Reviewers' comments:

Reviewer #1 (Remarks to the Author):

This article studies the role of informal mentorship on the future career of mentees. It analyzes a large set of mentor-mentee relationships across several disciplines and through a long time-frame. The authors analyze the effect of a mentor’s success on the mentee’s success. The authors also analyze the effect of gender homophily on the success of mentees. They find that more female mentor leads to a decrease in the success of female mentees. They implications for science policy.

Mentorship is an unstudied aspect of scientific research. While previous research has studied formal mentorship, this manuscript expands this previous research by studying informal mentorship. Thus, this is an important addition to the existing literature and I think deserves to continue be explored. I appreciate the scale of the analysis involving millions of publications and authors, which is necessary to capture subtle effects that mentorship might have.

I think however that the manuscript contains a number of major shortcomings. They use MAG which is known to have many problems with author disambiguation and tracking of citations. Also, they use co-authorship as synonymous of mentorship which is not well justified as there are many more reasons to be a co-author than to be a mentor. Finally, the conclusion that gender homophily in mentor-mentee relationships has negative effects for females ignores the historical aspects of this relationship as men have enjoyed significant advantages and access to resources for their mentees. In my view, there are societal aspects in the data that cannot be ignored no matter how clever the matching method is for doing causal inference on observational data. For these reasons, I think this paper needs major clarifications and revisions, the least of which is to tone down the claim that they are analyzing "mentorship" to something more accurate such as co-authorship.

To summarize, my top four problems with the paper are:

1) MAG data quality: the authors presumably used the author ID provided by MAG. In my own experience and also in previous research, it has been shown that there are significant challenges with this dataset, specifically with authors having multiple identifiers. Discussions of the source data quality should be expanded significantly, especially when compared to other data sources such as web of science, google scholar, and scopus. Also, estimation of gender based on names is significantly challenging and should be discussed much more extensively.

2) Selection of protégé and mentor: this is based on shared authorship in papers. there are many reasons why co-authorship occurs and I would think that only a small portion of times this is due to “mentoring”. The manuscript therefore is not about mentorship but rather co-authorship.

3) How would the analysis of historical data be related to the finding that opposite gender mentorship is better for females? Historically, for example, man had more privileges in universities and therefore

could give more support.

4) Selection of junior and senior researchers: The authors used the years since first publication as a measure of seniority but provide no justification for such selection. They choose 7 years as a threshold. Previous research have used other methods such as number of years since Ph.D. award. Significant justification of this choice is needed.

Reviewer #2 (Remarks to the Author):

The manuscript "The Impact of Informal Mentorship in Academic Collaborations," reports the findings of a study on the relationship between "informal mentorship, provided by senior coauthors to junior coauthors, and the protégé's future citation impact. The authors claim that when there is a junior author (7 years from PhD year) and a senior author on a team, the senior authors offer "informal mentorship" to junior authors during the research process. They find that junior coauthors' citations are positively correlated with senior coauthor collaborations. They also studied the relationship between the mentors' gender and their future performance. I like the large-scale aspect of the study, the research question (there is not enough work on mentorship in science), and the methods.

However, I have concerns about the research and the author's conclusions.

1. Labeling the phenomena under study as "informal mentorship" is unjustified and misleading. A senior member of a team does not autmatically or typically provide mentorship. In many teams, senior authors have their name on a paper because they are offering their name/reputation, funds, or salary (if they are paying a post doc) – none of which involves mentorship. It is also common for junior and senior faculty to work together and for neither to provide mentorship to the other. They work in parallel, having knowledge and skills about the research topic, writing, dealing with reviewers and the submission process. So, something might be going on but the claim that it is mentorship is unsubstantiated and in many cases nonsensical.

2. Contrary to the author's claim, it is common for junior faculty to offer the "informal mentorship" to the senior author. Junior authors are often more in touch with recent literature, hot topics, and new methods than senior authors. Consequently, many senior authors work with junior authors to learn from the junior author, not vice versa. I have often heard senior faculty proclaim their interest in hiring a junior colleague because they want to work with and learn from the junior colleague. Thus, an alternative explanation for the findings is that the superior success of junior faculty could be that they are innately superior and thus are chosen by senior faculty as coauthors vs junior faculty that are not chosen to be coauthors (by preferential attachment). This means when you see greater success of junior faculty after partnering with senior faculty, it may be due to the junior faculty's innate ability and not the senior faculty. (I understand you did matching but the matching is done on too few characteristics of junior faculty to eliminate confounds due to the innate qualities of the mentee.) How do you account for

the above relationship? Can you show that when a weak student coauthors with a Big Shot faculty, they still do as well as more talented students?

3. The measurement of senior and junior authors appears messy. Scientific teams often have multiple senior and multiple junior faculty on the same paper. If there are two senior faculty, should the junior faculty be expected to get a double “informal mentorship” boost? If not, why not? Why not count the other junior members of the team?

4. Some of the junior-senior relationships are in fact between *formal mentors* (e.g., PhD advisor) and mentees. Thus, to label these “formal mentorship” relationships as “informal mentorship” is just plain wrong and confuses the concept of mentorship and informal mentorship.

5. A recent Nature Communication piece, “Early coauthorship with top scientists predicts success in academic careers,” by Li et al. found that when collaborating with top scientists, junior researchers are more likely to be top scientists themselves in the future. How does your work on “Big Shot” senior scientists go beyond their findings? What is the null hypothesis related to big shots? Would your prior be that big shots have no effect? I think the null hypothesis is that big shots have an effect. So are you testing the null hypothesis?

6. There is no direct measurement of the transfer of any distinctive aspect of mentorship from senior authors to junior authors. This raises the issue that the effects are spurious.

The authors would be on more solid ground if they just referred to their analysis as the benefits to junior faculty from working on teams of junior and senior faculty. This would make the analysis closer to Li et al. (2018) and the Guimera (2005) analysis of “newcomers” and “incumbents” but the analysis would still be original in scope and scale and in an emphasis on how the success of junior authors is associated with teamwork between junior and senior faculty. (BTW: This change would require a complete change in title.)

Also, to make the mentorship idea more reasonable, what could you measure to show transfer learning? For example, a paper you cite on the Chaperone effect tries to do this with the publishing in a top journal. Can you show that benefits of the hub are that scholars in the “mentors” hub are now citing the protégé’s papers at a higher rate? That junior publish in new journals and on new topics connect to the research of the senior faculty that coauthor with?

7. The authors inappropriately use causal language. They claim to estimate “the average *causal* effect.” The claim is unwarranted. Observational data and CEM regression methods do not permit claims of causality among variables – just causal inference. Only an experiment can establish causality. CEM provides potentially better inferences (assuming matching is done on the necessary variables and groups are large). Remove all causal language and replace it with the language of correlations, associations, and probabilities. Given the importance the findings might have on changing people’s behavior, conservative interpretations are necessary.

8. I have methodological concerns.

a. The authors poorly explain the variable “number of years post mentorship.” It seems that a junior coauthor could collaborate with a senior coauthor on a paper in 2011 and the change in the junior’s citations is measured over the following five years until year 2015. However, what happens if the same junior coauthor published a paper in 2012 with a different senior author? Then from 2012, 2013, 2014, and 2015, the junior author could be getting additional citations from TWO senior collaborators. I did not see where the author’s adjusted their measurement to account for this change in state. Indeed, the overlap could be very large and have a significant effect on the results depending on how much the junior author publishes per year. It could also be the fact that what looks like a mentor effect in 2015 may be due to a paper in 2011 or a mentor in 2014. There is no sense of explaining this important process at the heart of the paper.

Relatedly, using the average of C5 (<C5>) to calculate the impact of a scientist is also not flawless, since the distribution of C10 of C5 follows a heavy-tailed distribution for an individual scientist (log-normal distribution, to be precise, see Sinatra et al. 2016 Science) (see Stringer et al). Therefore, the average of the C5 could be highly biased by the paper with large C5. In this sense, they should test their main findings by calculating the average of the logarithm of quantity (depends on the distribution of the quantity), rather than taking an average directly. What happens to the results if you use C8 or C10 as is done in other studies (e.g, Wuchty et al.)?

b. Matching criteria do not match on enough dimensions of mentee similarity to be informative. For example, aside from home run papers, citations are a function of productivity. Moreover, productivity is not a matched variable but is critical to understanding citations. This is a flaw that must be corrected.

c. Matching on the year of the protégé’s first publication is done correctly. However, the analogous measure should be measured for mentors, but it is not.

d. Average academic age of mentors. I do not understand the logic of this measure or how it measures what it purports to measure. This is computed for any given protege by first computing the academic age of each mentor in the year of their first publication *with* the protege and then averaging these numbers over all the mentors. Are you measure age or coauthorship tenure?

e. Team size positively correlates with a paper’s citation. Thus, team size should be matched.

f. Number of collaborators in total should be measured, not just the hub size of the mentor because it may very well be highly correlated with the network size of the protégé.

g. The authors have dropped a large amount of data, which could potentially lead to some biased conclusion. For example, the authors dropped all papers related to “Physics” in MAG, the second largest Discipline in MAG with mentor-mentee pairs (see their Supplementary Table S1), just because these papers seem to have a large average number of authors. The dropping of physics papers makes the paper suffer from selection bias. Why, for example, not just drop the physics papers over a certain team size or subfield (e.g., High Energy Physics, which tends to have large teams)? Or, by your analysis for different team sizes, something mentioned previously.

h. The gender effects are interesting but are not well integrated into the paper. They makeup one short paragraph. The analysis does not say anything about gender matching or other methodological issues. Hence, I cannot evaluate the analysis at this point.

In general, a lot more attention should be paid to matching dynamically so that matching is done on a yearly basis. This would ensure that the matched pairs remain similar over a long enough period of time to be able to observe a difference due to working with a senior author versus some unmeasured differences between pairs. Another alternative is to drop the matching and just compare groups of scholars that work with senior authors. Then, see how the proportion of work coauthored with senior scholars changes the rate of citation expected for each paper.

Reviewer #3 (Remarks to the Author):

This is a well done paper. The findings are not terribly surprising, but the analysis is unusually exhaustive. Instead of nit-picking, let me make one suggestion that I think would make a material improvement.

The authors should address the validity and comprehensiveness of their data source, the Microsoft Academic Graph network. Previous products that Microsoft has put out, apparently analogous to Google Scholar, seemed to me to be error ridden and far worse than Google Scholar (which itself is far from perfect of course). Is this one better? I'd like to see some evidence. To be clear, the existence of biases in a data source does not mean that the paper shouldn't be published. On the contrary: clarifying the sources, sizes, and directions of the bias would not only enable the paper's findings to be put on solid ground, but it would open up the data to future researchers to make many new discoveries.

For these reasons, I suggest that the authors do a small study of the validity of their data source focused on identifying the types of biases prevalent in these data, and either conduct sensitivity tests to show the robustness of their findings or, even better, perform bias corrections. Making clear what the biases are would serve as a valuable contribution in and of itself.

Reviewer #4 (Remarks to the Author):

The Impact of Informal Mentorship in Academic Collaborations

This is a well written and interesting paper. The authors study the intersting topic of mentorship in science by looking at informal mentorships as signaled through shared publications. Overall, I really like the work, but I have some practical/methodological questions that I would like to see the authors' responses to before I can recommend publication. Below, I go through these in the order in which they appear in the MS.

(1) The authors use the Microsoft Academic Graph (MAG) as a data source. Many of the other papers on scientific citations I've seen use other data sources (e.g. web of science). Thus I'm curious to understand that dataset a bit better. Does MAG have any biases in terms of topics covered? Is it of equal completeness to WOS? How does this data source impact the results?

(2) In order to get the analysis off the ground the authors make a lot of choices. In my opinion the paper lacks a justification for some of those choices. I also think that the authors need to show robustness of their results.
(2a) Junior period is defined as 7 years, anytime after that is senior period. Why seven years? (One could argue that the junior period is often longer), Why not have a gap? (Since experience of 7.01 years is not likely to be very different from 6.99 years of experience). How sensitive is the analysis to these choices?
(2b) Protege and mentor status is inferred when there is a co-publication AND the two scientists share a discipline AND they belong to the same US based institution. Is it so important that the protege and mentor are located at the same institution? To me it seems plausible that one could benefit from working with a successful mentor from a different institution. Why only US based institutions? And even more importantly, what about multi-author publications? It seems to me that a two-author publication might imply a different kind of mentorship than a 10 author publication. How sensitive is the analysis to these choices?
(2c) Average impact of mentors is used. To me that seems like a potentially problematic choice of impact-measure. We know that citation-success is distributed according to a power-law, thus the average may not reflect the typical quality of mentors (and the median might be more representative). It could, however, also be that it's only your top mentors that count, so perhaps using only citations (success) from the *top mentor* encountered is important? (see below for more on this) How sensitive is the analysis to these choices?

(3) In terms of measuring mentorship outcome, I don't quite understand the the wording "published post mentorship without their mentors". Does this mean that papers must be without more senior authors than the ego? Or is it only papers from the first 7 years? Or that it is without the set of authors that were at some point categorized as mentors? It would be good to elaborate.

(4) The authors use C_5. Is that the measure of citations used throughout?

(5) There is a typo in line 111: "protegpublished"

(6) The authors point out that the big-shot effect is increasing over the time. Could that have to do with overall growth in citations, or do they correct for that in the modeling?

(7) The authors look into the effect of having exactly N female mentors. Again, my hunch is that A) the number of coauthors matters - so a single female mentor on a paper with 8 male mentors is different than a paper with two authors and a single mentor. And B), my intuition is that the fraction of female mentors is a more useful metric than the number.

(8) Just a thought: In the discussion the authors lay out potential reasons for their findings. Personally, I think *visibility* is a big part of what drives the findings here. That if a young scientist publishes with a well known scientist people will notice that famous scientist X has a new paper with protege Y and notice protege Y more. If that hypothesis is correct, using the average citation count of the top mentor only should yield stronger results than average citations of all mentors. And the restriction that a mentor is from the same university shouldn't matter. The key thing is scientific visibility arising from publishing with a famous "name".

Dear Reviewers,

We thank the Editor and the Reviewers for their constructive feedback in guiding our process of revision. Indeed, your comments helped greatly improve and significantly expand our manuscript. Below, we specify our responses (highlighted in blue) to every comment. We also highlight all our modifications in blue in the newly submitted main manuscript and SI.

Reviewer #1:

This article studies the role of informal mentorship on the future career of mentees. It analyzes a large set of mentor-mentee relationships across several disciplines and through a long time-frame. The authors analyze the effect of a mentor's success on the mentee's success. The authors also analyze the effect of gender homophily on the success of mentees. They find that more female mentor leads to a decrease in the success of female mentees. They implications for science policy.

Mentorship is an unstudied aspect of scientific research. While previous research has studied formal mentorship, this manuscript expands this previous research by studying informal mentorship. Thus, this is an important addition to the existing literature and I think deserves to continue to be explored. I appreciate the scale of the analysis involving millions of publications and authors, which is necessary to capture subtle effects that mentorship might have.

I think however that the manuscript contains a number of major shortcomings. They use MAG which is known to have many problems with author disambiguation and tracking of citations. Also, they use co-authorship as synonymous of mentorship which is not well justified as there are many more reasons to be a co-author than to be a mentor. Finally, the conclusion that gender homophily in mentor-mentee relationships has negative effects for females ignores the historical aspects of this relationship as men have enjoyed significant advantages and access to resources for their mentees. In my view, there are societal aspects in the data that cannot be ignored no matter how clever the matching method is for doing causal inference on observational data. For these reasons, I think this paper needs major clarifications and revisions, the least of which is to tone down the claim that they are analyzing "mentorship" to something more accurate such as co-authorship.

Response:

Thank you for the constructive points that you have raised. Our detailed response to each point is provided below.

To summarize, my top four problems with the paper are:

1) MAG data quality: the authors presumably used the author ID provided by MAG. In my own experience and also in previous research, it has been shown that there are significant challenges with this dataset, specifically with authors having multiple identifiers. Discussions of the source data quality should be expanded significantly, especially when compared to other data sources such as web of science, google scholar, and scopus. Also, estimation of gender based on names is significantly challenging and should be discussed much more extensively.

Response:

We originally analyzed an older version of Microsoft Academic Graph (MAG), downloaded in 2016. A few months ago, MAG's research team published multiple papers showing how the latest version improves upon older versions in numerous ways (mainly in regards to the name disambiguation problem). With this in mind, we downloaded the latest version of MAG in March 2020, which increased the number of papers from 130 million to 222 million. Importantly, we re-did the entire analysis from scratch on this new dataset. Note that this is a significant undertaking, as we needed to start with the calculation of the main variables, and carry out the same analyses we have originally proposed, which we expanded in multiple ways that allows us to test the robustness of our findings.

Additionally, we have added a dedicated new section (Supplementary Note 2) to address the concerns regarding MAG and the name disambiguation problem. This section also describes the additional measures that we have taken to alleviate this problem.

As for the reviewer's comment on gender classification, we have added a dedicated new section (Supplementary Note 3) to discuss our methodology more extensively as the reviewer requested.

2) Selection of protégé and mentor: this is based on shared authorship in papers. there are many reasons why co-authorship occurs and I would think that only a small portion of times this is due to "mentoring". The manuscript therefore is not about mentorship but rather co-authorship.

