choices in their response.

The HVCM payment rule is somewhat complicated to the point that the resulting scores are not very intuitive so below we simply report the "prediction error" from the auxiliary questions. This error is simply a scientist's prediction about what percentile their response is at in the distribution of responses from their peers minus their actual location. As shown below, scientists tend to slightly overestimate their location.

For payments based on respondents ultimate HVCM score, we offered $100 Amazon gift cards to a randomly chosen subset of 75 respondents who scored in the top 10% of the full survey sample (which was communicated to participants at the beginning of the survey).

S2.3 Participation incentive

To encourage participation in the survey, the final question of the survey allowed the respondents to choose amongst six approved charities (all with active COVID-19 initiatives) to receive a donation on their behalf. We distributed $15,000 amongst the charities in accordance with the share of votes each charity received.

S3 The sample

S3.1 Distributions and responses

The recruitment for the survey occurred in two broad waves. The first occurred via a single mass e-mailing in mid-April 2020 to all 480,000 addresses. The second occurred via a series of weekly reminder e-mailings to a subset of those 480,000 addresses over ten weeks from early June to mid-August 2020. For reasons described below, all of the responses to the first wave were used as the basis for an earlier study,[vi] and all of the responses to the second wave form the basis for this study.

In the first mailing, out of a total of 480,000 emails sent, approximately 64,000 emails were directly bounced either due to incorrect spelling in the WoS data or the termination of the email account. Overall, 9,968 individuals entered the survey and 8,447 continued past the consent stage. Of those that did not, 412 were not an active scientist, post-doc, or graduate student and thus not within our population of interest, 81 did not consent, and 1,028 did not make any consent choice. When a respondent continued past the consent stage, we asked them to report the type of role they were in. Out of the 8,447 consenting (though not necessarily complete) responses, there were 5,728 responses from faculty or principal investigators (PIs), 1,023 responses from post-doctoral researchers, 701 from graduate students in a doctoral program, and 52 from retired scientists. 551 of the remaining respondents were some other type of position and another 392 did not report their position. Thus, the response rate to this single, first-wave email was roughly 2.0%.

We could not use any of this first-wave data in the analyses we conducted here for two reasons. First, the initial-wave survey instrument used did not include all of the research disruption forecast questions; it lacked the two questions regarding "startup" timing and costs, and we received input from respondents and colleagues that these two metrics would be important to capture. Second, the initial-wave instrument did not contain a theoretically correct implementation of the truth-telling incentives (described in the previous section). With the help of an author who developed these incentives (Drazen Prelec, whose help we appreciate), we were able to identify the flaw in our initial implementation and correct it in the instrument used as the basis of this data.

In the second mailing – which generated the data for this study – a total of 108,345 reminder emails were sent (to only addresses that did not participate in the first wave), of which 20,621 emails were undelivered (e.g., because they bounced). Overall, 7,263 individuals entered the survey and 6,653 continued past the consent stage. Of those that did not, 513 were not an active scientist, post-doc, or graduate student and thus not within our population of interest, 97 did not consent, and 0 did not make any consent choice. When a respondent continued past the consent stage, we asked them to report the type of role they were in. Out of the 6,653 consenting responses, there were 4,187 responses from faculty or principal investigators (PIs), 528 responses from post-doctoral researchers, 294 from graduate students in a doctoral program, 710 from research staff, 336 from retired but active scientists, and 11 from retired and inactive scientists. The response rate to this second-wave of emails was 7.6% for those that reported their position.

Our response rate may reflect the disruptive nature of the pandemic, but it also raises concerns for generalizability of our results. However, after we received feedback from the initial distribution that many individuals had received the email in their "junk" folder, we became concerned with our distribution being automatically flagged as spam. Based on spot-checking of five individuals that we ex-post identified as being randomly selected by our sample, and who we had professional relationships with, found that in four of the five cases the recruitment email had been flagged as spam. We know of no systematic way of estimating the true spam-flagging rate (nor how to avoid these spam filters when using email distributions at this scale) without using high-end, commercial-grade products.

For our analyses here, we focus entirely on the 4,187 responses from the sample of faculty or PIs. Our goal is to focus only on "academic" scientists – hence our reliance on the peer-reviewed publication record for sampling. To exclude industry scientists, we drop twelve respondents who reported working for any kind of private company or firm based on the name of their institution, flagging institutions as "industry" that include prefixes, suffixes, or any abbreviations that are clearly related to a for-profit entity (e.g., "company", "L.L.C.", "Corp", etc.)

We then drop observations that did not answer any of the research disruption questions or did not report a major field of study (983), so the final sample consists of 3,204 responses.

S3.2 Additional summary statistics

As evidence that our allocation of horizon assumptions was randomly presented to respondents, even after selecting this sample, the distribution of assumed future months of the pandemic are as follows: 1 month (15.7%), 2 months (17.2%), 3 months (16.6%), 4 months (16.7%), 6 months (17.0%) 8 months (16.8%).