Response:

We are studying the relationship between junior scientists and their senior collaborators, and we describe this relationship as mentorship. Broadly speaking, mentorship can be thought of as a relationship in which a more experienced person helps to guide a less experienced person. Thus, we feel that this definition applies in our case. Nevertheless, we agree that the paper would benefit from further evidence that such relationships indeed involve a form of mentorship. With this in mind, we ran a survey, the description of which is provided in the paragraph below which has been added to the main manuscript, along with two new figures summarizing the outcome (see Figure 1, Supplementary Figure S1, and Supplementary Note 4). **For the reviewer's convenience, we pasted these two figures right after the paragraph below**.

“… we interpret mentorship as the support that juniors receive from their senior collaborators, following its standard definition as `“the activity of giving a younger or less experienced person help and advice over a period of time” [41]. To verify whether the relationship between our identified mentor-protégé pairs involved some form of mentorship, we drew a random sample of 2000 scientists whom we identified as protégé, and manually extracted their emails from publicly available sources such as their personal web pages. We then contacted those scientists and asked them to fill a survey about their experience during their collaboration with one of the scientists whom we identified as their mentors. A detailed description of the survey is provided in Supplementary Note 4. Out of the 2000 scientists, 167 completed our survey; their responses are summarized in Figure 1. More specifically, Figure 1a presents the distribution of the responses to five questions, each asking whether the protégé has received advice from the mentor about a different career-building skill. As can be seen, for each skill, a high percentage of protégés agreed (strongly or otherwise) that they have received advice from the mentor about that skill, with the percentage ranging from 72% to 85% depending on the skill. In contrast, Figure 1b presents the percentage of protégés who agreed (strongly or otherwise) to x out of the 5 skills, where x ranges from 0 to 5. As can be seen, the vast majority of protégés agreed to all 5 skills. Moreover, adding up the percentage for x>0 reveals that 95% agreed (strongly or otherwise) that they have received advice from their mentor regarding at least one skill. Very similar trends were observed when considering only the protégés who stated that the identified mentor was not their thesis advisor nor a member of their thesis committee; see Supplementary Figure S1. Altogether, these findings indicate that the relationship between our identified prot\'eg\'es and mentors indeed involved some form of mentorship.”

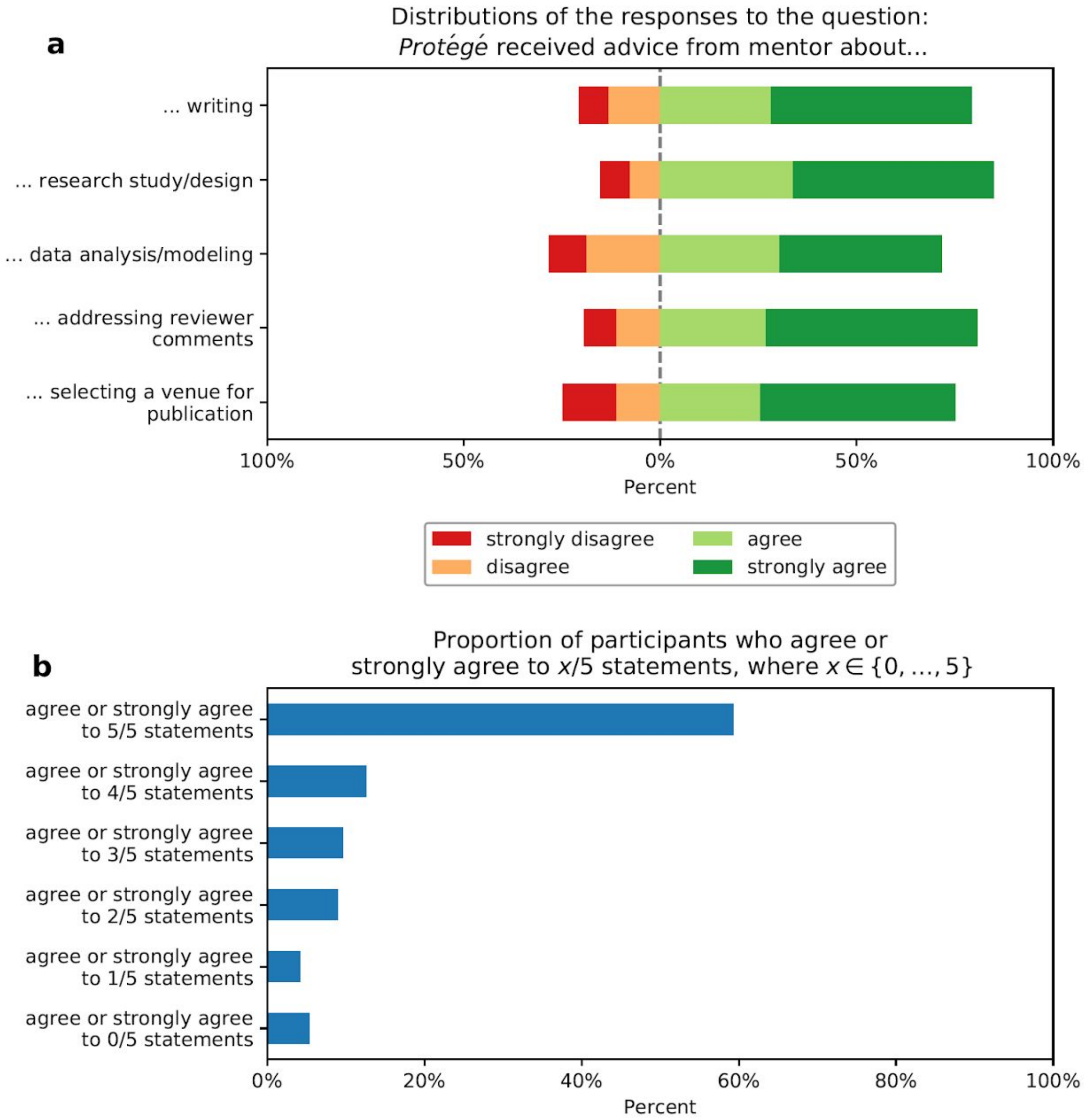
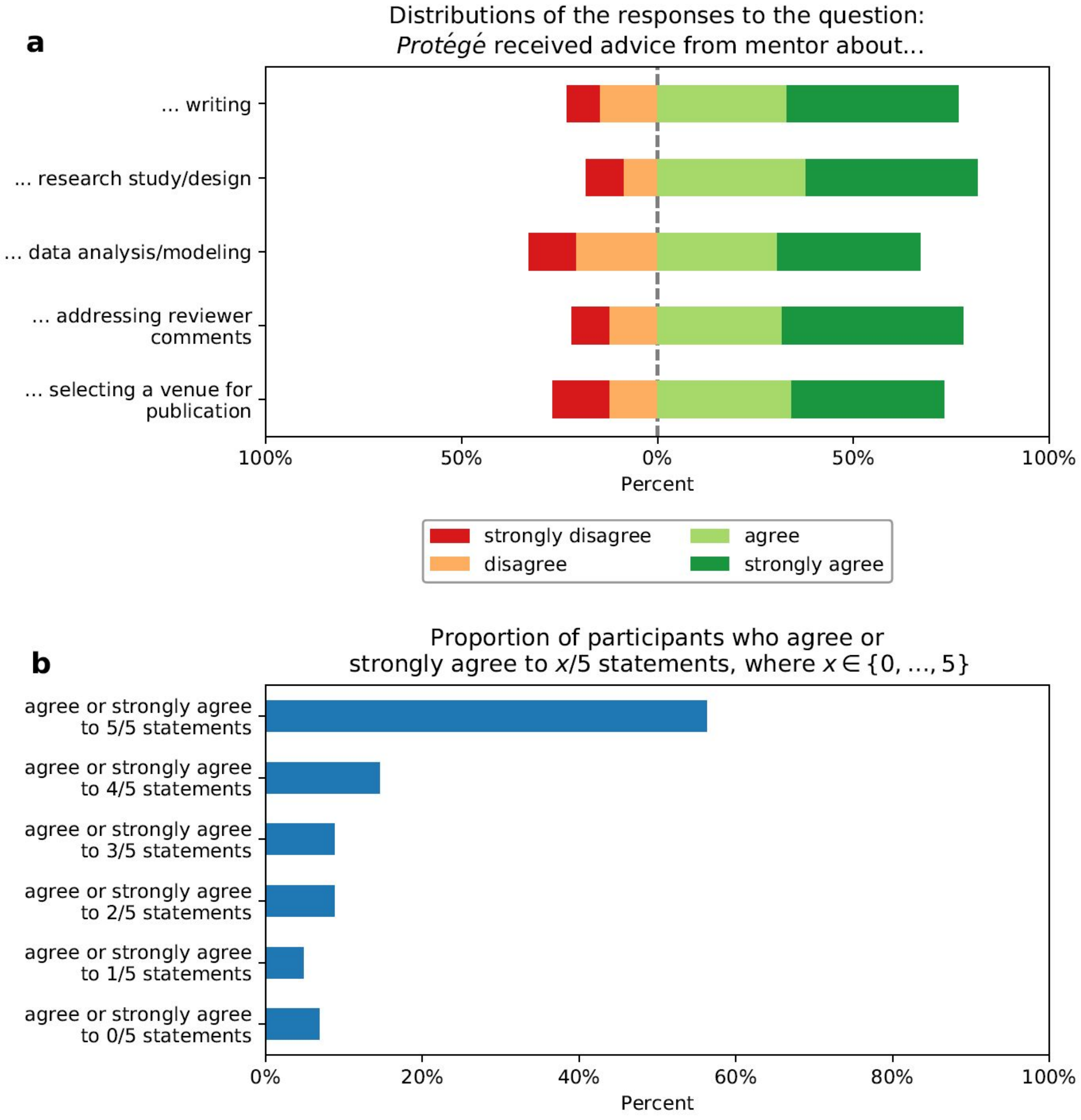


Figure 1: **Survey outcome.** Responses of 167 randomly-chosen scientists who were identified as protégés and asked about their relationship to a scientist who was identified as one of their mentors. **a**, Distributions of the responses to whether or not the protégé agrees with each of five statements regarding their mentors, where the statements take the form "*I received advice from him/her about...*" followed by five different skills: (i) writing; (ii) research study/design; (iii) data analysis/modeling; (iv) addressing reviewer comments; (v) selecting a venue for publication. **b**, Proportion of participants who either agree or strongly agree to x out of the 5 statements regarding their mentors, where $x \in \{0, \ldots, 5\}$.



Supplementary Figure S1: **The same as Figure 1 but for protégés who stated that the identified mentor was not their thesis advisor nor a member of their thesis committee.**

3) How would the analysis of historical data be related to the finding that opposite gender mentorship is better for females? Historically, for example, man had more privileges in universities and therefore could give more support.

Response:

We agree that there are indeed societal aspects that are not captured by our dataset nor by any other paper-citation dataset that we are aware of. Despite this common limitation of large scale observational data, they do provide the ability to analyze hundreds of millions of collaborations, a number that is several orders of magnitude greater than what is possible in traditional controlled experiments. For example, the findings in Figure 3, which provide insights based on carefully constructed comparisons, can be useful even if the precise mechanism is unknown. Nevertheless, we agree that this limitation should be mentioned explicitly in the paper. For this reason, we included the following paragraph:

> *"Our findings also suggest that mentors benefit more from working with male protégés rather than working with comparable female protégés, especially if the mentor is female. These conclusions are all deduced from careful comparisons between protégés who published their first mentored paper in the same discipline, in the same cohort, and at the very same institution. Having said that, it should be noted that there are societal aspects that are not captured by our observational data, and the specific mechanisms behind these findings are yet to be uncovered. One potential explanation could be that, historically, male scientists had enjoyed more privileges and access to resources than their female counterparts, and thus were able to provide more support to their protégés. Alternatively, these findings may be attributed to sorting mechanisms within programs based on the quality of protégés and the gender of mentors."*

4) Selection of junior and senior researchers: The authors used the years since first publication as a measure of seniority but provide no justification for such selection. They choose 7 years as a threshold. Previous research have used other methods such as number of years since Ph.D. award. Significant justification of this choice is needed.

Response:

The survey that we have now added to the paper (which is mentioned in more detail in our response to your second point) provides evidence that our identified pairs of mentor-protégés indeed involved some form of mentorship.

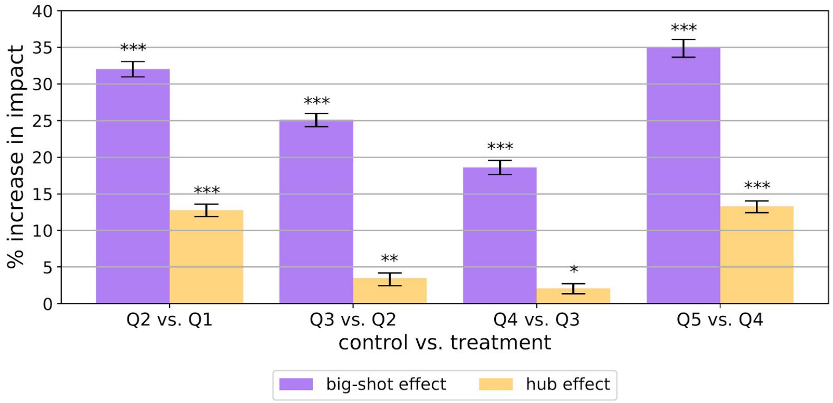
In addition, we have also done a robustness analysis whereby, instead of considering juniors and seniors to be those whose academic age is at most **7** and at least **8**, respectively:

- We considered juniors and seniors to be those whose academic age is at most **6** and at least **9**, respectively (see Supplementary Figure S7);

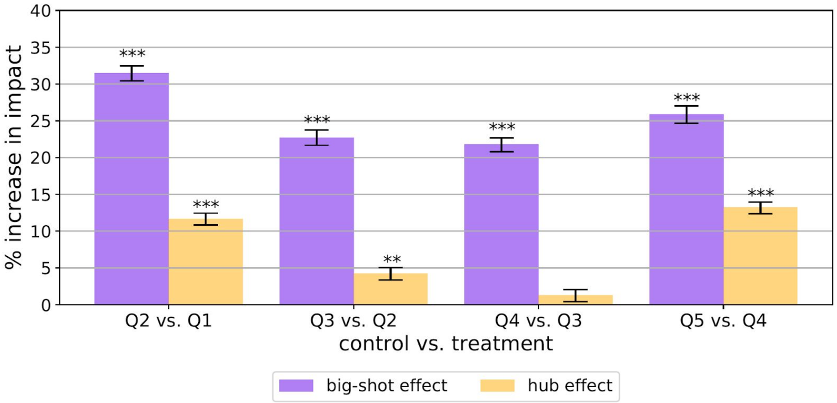
- We considered juniors and seniors to be those whose academic age is at most **5** and at least **10**, respectively (see Supplementary Figure S8).

We found that our main findings regarding the big-shot effect and the hub effect (which are now presented in Figure 2) persist, as shown in Supplementary Figures S7 and S8, as well as Supplementary Tables S14 to S17. These figures are pasted below for your convenience:

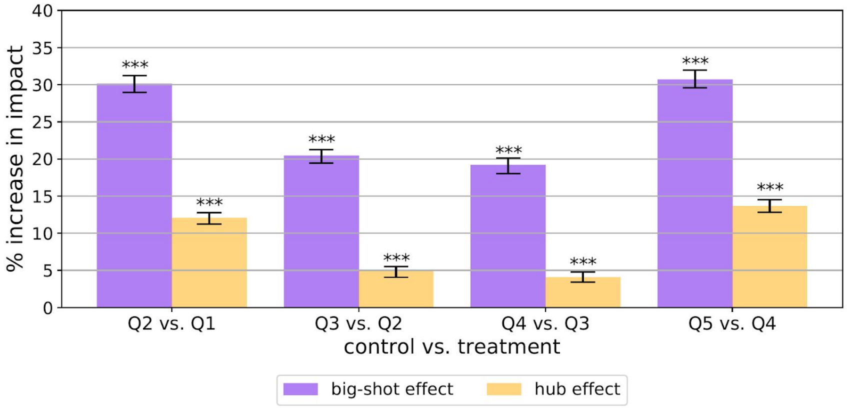
Figure 2:



Supplementary Figure S7:



Supplementary Figure S8:



The paragraph below, which has been added to the main manuscript, describes this robustness analysis, as well as many newly-added robustness analyses:

> *“Supplementary Figures S4 to S8 as well as Supplementary Tables S8 to S17 show similar trends when (i) c_5 with c_{10} as per Sinatra et al. [45]; (ii) computing our measures of mentorship quality using the maximum and median values instead of the average value; (iii) considering juniors and seniors to be those whose academic age is at most 6 and at least 9, respectively; and (iv) considering juniors and seniors to be those whose academic age is at most 5 and at least 10, respectively. Similar trends would also be observed if we replace the average with the sum in our measures of mentorship quality, since we are controlling for the number of mentors; see Supplementary Note 6.1 for more details. These findings imply that the scientific impact of the mentors matters more than their number of collaborators. Consequently, we restrict our attention to the big-shot effect throughout the remainder of our study. Supplementary Figures S9 to S14 as well as Supplementary Tables S18 to S23 suggest that this effect persists regardless of the discipline, the affiliation rank, the number of mentors, the average age of the mentors, the protégé's gender, and the protégé's first year of publication.”*

Reviewer #2:

The manuscript "The Impact of Informal Mentorship in Academic Collaborations," reports the findings of a study on the relationship between "informal mentorship, provided by senior coauthors to junior coauthors, and the protégé's future citation impact. The authors claim that when there is a junior author (7 years from PhD year) and a senior author on a team, the senior authors offer "informal mentorship" to junior authors during the research process. They find that junior co-authors' citations are positively correlated with senior co-author collaborations. They also studied the relationship between the mentors' gender and their future performance. I like the large-scale aspect of the study, the research question (there is not enough work on mentorship in science), and the methods.

However, I have concerns about the research and the author's conclusions.

1. Labeling the phenomena under study as "informal mentorship" is unjustified and misleading. A senior member of a team does not automatically or typically provide mentorship. In many teams, senior authors have their name on a paper because they are offering their name/reputation, funds, or salary (if they are paying a post doc) – none of which involves mentorship. It is also common for junior and senior faculty to work together and for neither to provide mentorship to the other. They work in parallel, having knowledge and skills about the research topic, writing, dealing with reviewers and the submission process. So, something might be going on but the claim that it is mentorship is unsubstantiated and in many cases nonsensical.

Response:

We are studying the relationship between junior scientists and their senior collaborators, and we describe this relationship as mentorship. Broadly speaking, mentorship can be thought of as a relationship in which a more experienced person helps to guide a less experienced person. Thus, we feel that this definition applies in our case. Nevertheless, we agree that the paper would benefit from further evidence that such relationships indeed involve a form of mentorship. With this in mind, we ran a survey, the description of which is provided in the paragraph below which has been added to the main manuscript, along with two new figures summarizing the outcome (see Figure 1, Supplementary Figure S1, and Supplementary Note 4). **For the reviewer's convenience, we pasted these two figures right after the paragraph below**.

> *"To verify whether the relationship between our identified mentor-protégé pairs involved some form of mentorship, we drew a random sample of 2000 scientists whom we identified as protégé, and manually extracted their emails from publicly available sources such as their personal web pages. We then contacted those scientists and asked them to fill a survey about their experience during their collaboration with one of the scientists whom we identified as their mentors. A detailed description of the survey*

is provided in Supplementary Note 4. Out of the 2000 scientists, 167 completed our survey; their responses are summarized in Figure 1. More specifically, Figure 1a presents the distribution of the responses to five questions, each asking whether the protégé has received advice from the mentor about a different career-building skill. As can be seen, for each skill, a high percentage of protégés agreed (strongly or otherwise) that they have received advice from the mentor about that skill, with the percentage ranging from 72% to 85% depending on the skill. In contrast, Figure 1b presents the percentage of protégés who agreed (strongly or otherwise) to x out of the 5 skills, where x ranges from 0 to 5. As can be seen, the vast majority of protégés agreed to all 5 skills. Moreover, adding up the percentage for x>0 reveals that 95% agreed (strongly or otherwise) that they have received advice from their mentor regarding at least one skill. Very similar trends were observed when considering only the protégés who stated that the identified mentor was not their thesis advisor nor a member of their thesis committee; see Supplementary Figure S1. Altogether, these findings indicate that the relationship between our identified prot\'eg\'es and mentors indeed involved some form of mentorship."

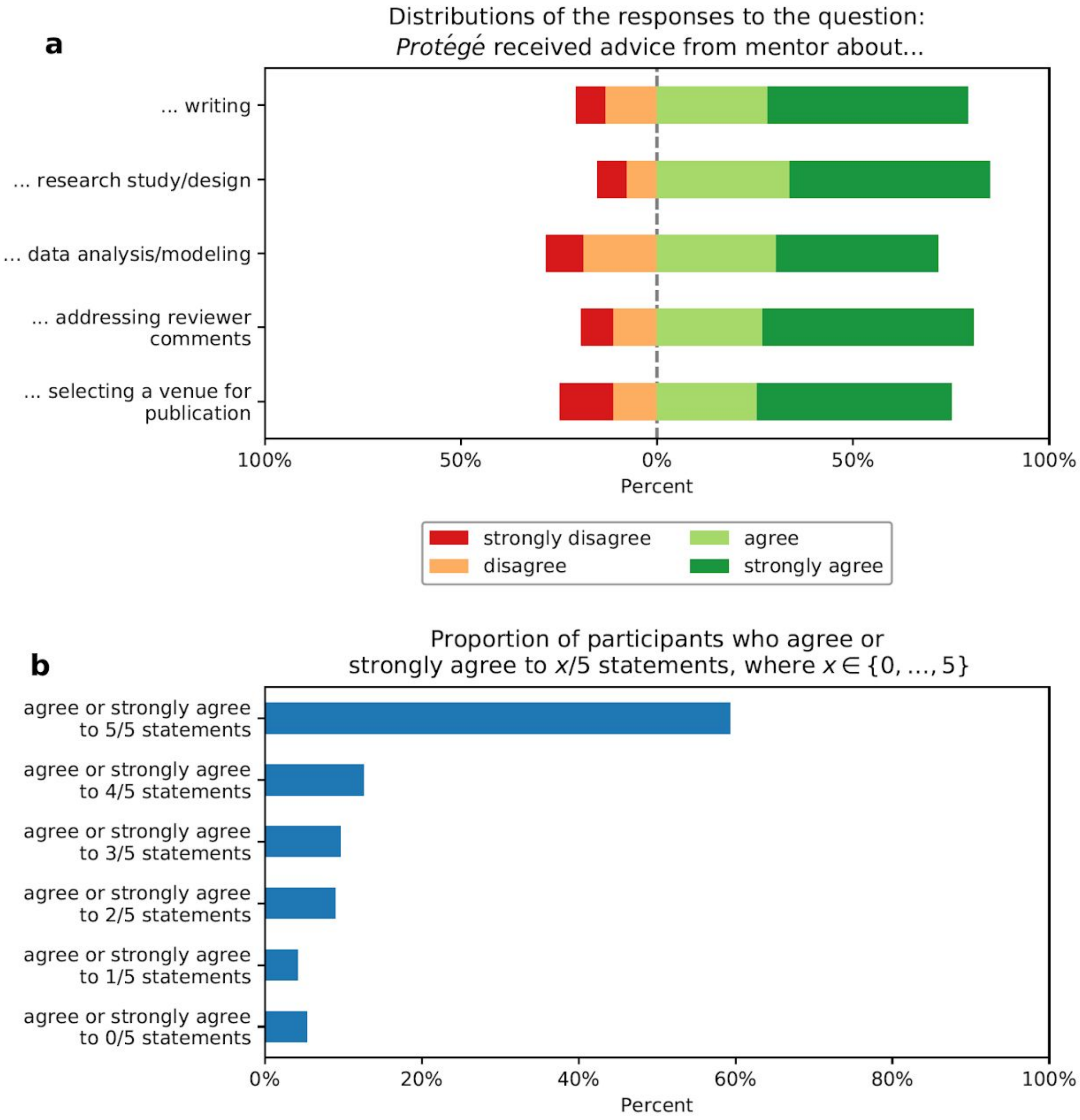
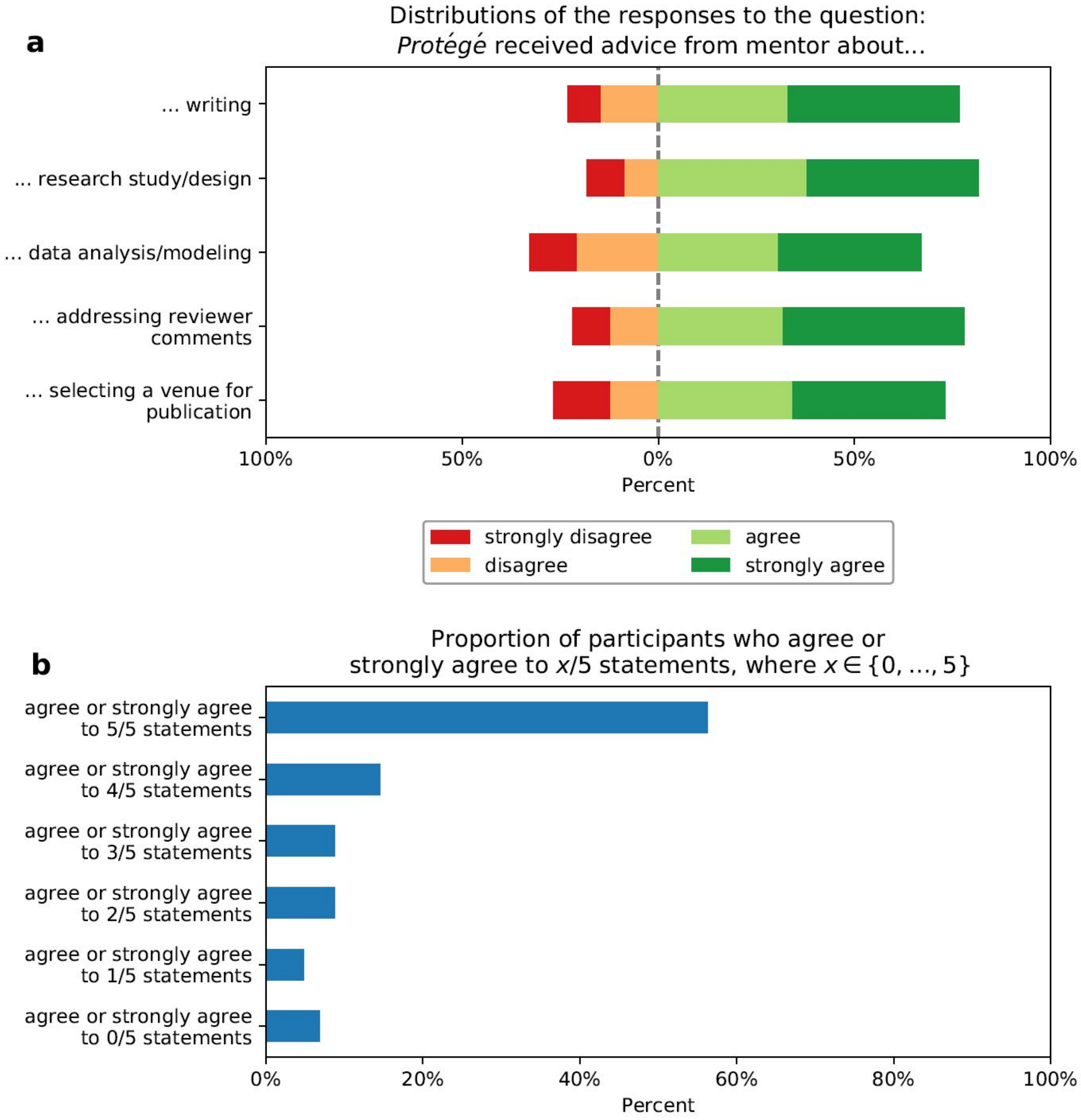


Figure 1: **Survey outcome.** Responses of 167 randomly-chosen scientists who were identified as protégés and asked about their relationship to a scientist who was identified as one of their mentors. **a**, Distributions of the responses to whether or not the protégé agrees with each of five statements regarding their mentors, where the statements take the form "*I received advice from him/her about...*" followed by five different skills: (i) writing; (ii) research study/design; (iii) data analysis/modeling; (iv) addressing reviewer comments; (v) selecting a venue for publication. **b**, Proportion of participants who either agree or strongly agree to x out of the 5 statements regarding their mentors, where $x \in \{0, \ldots, 5\}$.



Supplementary Figure S1: **The same as Figure 1 but for protégés who stated that the identified mentor was not their thesis advisor nor a member of their thesis committee.**

2. Contrary to the author’s claim, it is common for junior faculty to offer the “informal mentorship” to the senior author. Junior authors are often more in touch with recent literature, hot topics, and new methods than senior authors. Consequently, many senior authors work with junior authors to learn from the junior author, not vice versa. I have often heard senior faculty proclaim their interest in hiring a junior colleague because they want to work with and learn from the junior colleague. Thus, an alternative explanation for the findings is that the superior success of junior faculty could be that they are innately superior and thus are chosen by senior faculty as coauthors vs junior faculty that are not chosen to be coauthors (by preferential attachment). This means when you see greater success of junior faculty after partnering with senior faculty, it may be due to the junior faculty’s innate ability and not the senior faculty. (I understand you did matching but the matching is done on too few characteristics of junior faculty to eliminate confounds due to the innate qualities of the mentee.) How do you account for the above relationship? Can you show that when a weak student coauthors with a Big Shot faculty, they still do as well as more talented students?

Response:

We agree that these are very important points that need to be discussed in the manuscript. As for the possibility of the juniors being the ones who support (rather than receive support from) their senior collaborators, we now acknowledge this possibility and state that we do not consider it in the main manuscript. More specifically, we added the following sentence:

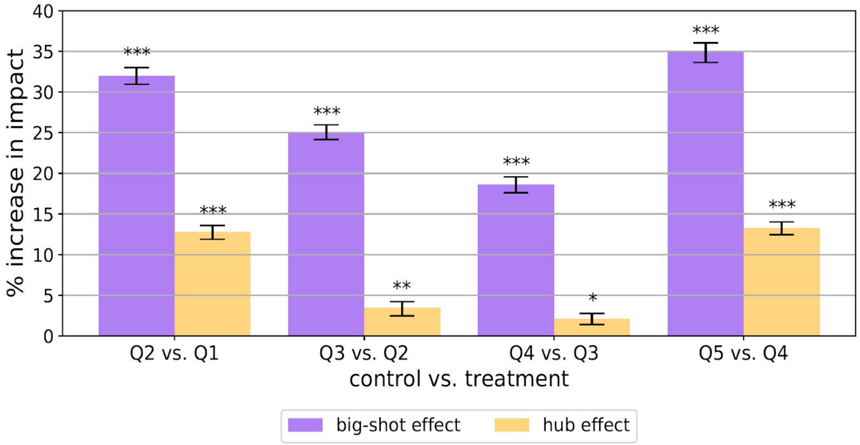
> *While we do acknowledge that it is possible for both juniors and seniors to receive support from their junior collaborators, we interpret mentorship as the support that juniors receive from their senior collaborators, following its standard definition as “the activity of giving a younger or less experienced person help and advice over a period of time” [41].*

As for the possibility that the protégés might be innately superior, while it is very challenging to completely rule out such a possibility in observational studies, we have added additional results to further increase our confidence that the observed effect is indeed attributed to mentorship quality and not innate ability; see Supplementary Note 6.3, Supplementary Figures S2 and S3, and Supplementary Tables S4 to S7. **For the reviewer’s convenience, we have pasted the text of this supplementary note below, followed by the three figures mentioned therein.**

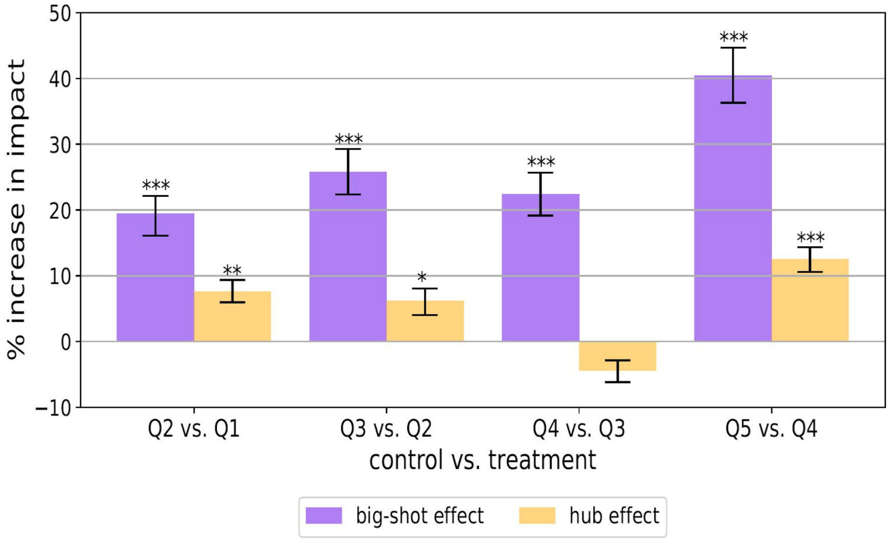
> *As we have seen in Supplementary Note 6.1, our CEM analysis controls for eight different criteria including the affiliation, which is a strong indicator of innate ability. Nevertheless, there might still be room for the protégés and their matches to be different in terms of their innate ability. This, in turn, opens the possibility that the observed differences between the protégés and their matches (in terms of post-mentorship outcome) are attributed to the differences in innate ability rather than mentorship quality. In this case, the explanation would be that the mentors with higher prior impact are more capable of selecting talented protégés, especially since*

these mentors are more likely to be good judges of innate ability in their area of expertise. However, if this is true, then the mentor's prior impact would actually serve as an indicator of the protégé's innate ability. In other words, if two protégés are selected by mentors whose prior impact is similar, then the ability of those two protégés should also be similar. This, in turn, implies that instead of controlling for the protégés' innate ability, it suffices to control for the prior impact of the mentors who selected the protégés. To identify those mentors, we focus on the papers published by the protégés during their first year of publication. We analyze the mentors who coauthored any of these papers, since one of them is likely to have selected the protégé. Here, the rationale is that it is unlikely to assume that protégés are typically selected by mentors who do not collaborate with them during their first year of publication. With this in mind, we reproduced Figure 2 but after controlling for the maximum (Supplementary Figure S2) and the average (Supplementary Figure S3) of all the prior impacts of the mentors who the protégé collaborated with during his/her first year of publication; see Supplementary Tables S4 to S7. As can be seen, both the big-shot effect and the hub effect largely persist, suggesting that the observed differences in performance between the protégés and their matches are attributed to the differences in mentorship quality rather than innate ability.