For these five disruption measures, the averages of the prediction error from the HVCM predictions are as follows (in the same order): 13.8 p.p., 1.7 p.p., 8.2 p.p., 28.6 p.p., and 37.3 p.p., respectively. These numbers reflect the percentage point difference between the scientist's prediction of where they lie within the distribution of responses from scientists who receive the same horizon assumption, are approximately the same age, and are in the same broad field of study. The positive errors for all metrics suggest that scientists tend to overestimate their location in the distribution.

Table S1 reports the summary statistics for the additional covariates we included in the "+Full" control specifications reported in Table 2 in the main table.

Table S1. Summary statistics for additional covariates.

	mean (s.d.)		**mean (s.d.)**
Number of dependents	1.17 (1.06)	Is going to work location {0,1}	0.25 (0.43)
Has dependent(s) aged 6-11 {0,1}	0.21 (0.41)	Total work hours per week, pre-pandemic	44.8 (13.1)
Has dependent(s) aged 12-17 {0,1}	0.20 (0.40)	Total work hours change, post-pandemic	-4.21 (12.1)

Has dependent(s) aged 18-65 {0,1}	0.19 (0.39)	Share time spent teaching, pre-pandemic (p.p.)	20.3 (17.1)
Has dependent(s) aged 65+ {0,1}	0.05 (0.22)	Share time spent teaching, change post-pandemic (p.p.)	1.2 (17.6)
Has a household partner {0,1}	0.84 (0.36)	Share time spent on other work tasks, pre-pandemic (p.p.)	19.8 (18.5)
Is a practicing physician {0,1}	0.09 (0.29)	Share time spent on other work, change post-pandemic (p.p.)	4.6 (15.1)
Institution is closed to non-essential personnel {0,1}	0.71 (0.45)		

Caption: Based on 3,204 scientist-level observations). (p.p.) indicates the variable is a percentage point, with the baseline being set to pre-pandemic levels. {0,1} indicates binary zero-one indicators.

S3.3 Comparison to the Survey of Doctorate Recipients

As with any opt-in survey, there may be correlations between which scientist choose to opt-in and their experiences about which they want to report. For example, scientists who felt strongly about sharing their situation, whether they experienced large positive or negative changes, may be more likely to respond, which would increase the heterogeneity of the sample. Furthermore, there may also be non-negligible gender differences that arise not due to actual differences in outcomes but due to differences in reporting known to occur across genders.[vii,viii,ix,x]

To estimate the generalizability of our respondent sample, we use the public microdata from The Survey of Doctorate Recipients (SDR) as the best sample estimates of the population of principal investigators in the U.S. The SDR is conducted by the National Center for Science and Engineering Statistics within the National Science Foundation, sampling from individuals who have earned a science, engineering, or health doctorate degree from a U.S. academic institution and are less than 76 years of age. The survey is conducted every two years, and we use the latest data available (2017 cycle). For this comparison, we focus only on university faculty in the SDR and only U.S. respondents in our survey.

First, we compared the distribution of fields of study. Due to variations in field definitions, for the purposes of this comparison we combined the Bio./Chem. fields with the Health/Medicine fields since the SDR field categories do not include many medical-specific fields. For each of the four major groups of disciplines the sample averages and differences are as follows: Biology or Chemistry or Health or Medicine or Related is 40% of our sample and 46% of SDR, Mathematics or Statistics or Economics or Related is 11% of our sample and 13% of SDR, Physics or

Engineering or Environmental Sciences or Related is 21% of our sample and 19% of SDR, and Sociology or Psychology or Other Social Sciences is 28% of our sample and 23% of SDR. All of the differences are within roughly two to six p.p., giving us confidence that our sample is not dramatically skewed away from or towards certain broad fields.

When we compare some of the common demographics in the two samples, we find no substantial differences in average age (47.5 in SDR versus 48.0 in our survey), whether the scientist has a partner at home (84% in SDR versus 85% in our survey), or the region of the U.S. where the scientist resides (Midwest: 22% in SDR versus 23% in our survey; West: 22% in SDR versus 27% in our survey; South: 33% in SDR versus 24% in our survey; Northeast: 22% in SDR versus 25% in our survey).

However, we do have a substantially larger number of female respondents, who account for only about 36% of the SDR sample but 50% of our sample, as well as a larger number of respondents with at least one child under 5 or one child between 6 and 11 years old, who account for 10-12% of the SDR sample but 13-20% of our sample. Since we estimate female scientists and those with these young children report larger disruptions on average, this difference in samples suggests that the overall level of disruptions in our sample may be larger than what the population is experiencing.

S4 Forecasted Disruption Index and regression approach

S4.1 Forecasted Disruption Index

For simplicity, we construct the "Forecasted Disruption Index" (FDI) based on scientists' responses to the five disruption questions. To do so, we first ensure the units are relatively comparable by standardizing the responses to each of 5 questions, sorting the values so that large values indicate what would generally be agreed as "worse" outcomes (e.g., less research funding, fewer publications, lower impact publications, more time to restart, more funding to restart). We then sum these five standardized variables and then again standardize that sum so that the magnitudes of the resulting FDI are interpretable in terms of standard deviations (i.e., a value of one indicates that a scientist's responses altogether are 1 standard deviation worse than average). This implicitly weights relative differences in each of the five metrics equally.