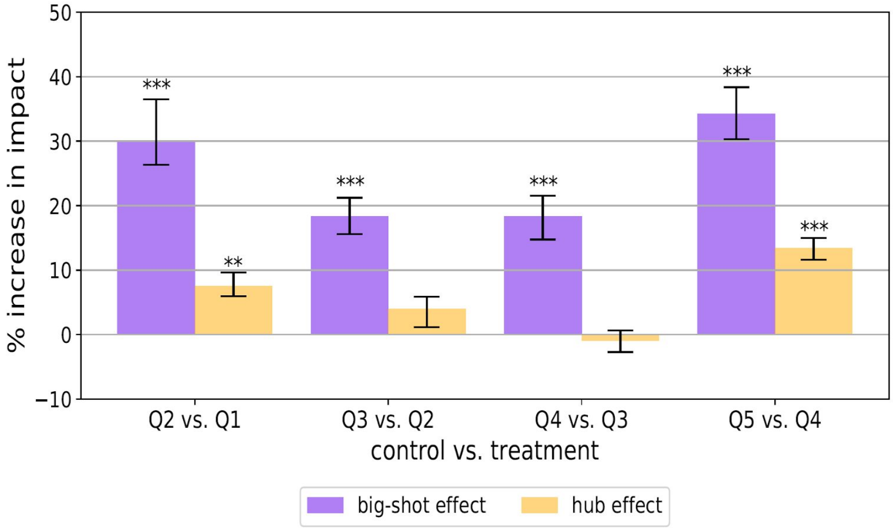
Figure 2:



Supplementary Figure S2



Supplementary Figure S3:



3. The measurement of senior and junior authors appears messy. Scientific teams often have multiple senior and multiple junior faculty on the same paper. If there are two senior faculty, should the junior faculty be expected to get a double "informal mentorship" boost? If not, why not? Why not count the other junior members of the team?

Response:

We agree that all of these important points need to be clarified in the manuscript.

Recall that when measuring the mentorship experience of a protégé, we take the average experience (be it big-shot or hub) over all the mentors. Now, in regards to the potential double boost in mentorship experience that a protégé may receive if they have double the number of mentors, this can be analyzed by replacing the average with the sum. However, since we control for the number of mentors, then by replacing the average with the sum, we would simply be scaling the experience by a constant factor (which is the number of mentors). For example, protégés with, say, 5 mentors are compared to other protégés who also have 5 mentors but with inferior average mentorship experiences. Thus, if we replace the average with the sum, then the mentorship experience of both the protégés and their matches would simply be multiplied by 5, resulting in similar findings. This discussion has been added to Supplementary Note 6.1, and this discussion is now referred to in the main manuscript.

In regards to the number of other juniors that are involved in the protégé's mentorship, we now add Supplementary Note 6.2, to discuss why this (along with many other variables) were deliberately excluded from our analysis.

4. Some of the junior-senior relationships are in fact between *formal mentors* (e.g., PhD advisor) and mentees. Thus, to label these "formal mentorship" relationships as "informal mentorship" is just plain wrong and confuses the concept of mentorship and informal mentorship.

Response:

We agree with this important observation. As a result, we have modified both the main manuscript as well as the Supplementary Materials to talk about mentorship in general, without specifying whether it is formal or informal. Additionally, our newly added survey results show that our identified mentor-protégé pairs involve some form of mentorship regardless of whether the mentor is the formal advisor or not. In fact, the results are remarkably similar when dropping formal mentors from the analysis. As a result, there is no longer a need to restrict our attention to just informal mentorship.

5. A recent Nature Communication piece, “Early coauthorship with top scientists predicts success in academic careers,” by Li et al. found that when collaborating with top scientists, junior researchers are more likely to be top scientists themselves in the future. How does your work on “Big Shot” senior scientists go beyond their findings? What is the null hypothesis related to big shots? Would your prior be that big shots have no effect? I think the null hypothesis is that big shots have an effect. So are you testing the null hypothesis?

Response:

Thank you for pointing out this very recent paper. Our work differs in multiple important ways.

Major differences:

1. Their study focuses on collaborators who are top scientists, i.e., among the 5% most impactful scientists in any given year, regardless of whether they are senior or junior) rather than focusing on collaborators who are mentors. In contrast, we focus on mentor-protégé pairs whose relationship involves some form of mentorship, regardless of whether they are among the top 5%, as we have demonstrated in Figure 1.

2. Their study does not address the fundamental question of whether the social capital of senior collaborators matters more than their impact; we address this limitation by comparing the hub-effect to the big-shot effect as we have demonstrated in Figure 2.

3. The authors do not consider gender in their study. In contrast, our analysis of the gender of both the mentor and the protégé yields important policy implications, as we have demonstrated in Figure 3.

4. In addition to the three main differences mentioned above, there are multiple technical differences, including the fact that we include 9 confounding factors (rather than just 3 confounders as in their study), and the fact that we measure post mentorship impact of the protégé *without their mentors*, to ensure that the observed impact is not attributed to the success of the mentors themselves, but rather to the success of the protégé.

We now cite this paper in the discussion section, and highlight the main differences, to emphasize how our contribution complements the results of this paper.

6. There is no direct measurement of the transfer of any distinctive aspect of mentorship from senior authors to junior authors. This raises the issue that the effects are spurious.

The authors would be on more solid ground if they just referred to their analysis as the benefits to junior faculty from working on teams of junior and senior faculty. This would make the analysis closer to Li et al. (2018) and the Guimera (2005) analysis of “newcomers” and “incumbents” but the analysis would still be original in scope and scale and in an emphasis on how the success of junior authors is associated with teamwork between junior and senior faculty. (BTW: This change would require a complete change in title.)

Also, to make the mentorship idea more reasonable, what could you measure to show transfer learning? For example, a paper you cite on the Chaperone effect tries to do this with the publishing in a top journal. Can you show that benefits of the hub are that scholars in the “mentors” hub are now citing the protégé’s papers at a higher rate? That junior publish in new journals and on new topics connect to the research of the senior faculty that coauthor with?

Response:

We hope that the newly added survey provides a more solid ground for our study of mentorship.

As for the other points raised, we agree that they are interesting, but we feel that the analysis has already been extended in multiple ways (thanks to the other numerous, constructive points raised by the reviewers) that the paper cannot include any more analysis. As a result, we now mention the points that you have raised in a closing remark in the first paragraph of our Discussion section. In particular, we have added the following:

> *“Future research could investigate the mechanisms that underlie this effect, e.g., (i) by comparing mentors who are “newcomers" to those who are ”incumbents" [17], (ii) by analyzing the papers that cite the protégés to see how many of those are authored by the mentors' collaborators, and (iii) by studying the topics that the protégés work on during, and after, the mentorship to understand the skills that are transferred from the mentors to their protégés. These would be welcome extensions to the study, but remain outside of its current scope.”*

7. The authors inappropriately use causal language. They claim to estimate “the average *causal* effect.” The claim is unwarranted. Observational data and CEM regression methods do not permit claims of causality among variables – just causal inference. Only an experiment can establish causality. CEM provides potentially better inferences (assuming matching is done on the necessary variables and groups are large). Remove all causal language and replace it with the language of correlations, associations, and probabilities. Given the importance the findings might have on changing people’s behavior, conservative interpretations are necessary.

Response:

We toned down the causal claims by introducing various edits throughout the manuscript.

8. I have methodological concerns.

a. The authors poorly explain the variable “number of years post mentorship.” It seems that a junior coauthor could collaborate with a senior coauthor on a paper in 2011 and the change in the junior's citations is measured over the following five years until year 2015. However, what happens if the same junior coauthor published a paper in 2012 with a different senior author? Then from 2012, 2013, 2014, and 2015, the junior author could be getting additional citations from TWO senior collaborators. I did not see where the author's adjusted their measurement to account for this change in state. Indeed, the overlap could be very large and have a significant effect on the results depending on how much the junior author publishes per year. It could also be the fact that what looks like a mentor effect in 2015 may be due to a paper in 2011 or a mentor in 2014. There is no sense of explaining this important process at the heart of the paper.

Response:

In the example you mentioned, the paper written in 2012 involves a mentor, which implies that the mentorship as a whole was not yet completed by 2012. Thus, we would not consider the impact of the protégé in 2011, since our outcome measure looks only at the citations of the papers written post mentorship without any of the mentors.

We agree that the way we described the outcome measure may be misunderstood in the way you suggested. To avoid any potential confusion, we changed it from:

> *"We measure this outcome by calculating the average impact of all the papers that the protégé published post mentorship without their mentors. "*

to:

> *"We measure this outcome by calculating the average impact of all the papers that satisfy the following two conditions: (i) they were published when the academic age of the protégé was greater than 7 years; (ii) the authors include the protégé but none of the scientists who were identified as their mentors."*

Relatedly, using the average of C5 (<C5>) to calculate the impact of a scientist is also not flawless, since the distribution of C10 of C5 follows a heavy-tailed distribution for an individual scientist (log-normal distribution, to be precise, see Sinatra et al. 2016 Science) (see Stringer et al). Therefore, the average of the C5 could be highly biased by the paper with large C5. In this sense, they should test their main findings by calculating the average of the logarithm of quantity (depends on the distribution of the quantity), rather than taking an average directly. What happens to the results if you use C8 or C10 as is done in other studies (e.g, Wuchty et al.)?

Response:

We agree that C10 is more widely used in the literature, but we chose C5 which allows us to analyze papers up to 2015 instead of 2010. Having said that, we agree it is important to show that our findings also persist when replacing C5 with C10. This is now done in the paper; see the newly-added Supplementary Figure S4 and Supplementary Tables S8 and S9. As for the log, while we understand the logic behind using it, unfortunately we cannot do so, as log(0) is undefined, and we have scientists with <C5> = 0.

b. Matching criteria do not match on enough dimensions of mentee similarity to be informative. For example, aside from home run papers, citations are a function of productivity. Moreover, productivity is not a matched variable but is critical to understanding citations. This is a flaw that must be corrected.

Response:

Indeed this needs to be clarified. We now address points 8.b, 8.e and 8.f in the newly added Supplementary Note 6.2.

c. Matching on the year of the protégé's first publication is done correctly. However, the analogous measure should be measured for mentors, but it is not.

Response:

Indeed, it is possible to control for the years in which the protégé met their mentors, e.g., if a protégé had three mentors, and collaborated with the first, second, and third mentors on years 1, 5 and 6 of the mentorship, respectively, then we can compare such a protégé to those who had three mentors with whom they collaborated on years 1, 5 and 6. The problem with this approach is the associated combinatorial aspect, e.g., given just 3 mentors, the possible combinations of years in which the protégé could have collaborated with the mentors are: {1,1,1}, …, {1,1,7}, {1,2,2}, ..., {1,2,7}, {1,3,3}, ..., {1,3,7}, {1,4,4}, ..., {1,4,7}, {1,5,5}, …, {1,5,7}, {1,6,6}, {1,6,7}, {1,7,7}, {2,2,2}, …, {2,2,7}, {2,3,3}, ..., {2,3,7}, and so on, all the way to {7,7,7}. Considering all those possibilities in our analysis would dramatically reduce the number of matches, especially for protégés who have a large number of mentors. Thus, for practical limitations, we exclude this from our analysis. We agree that this argument is not immediately obvious to the reader. Thus, we now discuss it explicitly in Supplementary Note 6.2.

d. Average academic age of mentors. I do not understand the logic of this measure or how it measures what it purports to measure. This is computed for any given protege by first computing the academic age of each mentor in the year of their first publication *with* the protege and then averaging these numbers over all the mentors. Are you measure age or coauthorship tenure?

Response:

Indeed, we first compute the academic age of each mentor in the year of their first publication with the protégé and then averaging these numbers over all the mentors. This is meant to reflect the average experience of the mentors during the mentorship. To avoid any potential confusion, we rephrased the following sentence:

> *… the average academic age of their mentors in the year of the first publication with the* protégé.

to

> *... the average academic age of their mentors, which is measured by first computing the academic age of each mentor in the year of their first publication with the protégé, and then averaging these numbers over all the mentors.*

e. Team size positively correlates with a paper's citation. Thus, team size should be matched.

Response:

Indeed this needs to be clarified. We now address points 8.b, 8.e and 8.f in the newly added Supplementary Note 6.2.

f. Number of collaborators in total should be measured, not just the hub size of the mentor because it may very well be highly correlated with the network size of the protégé.

Response:

Indeed this needs to be clarified. We now address points 8.b, 8.e and 8.f in the newly added Supplementary Note 6.2.

g. The authors have dropped a large amount of data, which could potentially lead to some biased conclusion. For example, the authors dropped all papers related to "Physics" in MAG, the second largest Discipline in MAG with mentor-mentee pairs (see their Supplementary Table S1), just because these papers seem to have a large average number of authors. The dropping of physics papers makes the paper suffer from selection bias. Why, for example, not just drop the physics papers over a certain team size or subfield (e.g., High Energy Physics, which tends to have large teams)? Or, by your analysis for different team sizes, something mentioned previously.

Response:

As per your suggestion, we no longer treat Physics differently from the other disciplines. More specifically, our analysis now includes Physics, and we now filter out the papers that have more than 20 coauthors, regardless of the disciplines (this filter is meant to increase the likelihood that the relationship between our identified mentor-protégé pair involve some form of mentorship; the filter works, as we have demonstrated in Figure 1). We obtained similar results compared to the previous analysis when we entirely filtered out Physics.

h. The gender effects are interesting but are not well integrated into the paper. They makeup one short paragraph. The analysis does not say anything about gender matching or other methodological issues. Hence, I cannot evaluate the analysis at this point.

Response:

Please note that, in addition to the two paragraphs related to Figure 3, we also discuss our gender related findings in the Discussion section; see lines 203 to 228 in the manuscript. As for the matching, we have slightly modified the following paragraph to improve the description. We hope that this description is now clear.

> *Next, we turn to a different exploratory analysis where we investigate the post-mentorship impact of protégés while taking into consideration their gender as well as the gender of their mentors. To this end, let F_i denote the set of protégés that have exactly i female mentors. We take the protégés in F_0 as our baseline, and match them to those in F_i for $i \in \{1,2,3,4,5\}$, while controlling for the protégé's average big-shot experience, number of mentors, gender, discipline, affiliation rank, and the year in which they published their first mentored paper. Then, we vary the fraction of female mentors to understand how this affects the protégé. More specifically, for any given $i > 0$, we compute the change in the post-mentorship impact of the protégés in F_i relative to the post-mentorship impact of those in F_0, which we refer to by writing F_i vs. F_0.*

Also note that our gender analysis is only meant to be exploratory, and we have toned down any statements that could have been interpreted as claiming causality.

In general, a lot more attention should be paid to matching dynamically so that matching is done on a yearly basis. This would ensure that the matched pairs remain similar over a long enough period of time to be able to observe a difference due to working with a senior author versus some unmeasured differences between pairs. Another alternative is to drop the matching and just compare groups of scholars that work with senior authors. Then, see how the proportion of work coauthored with senior scholars changes the rate of citation expected for each paper.

Response:

We thank the reviewer for this comment. Matching at a yearly-level would be prohibitive, even with such large-scale data set as MAG at our disposal (see a more detailed discussion of this point in our response to 8.c), which, if attempted, would yield a sample of comparisons that are not reflective of the distribution of mentorship experience in the population. However, yearly matching would be an analytically unwarranted step from a causal inference perspective as well, as it would amount to controlling for a post-treatment covariate. We believe that it is reasonable to assume that a senior-junior collaboration would endogenously alter a protégé’s publication and collaboration trajectory during their junior years. See the exposure of this issue in experimental contexts by Montgomery and colleagues (2018), the lessons of which translate to observational data. Throughout the paper we considered mentorship experience as the treatment, measured based on data about mentors prior to their first encounter with the protégés.

We thank the reviewer for pointing it out that there may be unobserved characteristics that we are not taking into account which may bias our results. While there is always the chance for unobservables, we match on a list of additional characteristics (note, for instance, we match for the exact same institution among the ones in the top 100, for instance, and control for mentors' networks when evaluating the effect of their impact), and have performed robustness analysis to address, to the best of our ability, the unmeasured characteristic of innate ability. We have also toned down the language of causality for improved accuracy, and we hope to the reviewer's satisfaction.

Montgomery, J. M., Nyhan, B. & Torres, M. How conditioning on post-treatment variables can ruin your experiment and what to do about it. *American Journal of Political Science* **62**, 760–775 (2018).

Reviewer #3:

This is a well done paper. The findings are not terribly surprising, but the analysis is unusually exhaustive. Instead of nit-picking, let me make one suggestion that I think would make a material improvement.

The authors should address the validity and comprehensiveness of their data source, the Microsoft Academic Graph network. Previous products that Microsoft has put out, apparently analogous to Google Scholar, seemed to me to be error ridden and far worse than Google Scholar (which itself is far from perfect of course). Is this one better? I'd like to see some evidence. To be clear, the existence of biases in a data source does not mean that the paper shouldn't be published. On the contrary: clarifying the sources, sizes, and directions of the bias would not only enable the paper's findings to be put on solid ground, but it would open up the data to future researchers to make many new discoveries.

Response:

We originally analyzed an older version of Microsoft Academic Graph (MAG), downloaded in 2016. A few months ago, MAG's research team published multiple papers showing how the latest version improves upon older versions in numerous ways (mainly in regards to the name disambiguation problem). With this in mind, we downloaded the latest version of MAG in March 2020 (which increased the number of papers from 130 million to 222 million), and we re-did the entire analysis from scratch on this new dataset. Moreover, we ran additional steps to further reduce the name disambiguation problem (following a process introduced in a paper published in Science). To evaluate the accuracy of this process, we followed the same approach used in the Science paper, and found the error rate to be negligible. Details of this evaluation, as well as a discussion of the literature supporting MAG, can be found in the newly-added Supplementary Note 2.

For these reasons, I suggest that the authors do a small study of the validity of their data source focused on identifying the types of biases prevalent in these data, and either conduct sensitivity tests to show the robustness of their findings or, even better, perform bias corrections. Making clear what the biases are would serve as a valuable contribution in and of itself.

Response:

We agree that additional robustness analysis would improve the manuscript. With this in mind, we now take many additional steps to demonstrate the robustness of our findings; these steps are summarized in the following paragraph which has been added to the main manuscript:

> *"Supplementary Figures S4 to S8 as well as Supplementary Tables S8 to S17 show similar trends when (i) c_5 with c_{10} as per Sinatra et al. [45]; (ii) computing*

our measures of mentorship quality using the maximum and median values instead of the average value; (iii) considering juniors and seniors to be those whose academic age is at most 6 and at least 9, respectively; and (iv) considering juniors and seniors to be those whose academic age is at most 5 and at least 10, respectively. Similar trends would also be observed if we replace the average with the sum in our measures of mentorship quality, since we are controlling for the number of mentors; see Supplementary Note 6.1 for more details. These findings imply that the scientific impact of the mentors matters more than their number of collaborators. Consequently, we restrict our attention to the big-shot effect throughout the remainder of our study. Supplementary Figures S9 to S14 as well as Supplementary Tables S18 to S23 suggest that this effect persists regardless of the discipline, the affiliation rank, the number of mentors, the average age of the mentors, the protégé's gender, and the protégé's first year of publication.”

Reviewer #4:

This is a well written and interesting paper.

(1) The authors use the Microsoft Academic Graph (MAG) as a data source. Many of the other papers on scientific citations I’ve seen use other data sources (e.g. web of science). Thus I'm curious to understand that dataset a bit better. Does MAG have any biases in terms of topics covered? Is it of equal completeness to WOS? How does this data source impact the results?

Response:

Indeed, like other massive datasets on scientific collaborations, MAG may unintentionally introduce some biases. Nevertheless, it provides the ability to analyze hundreds of millions of collaborations. We believe that, compared to other existing alternatives, MAG is as good, if not superior. We have added to Supplementary Note 2 the following sentence, which provides multiple references that testify to the quality of the constantly improving MAG:

> *“Since its relaunch in 2015 into Microsoft Academic Services (2), many independent studies have suggested that the MAG dataset is in many aspects as accurate, if not more, than manually curated data (3–11).”*

To build on that, MAG’s research team published multiple papers very recently (just a few months ago) showing how the latest version improves upon older ones in numerous ways. With this in mind, we downloaded the latest version of MAG in March 2020 (which increased the number of papers from 130 million to 222 million); we re-did the entire analysis from scratch on this new dataset, and found the results to remain robust. Moreover, we took additional steps to further reduce the name disambiguation problem following a process introduced in a paper published recently in Science. To evaluate the accuracy of this process, we followed the same approach used in the Science paper, and found the error rate to be negligible. Details on how MAG improved its dataset, as well as the disambiguation process we introduced can also be found in Supplementary Note 2.

(2) In order to get the analysis off the ground the authors make a lot of choices. In my opinion the paper lacks a justification for some of those choices. I also think that the authors need to show robustness of their results.

Response:

We agree that additional robustness analysis would improve the manuscript. With this in mind, we now take many additional steps to demonstrate the robustness of our findings; these steps are summarized in the following paragraph which has been added to the main manuscript:

> *"Supplementary Figures S4 to S8 as well as Supplementary Tables S8 to S17 show similar trends when (i) c_5 with c_{10} as per Sinatra et al. [45]; (ii) computing our measures of mentorship quality using the maximum and median values instead of the average value; (iii) considering juniors and seniors to be those whose academic age is at most 6 and at least 9, respectively; and (iv) considering juniors and seniors to be those whose academic age is at most 5 and at least 10, respectively. Similar trends would also be observed if we replace the average with the sum in our measures of mentorship quality, since we are controlling for the number of mentors; see Supplementary Note 6.1 for more details. These findings imply that the scientific impact of the mentors matters more than their number of collaborators. Consequently, we restrict our attention to the big-shot effect throughout the remainder of our study. Supplementary Figures S9 to S14 as well as Supplementary Tables S18 to S23 suggest that this effect persists regardless of the discipline, the affiliation rank, the number of mentors, the average age of the mentors, the protégé's gender, and the protégé's first year of publication."*

(2a) Junior period is defined as 7 years, anytime after that is senior period. Why seven years? (One could argue that the junior period is often longer), Why not have a gap? (Since experience of 7.01 years is not likely to be very different from 6.99 years of experience). How sensitive is the analysis to these choices?

Response:

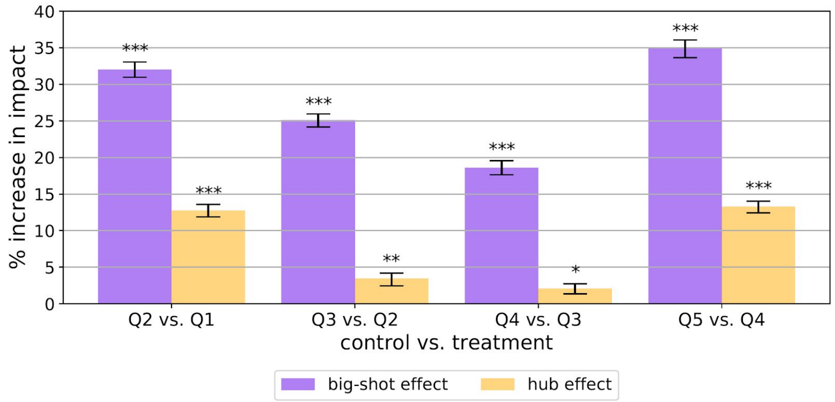
We have done a robustness analysis whereby, instead of considering juniors and seniors to be those whose academic age is at most **7** and at least **8**, respectively:

- We considered juniors and seniors to be those whose academic age is at most **6** and at least **9**, respectively (see Supplementary Figure S7);

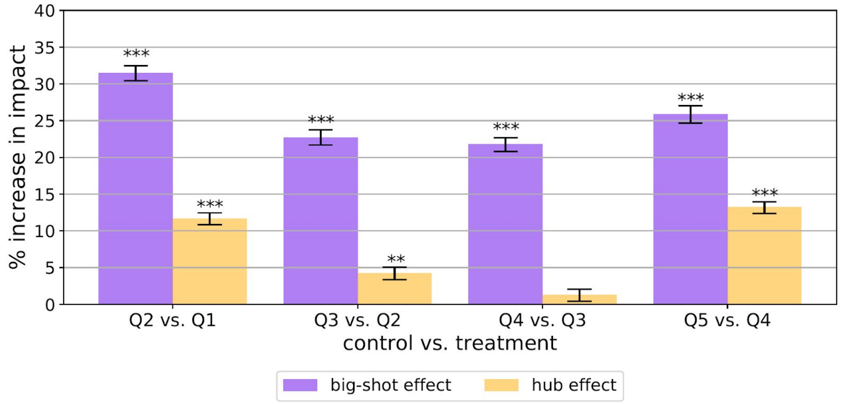
- We considered juniors and seniors to be those whose academic age is at most **5** and at least **10**, respectively (see Supplementary Figure S8).

We found that our main findings regarding the big-shot effect and the hub effect (which are now presented in Figure 2) persist, as shown in Supplementary Figures S7 and S8, as well as Supplementary Tables S14 to S17. These figures are pasted below for your convenience:

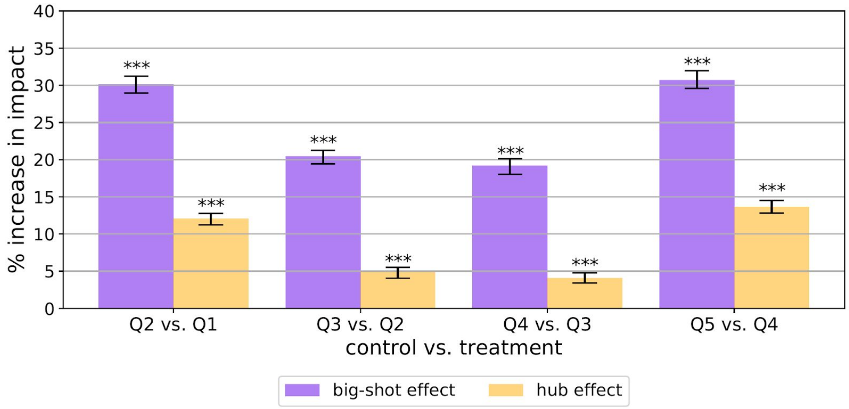
Figure 2:



Supplementary Figure S5:



Supplementary Figure S6:



Broadly speaking, mentorship can be thought of as a relationship in which a more experienced person helps to guide a less experienced person. Thus, we feel that this definition applies in our case. Nevertheless, we agree that the paper would benefit from further evidence that our identified mentor-protégé pairs indeed involve a form of mentorship. With this in mind, we ran a survey, the description of which is provided in the paragraph below which has been added to the main manuscript, along with two new figures summarizing the outcome (see Figure 1, Supplementary Figure S1, and Supplementary Note 4). **For the reviewer's convenience, we pasted these two figures right after the paragraph below**.

> *"… we interpret mentorship as the support that juniors receive from their senior collaborators, following its standard definition as `"the activity of giving a younger or less experienced person help and advice over a period of time" [41]. To verify whether the relationship between our identified mentor-protégé pairs involved some form of mentorship, we drew a random sample of 2000 scientists whom we identified as protégé, and manually extracted their emails from publicly available sources such as their personal web pages. We then contacted those scientists and asked them to fill a survey about their experience during their collaboration with one of the scientists whom we identified as their mentors. A detailed description of the survey is provided in Supplementary Note 4. Out of the 2000 scientists, 167 completed our survey; their responses are summarized in Figure 1. More specifically, Figure 1a presents the distribution of the responses to five questions, each asking whether the protégé has received advice from the mentor about a different career-building skill. As can be seen, for each skill, a high percentage of protégés agreed (strongly or otherwise) that they have received advice from the mentor about that skill, with the percentage ranging from 72% to 85% depending on the skill. In contrast, Figure 1b presents the percentage of protégés who agreed (strongly or otherwise) to x out of the 5 skills, where x ranges from 0 to 5. As can be seen, the vast majority of protégés agreed to all 5 skills. Moreover, adding up the percentage for x>0 reveals that 95% agreed (strongly or otherwise) that they have received advice from their mentor regarding at least one skill. Very similar trends were observed when considering only the protégés who stated that the identified mentor was not their thesis advisor nor a member of their thesis committee; see Supplementary Figure S1. Altogether, these findings indicate that the relationship between our identified prot\'eg\'es and mentors indeed involved some form of mentorship."*

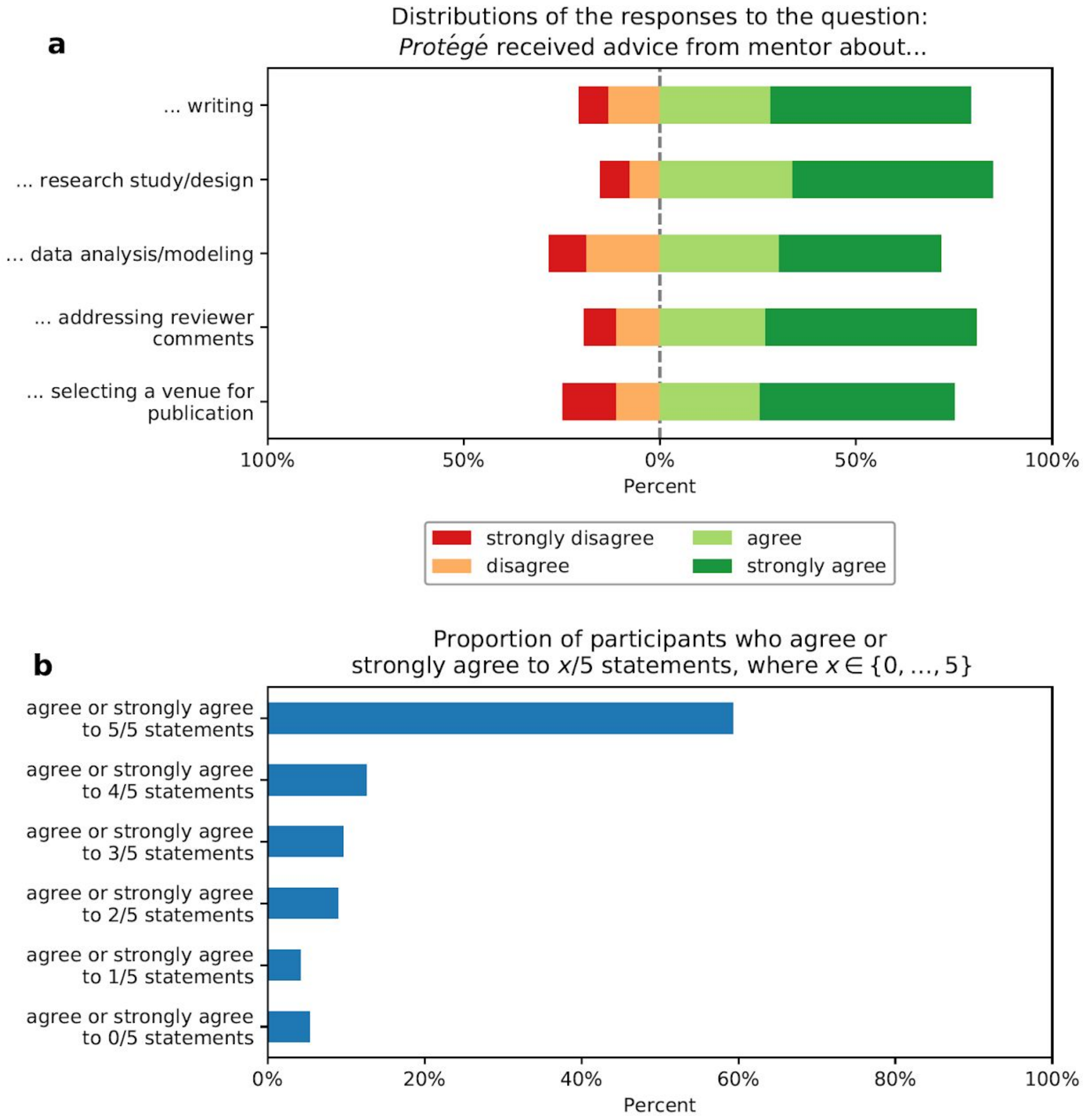
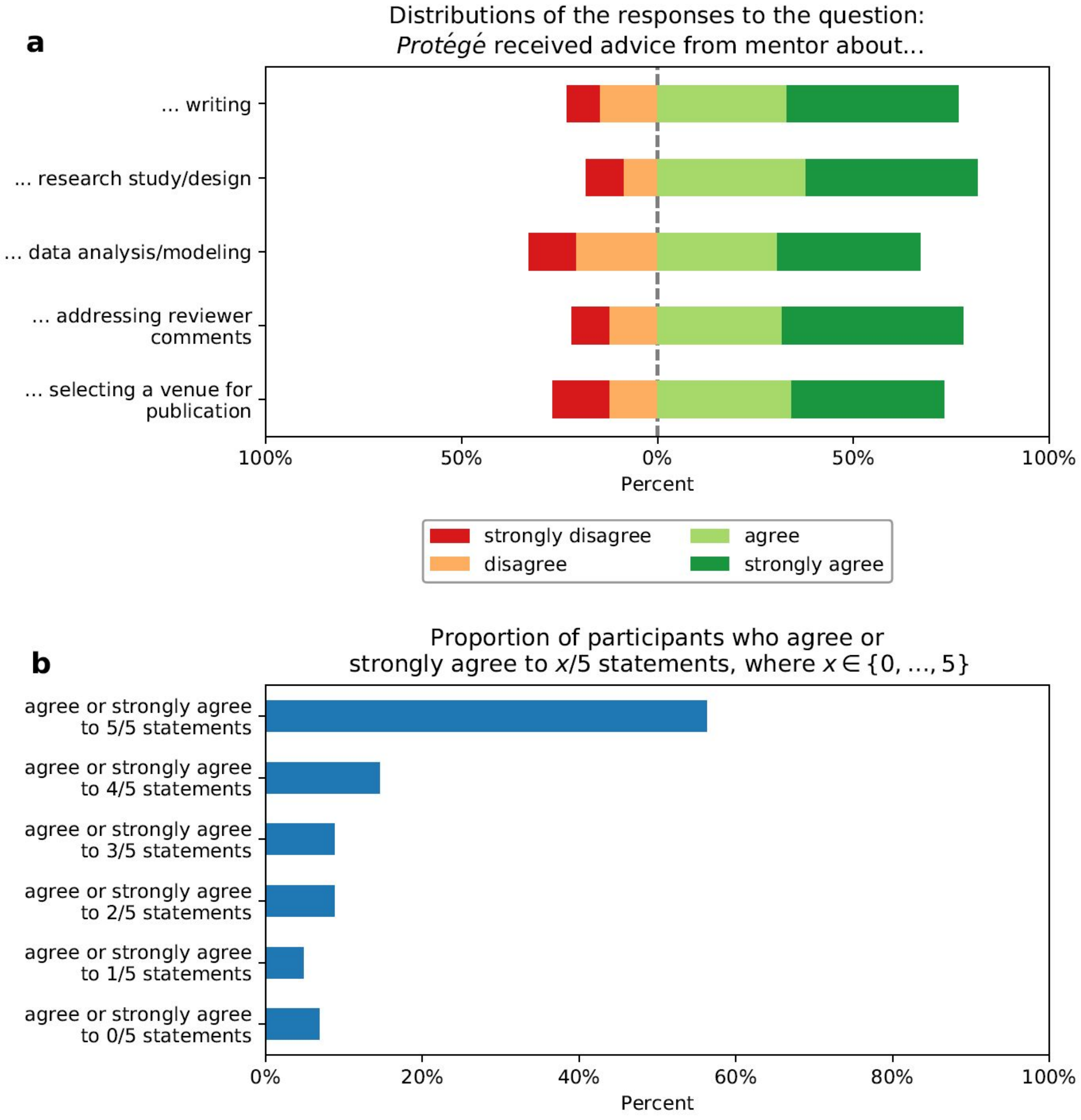