For reference, a 1 s.d. change (a unit increase in the FDI) is roughly equivalent to either a 20% decline in research funding, a 16 or 22% decline in publication quantity or impact, or needing two months longer (or 35% more funding) to recover from the pandemic once it is over.

S4.2 Regression approach, control variable transformations, and Lasso selection

The regression specifications and are discussed in the main paper. The additional covariates summarized above in Table S1 are those that are included in the "+Full" but not the "+Baseline" specifications reported in Table 2 in the main paper. In addition to adding these covariates to the potential control to be used in the Lasso approach (explained below), we transform all of the continuous covariates using both squared and log transformations to better approximate non-linearities in the data. Furthermore, in the specifications where we estimate group-specific coefficients (columns 4-6 in Table 2 in the main paper), we also interact all of the possible control variables with group-specific indicators to allow for relationships to be group-specific.

We estimate all regressions using OLS in Stata © software, using heteroskedasticity robust standard errors in all models. When using the Lasso algorithm as a part of variable selection, we use the `lasso linear` command in Stata © with the default settings associated with the "adaptive" cross-validation selection method in order to first select the relevant controls, and then follow that with a post-Lasso OLS regression.

S5 Additional results: Extension policies

Table S2 presents the results from the regressions described in the section titled "An early look at extension policies" in the main paper. The regression specification is the same as Equation 1 in the main paper, except the dependent variable in columns (1-2) is the institutional rating itself. The results indicate that some of the variation in institution ratings and the FDI is associated with scientists reporting extensions on "deliverables", whereas tenure-clock extensions is not significantly associated with any differences in either institution ratings or the FDI for scientists that are tenure-track but untenured (and are thus at risk of such a policy).

Table S2. Extension policy regression results

	Dependent variable: $InstitutionRating$		Dependent variable: $ForecastedDisruptionIndex$	
	(1)	(2)	(3)	(4)
Extension, deliverables	0.094***	0.092***	-0.092***	-0.088***
	(0.0332)	(0.0323)	(0.0326)	(0.0318)
Extension, tenure clock if tenure-track and untenured	-0.081	-0.058	0.085	0.089
	(0.0631)	(0.0621)	(0.0689)	(0.0677)
Pandemic Horizon	-0.002	-0.004	0.022***	0.029***
	(0.0065)	(0.0064)	(0.0069)	(0.0066)
Control inclusion	+Baseline	+Full	+Baseline	+Full
Control selection	Manual	Lasso	Manual	Lasso

Caption: Reports results from estimating versions of Eq. 1 with two different dependent variables. Based on 3,204 scientist-level observations; heteroskedasticity robust standard errors in parentheses, *p<0.1, **p<0.05, ***p<0.01. "+Baseline" indicates that all of the non-forecast variables in Table 2 in the main paper are included as covariates (using indicator variables as needed) in addition to the horizon control and likewise, "+Full" indicates that all possible covariates, including transformations thereof are included as covariates. "Manual" control selection indicates that all possible covariates are in the regression model, while "Lasso" selection relies on the output of the Lasso routine and performs a post-Lasso regression using only the selected covariates.

S6 References for Supplementary information

[i] Birkle, C., Pendlebury, D. A., Schnell, J. & Adams, J. Web of Science as a data source for research on scientific and scholarly activity. *Quantitative Science Studies* 1, 363–376 (2020).

[ii] Roser, M., Ritchie, H., Ortiz-Ospina, E. & Hasell, J. Coronavirus (COVID-19) deaths. *Our World in Data* https://ourworldindata.org/covid-deaths.

[iii] Ritchie, H., Roser, M., Ortiz-Ospina, E. & Hasell, J. Policy responses to the coronavirus pandemic. *Our World in Data* https://ourworldindata.org/policy-responses-covid.

[iv] Powell, K. Science-ing from home. *Nature* 580, 419–421 (2020).

[v] Cvitanić, J., Prelec, D., Riley, B., & Tereick, B. “Honesty via choice-matching.” *American Economic Review: Insights*, 1.2 (2019): 179-192.

[vi] Myers, K.R., et al. "Unequal effects of the COVID-19 pandemic on scientists." *Nature Human Behaviour*, 4.9 (2020): 880-883.

[vii] Lundeberg, M. A., Fox, P. W. & Punćcohaŕ, J. Highly confident but wrong: gender differences and similarities in confidence judgments. *Journal of Educational Psychology* 86, 114–121 (1994).

[viii] Bollinger, C. R., Hirsch, B. T., Hokayem, C. M. & Ziliak, J. P. Trouble in the tails? What we know about earnings nonresponse 30 years after Lillard, Smith, and Welch. *Journal of Political Economy* 127, 2143–2185 (2019).

[ix] Bound, J., Brown, C., & Mathiowetz, N. Measurement error in survey data. In *Handbook of econometrics* (Vol. 5, pp. 3705-3843). Elsevier (2001).

[x] Kim, C. & Tamborini, C. R. Response error in earnings: an analysis of the survey of income and program participation matched with administrative data. *Sociological Methods & Research* 43, 39–72 (2014).