Figure 1: **Survey outcome.** Responses of 167 randomly-chosen scientists who were identified as protégés and asked about their relationship to a scientist who was identified as one of their mentors. **a**, Distributions of the responses to whether or not the protégé agrees with each of five statements regarding their mentors, where the statements take the form "*I received advice from him/her about...*" followed by five different skills: (i) writing; (ii) research study/design; (iii) data analysis/modeling; (iv) addressing reviewer comments; (v) selecting a venue for publication. **b**, Proportion of participants who either agree or strongly agree to x out of the 5 statements regarding their mentors, where $x \in \{0, \ldots, 5\}$.



Supplementary Figure S1: **The same as Figure 1 but for protégés who stated that the identified mentor was not their thesis advisor nor a member of their thesis committee.**

(2b) Protege and mentor status is inferred when there is a co-publication AND the two scientists share a discipline AND they belong to the same US based institution. Is it so important that the protege and mentor are located at the same institution? To me it seems plausible that one could benefit from working with a successful mentor from a different institution. Why only US based institutions? And even more importantly, what about multi-author publications? It seems to me that a two-author publication might imply a different kind of mentorship than a 10 author publication. How sensitive is the analysis to these choices?

Response:

We agree that these are all questions that should be clarified explicitly in the paper.

- In regards to our selection criteria of the mentor-protégé pairs, we now have a lengthy discussion in the newly added Supplementary Note 5 to address these concerns.

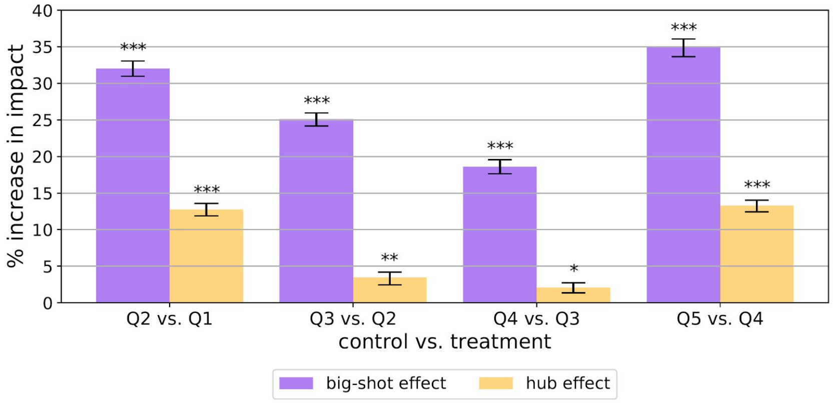
- In regards to considering multi-authored publications, we now address this concern in the newly added Supplementary Note 6.2.

(2c) Average impact of mentors is used. To me that seems like a potentially problematic choice of impact-measure. We know that citation-success is distributed according to a power-law, thus the average may not reflect the typical quality of mentors (and the median might be more representative). It could, however, also be that it's only your top mentors that count, so perhaps using only citations (success) from the *top mentor* encountered is important? (see below for more on this) How sensitive is the analysis to these choices?

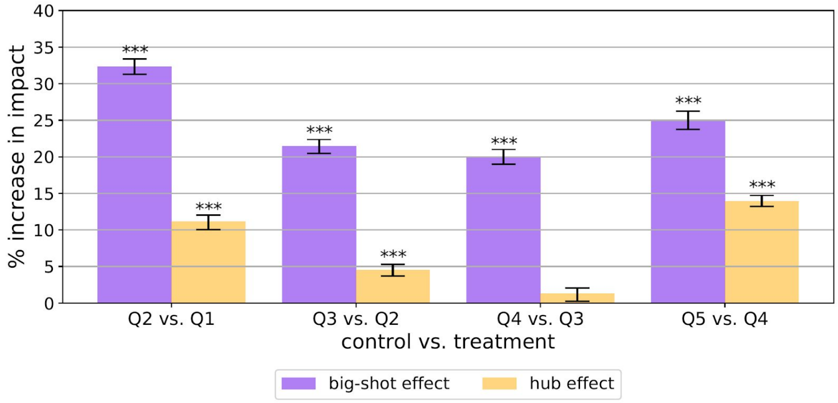
Response:

As part of our newly introduced robustness analysis, we now reproduce Figure 2 using: (i) the impact of the top mentor encountered, and (ii) the median impact of the mentors. The results can be found in Supplementary Figures S5 and S6, respectively, as well as, Supplementary Tables S10 to S13. These figures are pasted below (along with Figure 2) for your convenience:

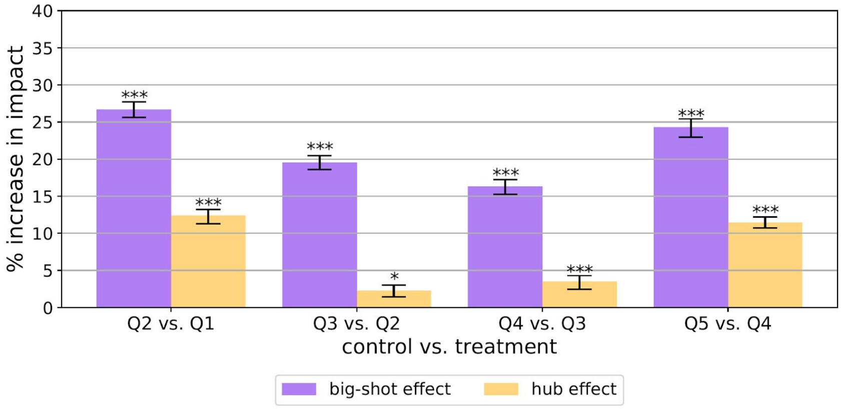
Figure 2:



Supplementary Figure S3:



Supplementary Figure S4:



(3) In terms of measuring mentorship outcome, I don’t quite understand the wording “published post mentorship without their mentors”. Does this mean that papers must be without more senior authors than the ego? Or is it only papers from the first 7 years? Or that it is without the set of authors that were at some point categorized as mentors? It would be good to elaborate.

Response:

We agree that the sentence may be mis-interpreted. To avoid potential confusion, we changed it from:

> *“We measure this outcome by calculating the average impact of all the papers that the protégé published post mentorship without their mentors. “*

to:

> *“We measure this outcome by calculating the average impact of all the papers that satisfy the following two conditions: (i) they were published when the academic age of the protégé was greater than 7 years; (ii) the authors include the protégé but none of the scientists who were identified as their mentors.”*

(4) The authors use C_5. Is that the measure of citations used throughout?

Response:

Yes, this is the measure of publication impact that we use throughout the paper. To avoid any potential confusion, we now state this explicitly in the main paper by writing:

> *The impact of each such paper is calculated as the number of citations that it accumulated 5 years post publication, denoted by c_5 [15]; this is the measure of scientific impact that will be used throughout the article.*

(5) There is a typo in line 111: “protegpublished”

Response:

Thank you for pointing this out. The typo has been fixed.

(6) The authors point out that the big-shot effect is increasing over time. Could that have to do with overall growth in citations, or do they correct for that in the modeling?

Response:

Among the main modifications that we have made in response to the reviewers were:

1. Running a survey to establish that our identified mentor-protégé pairs involved mentorship

2. Performing extensive robustness analysis

Based on the above two modifications, we omitted the figure presenting the big-shot over time, and replaced it with the figure summarizing the survey outcome. This was for two reasons:

1. We felt our analysis of the big-shot effect over time should be part of the robustness analysis; we now only comment on it by writing: “*Note that this effect persists regardless of the discipline, ..., and the protégé's first year of publication*”.

2. We felt that the figure summarizing the survey outcome is much more interesting than the figure presenting the big-shot effect over time.

That being said, in response to your comment, yes we do control for the protégé's first year of publication. We hope that the newly added text on our robustness analysis will help avoid any future confusion; see the newly-added text describing our robustness analysis, which is quoted in our response to your Comment (2).

(7) The authors look into the effect of having exactly N female mentors. Again, my hunch is that A) the number of coauthors matters - so a single female mentor on a paper with 8 male mentors is different than a paper with two authors and a single mentor. And B), my intuition is that the fraction of female mentors is a more useful metric than the number.

Response:

We agree that the fraction of female mentors is a more useful metric, and this is exactly what we plot in Figure 3 (i.e., for every N = 1, …, 5, we vary the fraction of female to male mentors). Perhaps our writing did not make this point explicit. To avoid any potential confusion, we now mention this explicitly in the manuscript, by writing the following:

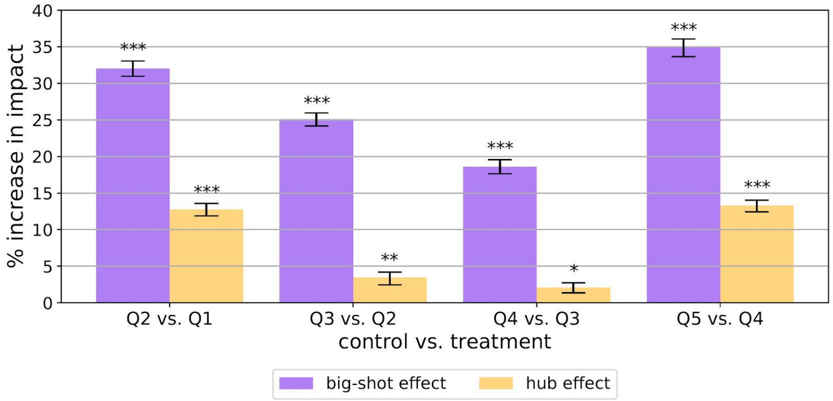
> *"... Then, we vary the fraction of female mentors to understand how this affects the protégé. More specifically, for any given i>0, we compute the change in the post-mentorship impact of the protégés in F_i relative to the post-mentorship impact of those in F_0, which we refer to by writing F_i vs. F_0."*

(8) Just a thought: In the discussion the authors lay out potential reasons for their findings. Personally, I think *visibility* is a big part of what drives the findings here. That if a young scientist publishes with a well known scientist people will notice that famous scientist X has a new paper with protege Y and notice protege Y more. If that hypothesis is correct, using the average citation count of the top mentor only should yield stronger results than average citations of all mentors. And the restriction that a mentor is from the same university shouldn't matter. The key thing is scientific visibility arising from publishing with a famous "name".

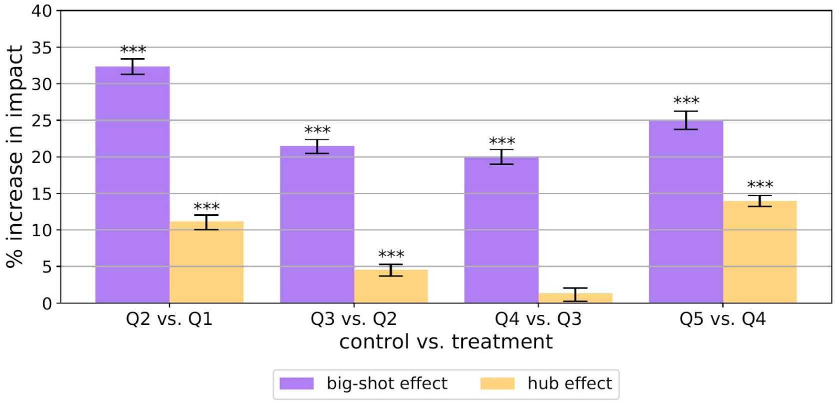
Response:

We agree that such analysis is needed. To address this point, as part of our newly introduced robustness analysis, we reproduce Figure 2 using the impact of the *top* mentor encountered instead of the *average* over all mentors. The results can be found in Supplementary Figure S5 and Supplementary Tables S10 and S11. The results exhibit very similar trends compared to those in Figure 2. For the reviewer's convenience, both figures are pasted below to facilitate the comparison between two.

Figure 2:



Supplementary Figure S3:



<b>REVIEWER COMMENTS</B>

Reviewer #1 (Remarks to the Author):

Thanks to the authors for their response.

I still maintain that calling co-authorship as mentorship is problematic. While I appreciate the survey was done and the efforts in processing the new MAG dataset (which I agree is better quality than the one from 2016), the are several problems with the survey. First, it does not discard the possibility that respondents who had a mentorship relationship were the ones who replied. In contrast, people who did not feel attached to the relationship did not respond. Approximately a 9% response rate (167 out of 2000 people reply) seems to signal this. Second, you only asked the "protegé" but not the mentor. At least by asking both, you can control for some factors. And third, you are not controlling for the field. Since your study is large, there are enormous differences in mentoring and co-authorship patterns, from close one-on-one but largely independent work in Economics/Sociology to large groups and close interaction seen in Biomedical sciences. Therefore, I do not think this paper is about mentorship.

I think you need to start from grad student/postdoc relationships to study mentorship. I would suggest looking at the Proquest doctoral thesis dataset as a start.

Reviewer #2 (Remarks to the Author):

The authors have sufficiently addressed some of the comments and criticisms of the reviewers and did a commendable job in combining the large-scale data with a survey (though the survey questions are relatively weak measures of mentorship. For example, when asking a jr author, did you “receive advice,” the answer almost certain has to yes. A better question would ask whether the advice was valuable for career advancement and could not have been gained other than through their senior collaborator).

With that said, I recommend publication conditional on the following changes first being made to the paper.

1) Delete all the causal language in the paper. Another Reviewer makes this point very forcefully and I agree with it as noted in my round 1 comments. It is to the benefit of your readers and you to avoid playing fast and loose with causal language – it only gets you into trouble after the paper publishes.

Claiming causal relationships requires a control and a treatment group and random assignment. YOU HAVE NONE OF THAT. First, control and treatment groups require the treatment group to get the treatment and the control groups not to get the treatment. Your “treatment” and “control” groups both get the treatment and it doesn’t matter that the treatment is at different levels. It is illogical to conclude that smoking causes cancer if everyone in the study smokes, even if some 1 pack a day and others are chain smokers.

To fix this problem, completely remove words like “causal”, “causality”, “impact”, “antecedent” ,“determine” “effect”, “stimulate”, “affect”, and “effects” or related words from the paper. For example, replace the phrase (line 118) “average causal effect” with “average correlation or average association.” Or, say “informal mentorship predicts junior faculty future success.”

2) Correct your presentation of CEM – CEM does not provide evidence for a causal relationship only “causal inferences.” CEM helps when observational data that have clear treatment and control groups but no random assignment. The matching is a surrogate for the lack of random assignment but without a true treatment and control group the matching doesn’t really do anything that a fixed effects regression does in terms of buying you causal claims. So, your CEM in a technical sense doesn’t make your analysis wrong or bias your estimates it just means that the estimates are basically the same estimates achieved by using a fixed effect model. If you run a fixed effect model, you will get nearly identical estimates. Try it, you’ll see. This is issue is especially important to rectify because you matching is relatively weak match (e.g., big shots vs non-big shots is a subjective measure sensitive to changes in thresholdin).

To fix this problem removing all the wording about causality and state that “CEM is a form of regression that can improve “causal inference” even though the results cannot be claimed to establish causality.”

3) Revise the paper’s title. Remove the word impact from the title and make the title accurately reflect your analysis of informal mentorship through coauthorship.

“The Association between Early Career Informal Mentorship in Academic Collaborations and Junior Author Performance.”

4) Update the references. The paper, “Early coauthorship with top scientists predicts success in academic careers,” by Li et al. does not get the credit it deserves given its overlap with your work. The Li paper literally – just like you -- looks at early coauthors collaborating with top scientist and their subsequent academic performance. So, they examine the same concepts as you but with different measures. Thus, referencing their paper only in the Discussion is unfair. You need to reference the paper in your intro/literature review and state how your paper addresses limitations in their paper, extends their paper, or both. Use their paper to show why your paper is a contribution.

The recent literature also has a new paper on mentorship and protégé success published in PNAS by Ma et al. Like you, Ma et al. use big data, examine link between mentorship and student performance, including coauthorship, and big shots, but differ in that they examine formal mentorship. To make your paper up-to-date, you need to discuss why their results differ from your results especially in regard to their finding that coauthoring with one’s mentor is inversely correlated with student success because it suggests a lack of the student’s intellectual originality. Here is an opportunity to connect the concept of mentorship and informal mentorship in a meaningful way, which would be another contribution of your paper.

Reviewer #3 (Remarks to the Author):

I think the authors did a fine job revising. I think the choices made are reasonable even if not always obviously correct (if there is such a thing in these cases). I recommend publication.

Reviewer #4 (Remarks to the Author):

I have reviewed the author's rebuttal and found the answers convincing. I am happy with publication in the current form.

Dear Reviewers,

We thank the Editor and the Reviewers for their constructive feedback in guiding our process of revision. Indeed, your comments helped greatly improve and significantly expand our manuscript. Below, we specify our responses (highlighted in blue) to every comment. We also highlight all our modifications in blue in the newly submitted main manuscript and SI.

Reviewer 1:

Thanks to the authors for their response.

I still maintain that calling co-authorship as mentorship is problematic. While I appreciate the survey was done and the efforts in processing the new MAG dataset (which I agree is better quality than the one from 2016), there are several problems with the survey. First, it does not discard the possibility that respondents who had a mentorship relationship were the ones who replied. In contrast, people who did not feel attached to the relationship did not respond. Approximately a 9% response rate (167 out of 2000 people reply) seems to signal this. Second, you only asked the "protegé" but not the mentor. At least by asking both, you can control for some factors. And third, you are not controlling for the field. Since your study is large, there are enormous differences in mentoring and co-authorship patterns, from close one-on-one but largely independent work in Economics/Sociology to large groups and close interaction seen in Biomedical sciences. Therefore, I do not think this paper is about mentorship.

Response:

We agree that the response rate should be discussed further in the article, thank you for pointing this out. While selective non-response could be a possible explanation, we believe there are other factors at work. First, when asking a junior scientist, X, about a senior scientist, Y, in the context of their collaboration on paper Z, we do not mention Y nor Z until the junior starts taking the survey. More specifically, when we emailed 2000 protégés, the email included a survey link leading them to a Qualtrics website, stating that we are conducting a study on “scientific collaboration”, *without providing any further details* such as the fact that the survey will ask the respondent about a senior collaborator. This eliminates the possibility that participants who did not click on the survey link could have done so due to the fact that they did not feel attached to their senior collaborator, or felt they have not received the necessary support. As for the fact that, out of the 2000 scientists that we contacted, only 167 filled the survey, we agree that this is a small percentage. Nevertheless, we believe this is expected in such cases. After all, very few scientists are willing to participate in a survey mentioned in an email not coming from their own institution, which they have received from complete strangers. We agree that these details are important, and should be mentioned in our study. Based on this, we added the following text to Supplementary Note 4:

> … For each such scientist, s_i, we randomly selected one of the scientists whom we identified as their mentors, denoted as s_j. The survey focused on a randomly selected paper, p_k, that s$_{\mathrm{i}}$ wrote with s_j. We manually extracted the emails of the 2000 scientists from publicly available sources, such as their personal web pages or the website of their latest affiliation. The scientists were subsequently sent an email asking them to participate in a short survey about “scientific collaboration”, without mentioning scientist s_j nor paper p_k, nor the fact that the survey was about mentorship. This eliminates the possibility that participation rates were greater among those who received greater support from the selected mentor, $s_{j,}$ or among those who had a more positive experience working on the selected paper, p_k. A Qualtrics link was provided at the end of the email directing them to the survey.

Two sentences later, we add the following text:

> … The survey questions we analyze can be found below. Out of the 2000 scientists, only 179 clicked on the Qualtrics link. This is expected given the typically low participation rates that are observed when receiving survey requests via emails outside of one’s organization. For each scientist, s_i, who clicked on the link, we first asked two questions, verifying whether they were an author on paper p_k and whether they collaborated with scientist s_j, respectively. This was primarily to verify whether the email address we extracted indeed belonged to s_i. Out of the 179 scientists, 93.3% (i.e., 167) answered “Yes” to the first question, i.e., the one about paper p_k. This implies that the remaining 6.7% were incorrectly identified during our process of manual email extraction. Out of those who answered “Yes" to the first question, 100% answered “Yes” to the second question, i.e., the one about s_j. Note that none of the scientists who clicked on the Qualtrics link dropped out of the survey willingly when they learnt that it was about paper p_k and collaborator s_j. All of them were willing to complete

the survey, but those who answered “No” to the first question were prevented from participating in it. A summary of their responses to questions 3 and 4 is provided in Figure 1 in the main manuscript.

Regarding the fact that we surveyed the protégés but not their mentors, we made a deliberate choice, since our analysis takes the perspective of the protégé. The purpose of the survey was to verify that the junior scientists recognize that their senior collaborators provide them with mentorship not only in the context of their collaboration, but also in the context of their career development in general. That being said, we agree that the paper would benefit from further evidence that the protégé received some form of mentorship from the senior collaborator. We address this concern as follows:

1. We now elaborate on our definition of mentorship, which justifies our survey. Specifically, we added the following text to the main manuscript:

   > *While we acknowledge that it is possible for juniors to receive support from their junior collaborators, we interpret mentorship as the support that juniors receive from their senior collaborators, following the standard definition of mentorship as “the activity of giving a younger or less experienced person help and advice over a period of time” [43]. Based on this definition, the difference in experience between the protégé and their mentor seems to be a necessary, albeit not sufficient, condition for the relationship to be considered mentorship. In addition to the difference in experience, the relationship also needs to involve some form of support from the mentor to the protégé. Arguably, the fact that the mentor has coauthored a paper with the protégé provides evidence that the former indeed supported the latter. Nevertheless, it would be desirable to provide further evidence that the mentor supported the protégé in ways related not only to the paper on which they are collaborating, but also to career development in general. To verify whether this is the case, we sampled 2000 scientists ...*

2. We now revisit questions that we have included in the survey but have not analyzed, as we did not believe that this was what the reviewers were looking for. Importantly, these newly-added questions focus on any support that the junior may have received from the senior in terms of career development in general, i.e., *outside of the context of their collaboration*. The questions are

now mentioned in Supplementary Note 4, and the responses are summarized in Figure 1 of the main manuscript. Both the questions and the new version of Figure 1 are pasted below for the reviewer’s convenience.

> *Which of these statements are true about your collaborator, <Senior Collaborator Name>?*
>
> *Keep in mind that these statements do not necessarily describe events during the time of your collaboration.*
>
> - *I received grant writing advice from him/her. (T/F)*
> - *I received a letter of recommendation from him/her for a fellowship/award or job application (T/F)*
> - *I received career planning advice from him/her (T/F)*
> - *He/she put me in touch with an important person in my field (T/F)*

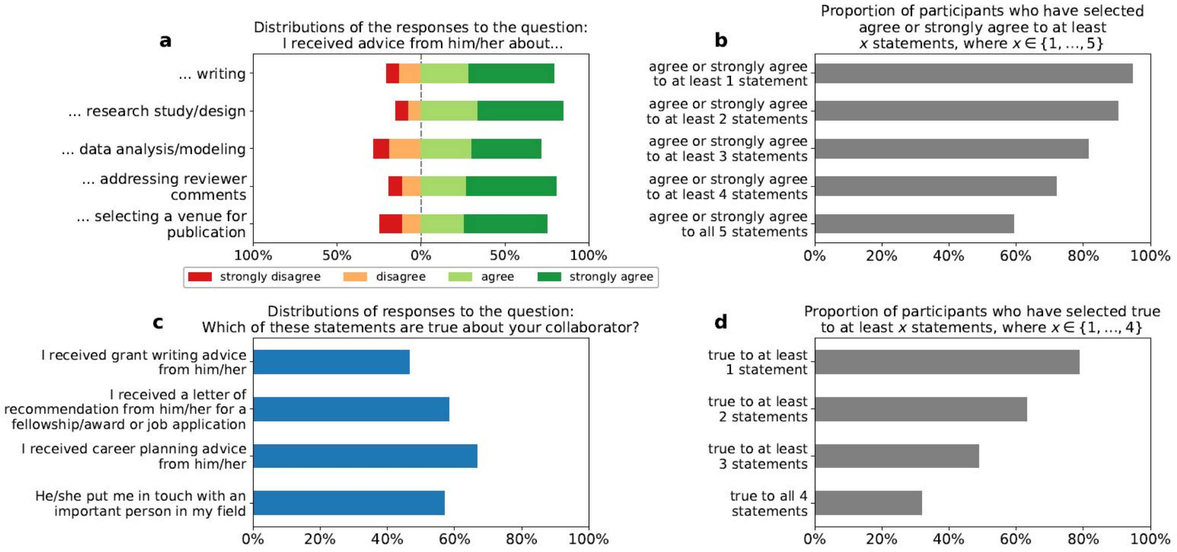
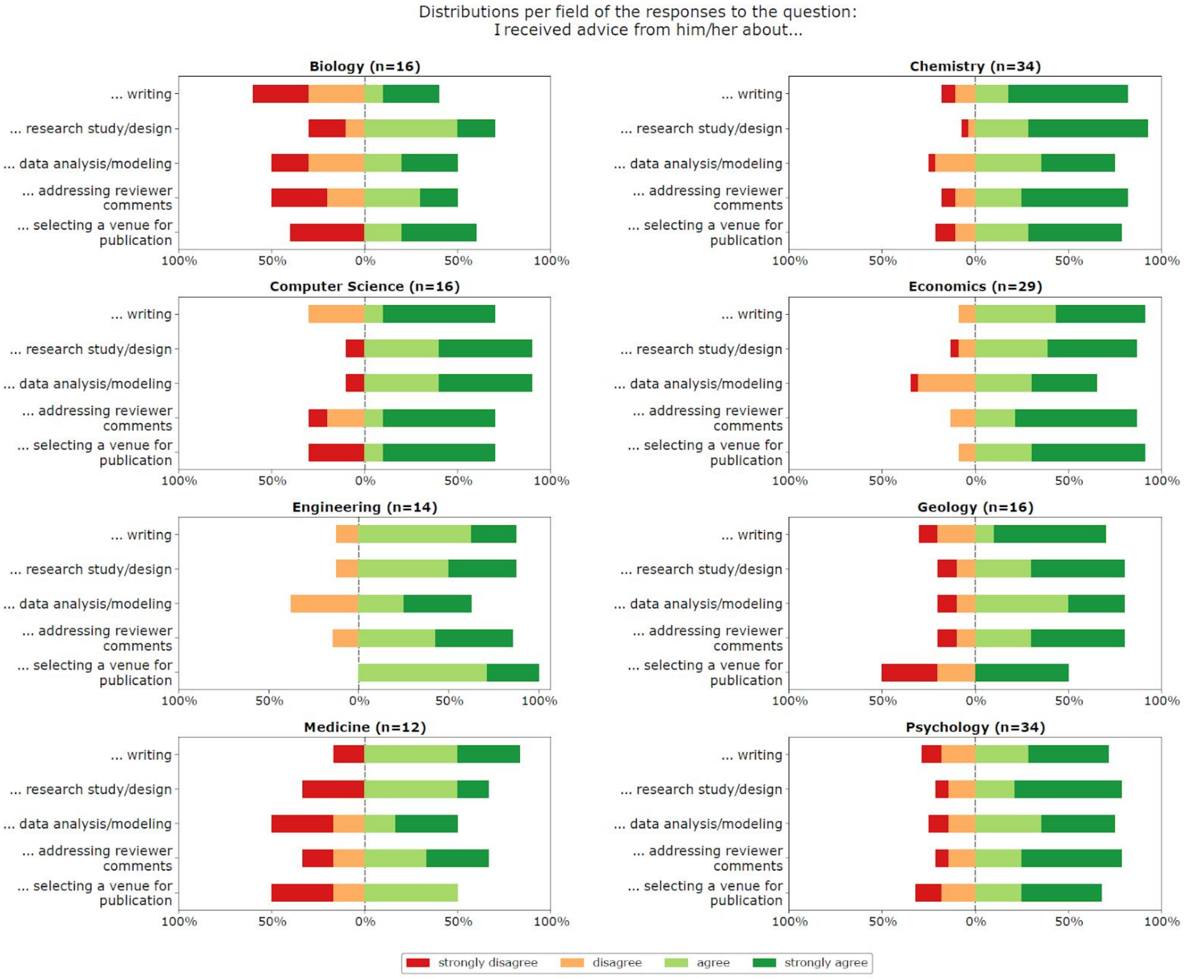
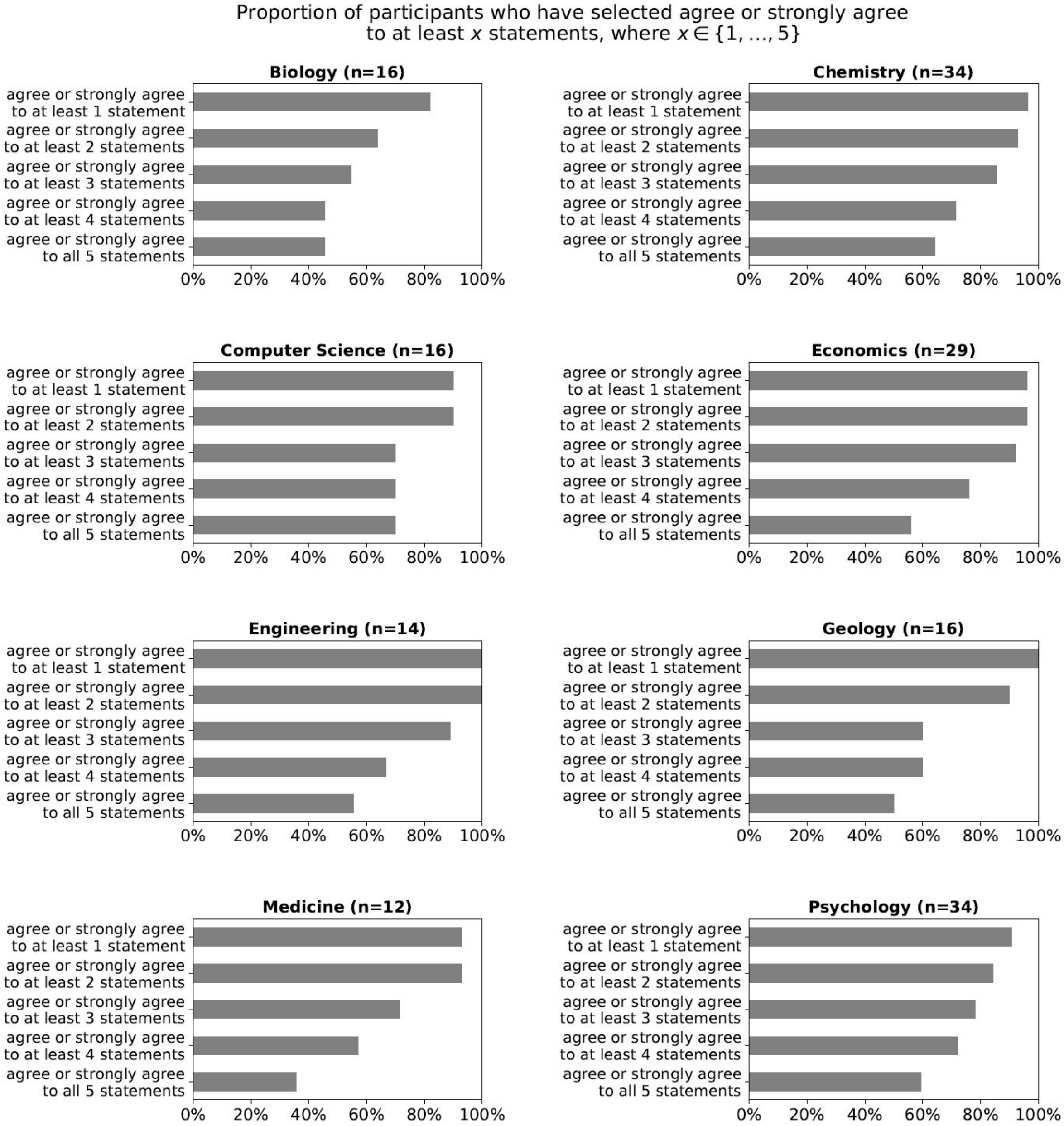


Figure 1: **Survey outcome.** Responses of 167 randomly-chosen scientists who were identified as protégés and asked about their relationship to a scientist who was identified as one of their mentors. **a**, Distributions of the responses to each of five statements regarding their senior collaborator, where the statements take the form "*I received advice from him/her about...*" followed by five different skills: (i) writing; (ii) research study/design; (iii) data analysis/modeling; (iv) addressing reviewer comments; (v) selecting a venue for publication. **b**, A different way of summarizing the responses in (**a**), showing the proportion of participants who either agree or strongly agree to at least x out of the 5 statements regarding their senior collaborator, where $x \in \{1, \ldots, 5\}$. **c**, The percentage of protégés who selected "true" for each of the following four statements regarding their senior collaborator: (i) I received grant writing advice from him/her; (ii) I received a letter of recommendation from him/her for a fellowship/award or job application; (iii) I received career planning advice from him/her; (iv) He/she put me in touch with an important person in my field. **d**, A different way of summarizing the responses in (**c**), showing the proportion of participants who have selected true to at least x out of the 4 statements regarding their senior collaborator, where $x \in \{1, \ldots, 4\}$.

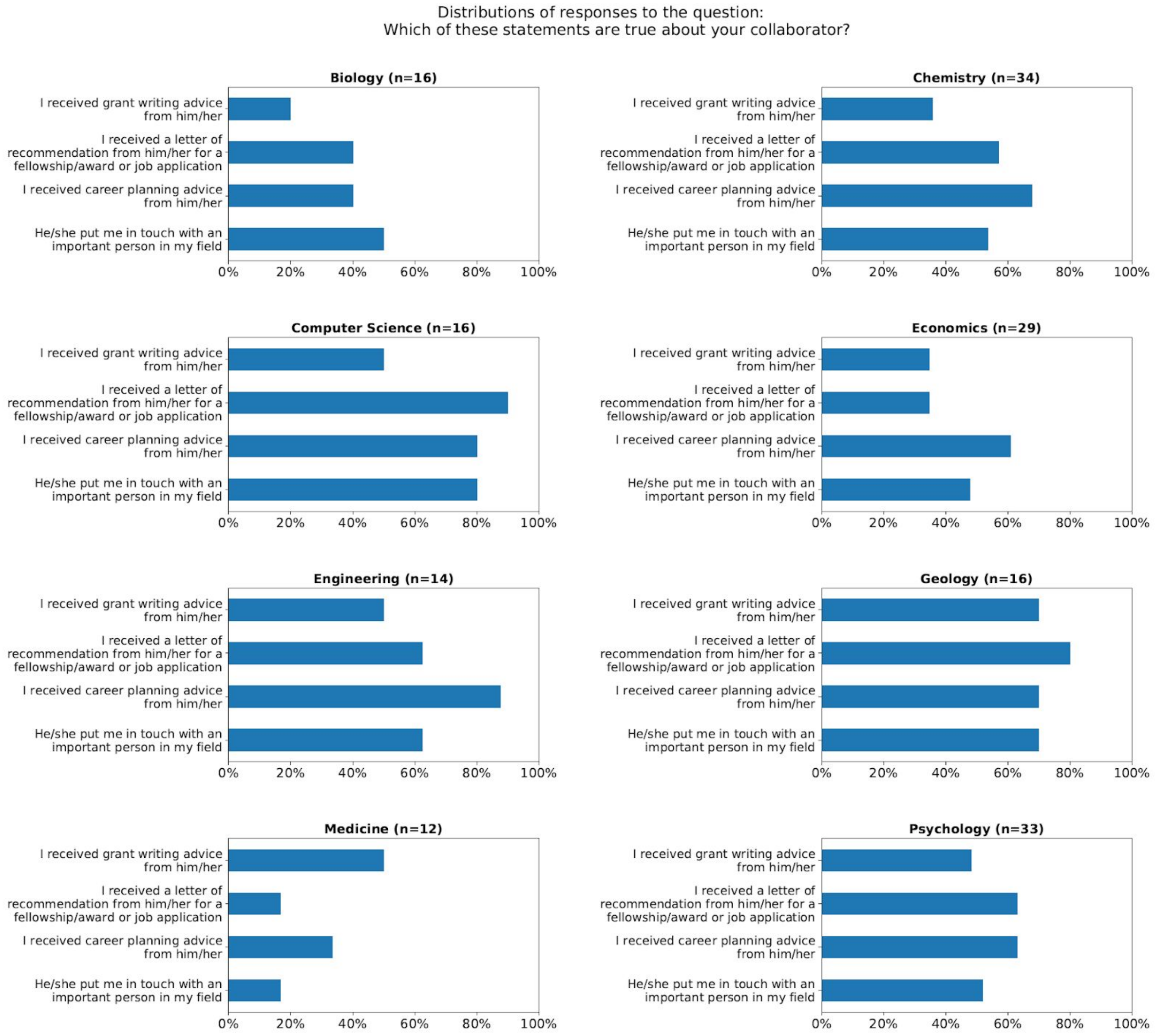
Lastly, we agree that it would be interesting to see the difference across disciplines. We now show the breakdown of the results from Figure 1 across 8 out of the 10 disciplines; see Supplementary Figures S2 to S5. The two missing disciplines were the result of updating our dataset as per the suggestions of multiple reviewers; an update that was carried out after running our survey. As can be seen from these newly-added figures, the results vary somewhat from one discipline to another, as would be expected. Despite these differences, the broad trend is similar to the one observed in Figure 1. The figures are pasted below for the reviewer's convenience:



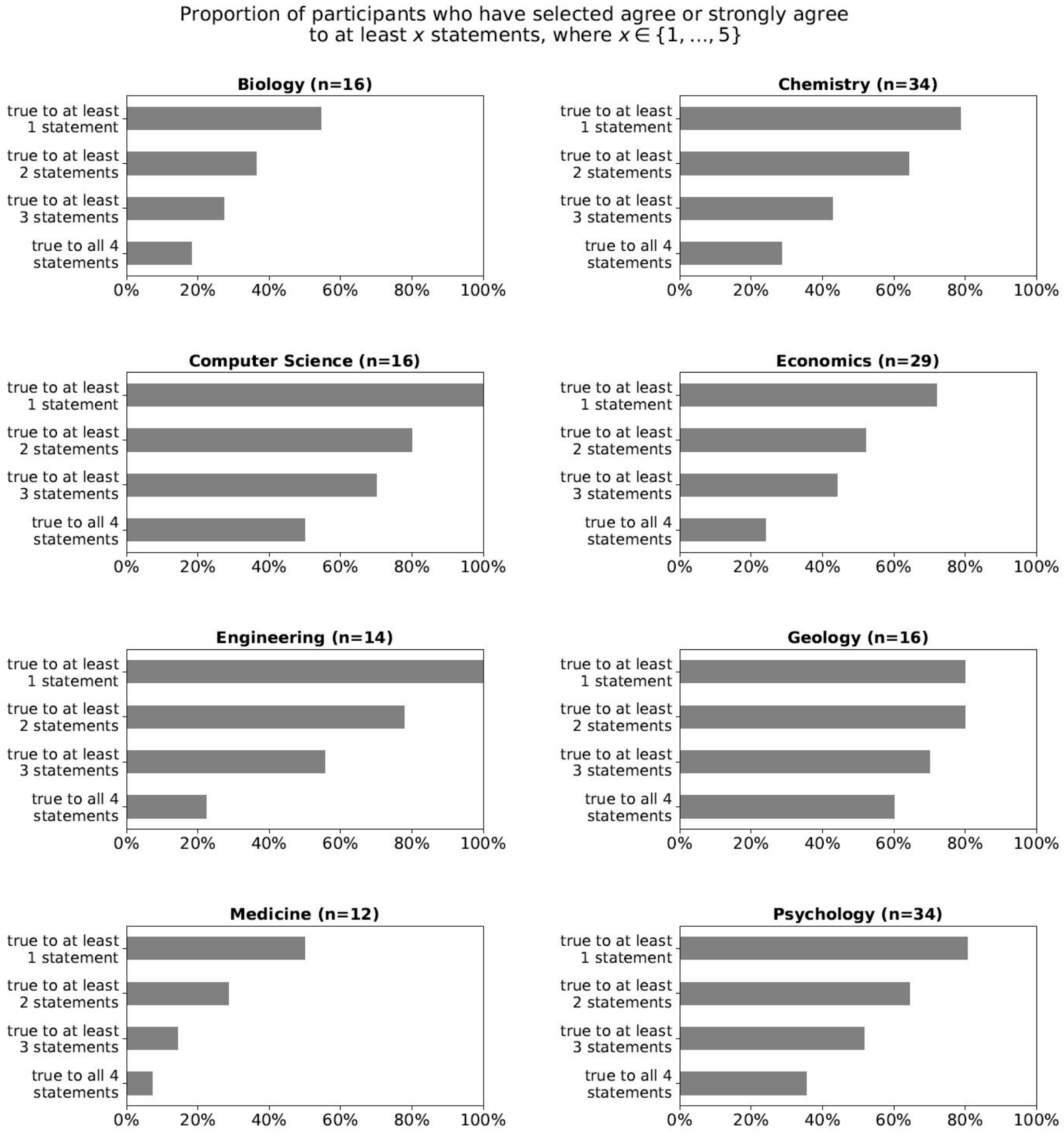
Supplementary Figure S2: **The same as Figure 1a but across different disciplines.**



Supplementary Figure S3: **The same as Figure 1b but across different disciplines.**



Supplementary Figure S4: **The same as Figure 1c but across different disciplines.**



Supplementary Figure S5: **The same as Figure 1d but across different disciplines.**

I think you need to start from grad student/postdoc relationships to study mentorship. I would suggest looking at the Proquest doctoral thesis dataset as a start.

Response:

The Proquest doctoral thesis dataset was used in the very recent study by Ma et al. [PNAS-2020]. Thus, studying this dataset would render our study similar to theirs. Our contribution lies in the fact that we study informal mentorship. We now highlight the main differences between our study and theirs in the following paragraph, which was newly added to the introduction:

> *Another recent paper that is closely related to ours is the one by Ma et al. [41], who study how the success of junior scientists is related to the ability of their mentors to create and communicate prizewinning research. As such, their work resembles ours in the sense that they also study some form of academic success and how it is related to mentorship. However, they study formal mentorship, where the mentor is the official PhD advisor of the protégé. In contrast, our study covers informal mentorship whereby juniors are mentored by multiple senior colleagues without them necessarily having formal supervisory roles. Furthermore, their analysis of the protégé's performance post mentorship includes papers written with the mentors, leading to their finding that coauthoring with one’s advisor is inversely correlated with one's success. In contrast, our analysis excludes papers written with any of the scientists who served as mentors during the mentorship experience; this ensures that the observed impact is not attributed to the mentors but rather to the protégés.*

Reviewer 2:

The authors have sufficiently addressed some of the comments and criticisms of the reviewers and did a commendable job in combining the large-scale data with a survey (though the survey questions are relatively weak measures of mentorship. For example, when asking a jr author, did you “receive advice,” the answer almost certain has to yes. A better question would ask whether the advice was valuable for career advancement and could not have been gained other than through their senior collaborator).

With that said, I recommend publication conditional on the following changes first being made to the paper.

1) Delete all the causal language in the paper. Another Reviewer makes this point very forcefully and I agree with it as noted in my round 1 comments. It is to the benefit of your readers and you to avoid playing fast and loose with causal language – it only gets you into trouble after the paper publishes.

Claiming causal relationships requires a control and a treatment group and random assignment. YOU HAVE NONE OF THAT. First, control and treatment groups require the treatment group to get the treatment and the control groups not to get the treatment. Your “treatment” and “control” groups both get the treatment and it doesn’t matter that the treatment is at different levels. It is illogical to conclude that smoking causes cancer if everyone in the study smokes, even if some 1 pack a day and others are chain smokers.

To fix this problem, completely remove words like “causal”, “causality”, “impact”, “antecedent” ,“determine” “effect”, “stimulate”, “affect”, and “effects” or related words from the paper. For example, replace the phrase (line 118) “average causal effect” with “average correlation or average association.” Or, say “informal mentorship predicts junior faculty future success.”

Response:

Thank you for your comment. We have now removed all causal language from both the main manuscript and the supplementary notes.

2) Correct your presentation of CEM – CEM does not provide evidence for a causal relationship only “causal inferences.” CEM helps when observational data that have

clear treatment and control groups but no random assignment. The matching is a surrogate for the lack of random assignment but without a true treatment and control group the matching doesn't really do anything that a fixed effects regression does in terms of buying you causal claims. So, your CEM in a technical sense doesn't make your analysis wrong or bias your estimates, it just means that the estimates are basically the same estimates achieved by using a fixed effect model. If you run a fixed effect model, you will get nearly identical estimates. Try it, you'll see. This issue is especially important to rectify because your matching is relatively weak (e.g., big shots vs non-big shots is a subjective measure sensitive to changes in threshold).

To fix this problem remove all the wording about causality and state that "CEM is a form of regression that can improve "causal inference" even though the results cannot be claimed to establish causality."

Response:

Thank you for your comment. In addition to removing all causal language in response to your first comment, we added the following sentences to the beginning of the paragraph in which we introduce CEM in the main manuscript:

> *"We aim to establish whether mentorship quality (measured by big-shot experience or network experience) is associated with the post mentorship outcome. To this end, we use coarsened exact matching (CEM) [44]. While this technique does not establish the existence of causal effect, it is commonly used to infer causality from observational data."*

3) Revise the paper's title. Remove the word impact from the title and make the title accurately reflect your analysis of informal mentorship through coauthorship.

"The Association between Early Career Informal Mentorship in Academic Collaborations and Junior Author Performance."

Response:

The title has been updated as per your suggestion.

4) Update the references. The paper, "Early coauthorship with top scientists predicts success in academic careers," by Li et al. does not get the credit it deserves given

its overlap with your work. The Li paper literally – just like you -- looks at early coauthors collaborating with top scientists and their subsequent academic performance. So, they examine the same concepts as you but with different measures. Thus, referencing their paper only in the Discussion is unfair. You need to reference the paper in your intro/literature review and state how your paper addresses limitations in their paper, extends their paper, or both. Use their paper to show why your paper is a contribution.

Response:

We now cite the paper by Li et al. in the introduction and give them credit for studying how the impact of junior scientists is related to the impact of their past collaborators. We also highlight the main differences between our study and theirs, to emphasize how our contribution complements theirs. More specifically, we added the following text to the introduction:

> *It should be noted that we are not the first to study how the impact of junior scientists is related to the impact of their past collaborators. A recent study by Li et al. [40] found that juniors who publish with "top scientists" enjoy a persistent competitive advantage throughout the rest of their careers. More specifically, they focus on collaborators who are among the 5% most impactful scientists in any given year, regardless of whether they are senior or junior. In contrast, as we will show, our study focuses on collaborators who are likely to have served as mentors, regardless of whether they are among the top 5%. In other words, Li et al. study coauthorship with top scientists, while we study coauthorship with mentors. Another difference between their study and ours is that they do not address the fundamental question of whether the social capital of collaborators matters more than their impact; we address this question by analyzing not only the mentors' impact but also their collaboration network. Finally, unlike their paper, our study complements existing literature on women in science, by analyzing the gender of both the protégés and their mentors, and how these shape mentorship experiences.*

The recent literature also has a new paper on mentorship and protégé success published in PNAS by Ma et al. Like you, Ma et al. use big data, examine link between mentorship and student performance, including coauthorship, and big shots, but differ in that they examine formal mentorship. To make your paper up-to-date, you need to discuss why their results differ from your results especially in regard to their finding that coauthoring with one’s mentor is inversely correlated with student success because it suggests a lack of the student’s intellectual originality. Here is an opportunity to connect the concept of mentorship and informal mentorship in a meaningful way, which would be another contribution of your paper.

Response:

We have added the following text to the introduction to cover the work by Ma at al.:

> *Another recent paper that is closely related to ours is the one by Ma et al. [41], who study how the success of junior scientists is related to the ability of their mentors to create and communicate prizewinning research. As such, their work resembles ours in the sense that they also study some form of academic success and how it is related to mentorship. However, they study formal mentorship, where the mentor is the official PhD advisor of the protégé. In contrast, our study covers informal mentorship whereby juniors are mentored by multiple senior colleagues without them necessarily having formal supervisory roles. Furthermore, their analysis of the protégé's performance post mentorship includes papers written with the mentors, leading to their finding that coauthoring with one’s advisor is inversely correlated with one's success. In contrast, our analysis excludes papers written with any of the scientists who served as mentors during the mentorship experience; this ensures that the observed impact is not attributed to the mentors but rather to the protégés.*

Reviewer 3:

I think the authors did a fine job revising. I think the choices made are reasonable even if not always obviously correct (if there is such a thing in these cases). I recommend publication.

Response:

Thank you for the time and effort you have put into reviewing our manuscript, and for your constructive feedback which certainly improved the paper.

Reviewer 4:

I have reviewed the author's rebuttal and found the answers convincing. I am happy with publication in the current form.

Response:

Thank you for the time and effort you have put into reviewing our manuscript, and for your constructive feedback which certainly improved the paper.

**REVIEWERS' COMMENTS:

Reviewer #1 (Remarks to the Author):

Thanks for the further details. I appreciate that authors have added language toning down the claims in the paper. I think the clarifications and further drilling down on the survey's data added significant value and nuance to the results. I thank the authors for such details.