Researchers' Goals When Assessing Credibility and Impact in Committees and in Their Own Work

James Harney[1], Louise Mayville[2], Iain Hrynaszkiewicz[2], Veronique Kiermer[1]

Correspondence to jharney@plos.org

1. PLOS, 1265 Battery Street, Suite 200, San Francisco, CA 94111, USA
2. PLOS, Carlyle House, Carlyle Road, Cambridge CB4 3DN, UK

Abstract

In a series of 52 semi-structured interviews with researchers in cell biology, we sought to characterize researchers' goals when evaluating the credibility (or trustworthiness) and impact of research outputs in two contexts: during researchers' own work (the Discovery context) and when researchers participate in research assessment committees for grant review and hiring and promotion (the Committee context). We have compiled a list of researchers' goals in these contexts, expressed as desired outcome statements and standardized across the two contexts, which will inform a quantitative survey to validate and prioritize these goals and to identify opportunities for new or improved solutions for research assessment. Based on the qualitative data, we examined how these needs intersect in the two contexts. We find that the goals of researchers in the Discovery and Committee context overlap significantly. Both impact and credibility matter in each context. In particular, credibility is the dominant factor in the Discovery context and somewhat less represented but still strongly relevant in the Committee context. Researchers use proxy methods, in particular journal-based proxies, to evaluate all attributes of research outputs and these proxies were reported with similar frequency in both contexts. We also find that researchers seek to understand reproducibility, quality and novelty of research outputs in both contexts, in addition to credibility and impact. While publications remain the dominant unit of research assessment, researchers in our sample also evaluate research data, code and preprints, in both contexts. Our preliminary findings suggest potential opportunities to reduce time, reduce error, or improve the quality of assessment practices, in a manner that avoids journal-based proxies. Amongst these improvements are potential opportunities to (i) provide more reliable signals of credibility, quality, and impact, (ii) apply these signals to publications and preprints, and (iii) improve research assessment guidelines.

Introduction

Researchers routinely make judgments about the credibility and impact of research outputs (e.g., preprints, publications, data, code) during their own research (the Discovery context) and when participating in research assessment committees for grant review and hiring and promotion (the Committee context). We set out to characterize researchers' goals when they evaluate the credibility and impact of research outputs in these two situational contexts. We wanted to examine how these needs intersect in the two contexts and to discern the relative importance of credibility, impact and related concepts. This characterization will be the basis of further examination through survey

research. Our ultimate aim is to increase understanding of how researchers assess impact and credibility of research outputs and to help develop better solutions for research assessment.

For the purpose of this project, we are defining these concepts as follows:

- Credibility, or trustworthiness, reflects the likelihood that the work is robust and reliable. In previous research, diverse labels have been used to refer to the same concept of trustworthiness including reliability and cognitive authority. To the extent that it often refers to adherence to norms of scientific practice and rigor [ref. 1], logical reasoning and robust data, credibility can be related to quality.
- Impact reflects influence either in academia and the research sector (for example by advancing knowledge of a phenomenon in a meaningful way or significantly expanding capabilities) or influence on society or the economy (for example through policies or commercial developments). Influence is also often related to visibility and renown. In the context of research assessment, impact is a main preoccupation of university administrators and funding agencies because it relates to return on investment [ref. 2].

Previous studies have examined how researchers make judgments about the trustworthiness of research outputs. In particular, researchers have focused in the past two decades on trustworthiness of publications in light of the digital transformation of scholarly publishing that has given researchers unprecedented access to an increasing number of scholarly outputs. Quality and trustworthiness have been consistently recognized as prime criteria, in addition to topical relevance, that scholars use when discovering new information on the web and deciding which material to engage with [ref. 8,9]. In the case of scholarly publications, these criteria have been shown to be influenced by a complex array of characteristics and clues [ref. 8,10]. Quality tends to be best determined by personal inspection (like reading the abstract, assessing the methodology, checking for sound logic and credible data) and to some extent are influenced by practicalities such as accessing the material [ref. 11]. Perceived quality is also typically associated with peer review [ref. 9]. Both quality and trustworthiness determinations have been shown to be influenced by social and traditional clues. Social clues include colleagues' recommendations (including through social media) and familiarity with the authors. Traditional clues center on the reputation or brand of the journal and extend through metrics like impact factor used as proxies when dealing with information outside of their own field of study [ref. 11,9]. Researchers who have studied these judgements also stress the influence of academic realities in complicating the trust picture. For example, citing behaviors are guided by networks of social and research influence and show a greater influence on the reputation of authors, journals or institutions [ref. 12].

A large body of literature has focused on the influence of various metrics in research assessment, described flaws of commonly used metrics like the journal impact factor and called for responsible practices [reviewed in ref. 2, see also ref. 22, 23, 24, 25, 26]. However, despite awareness of the flaws of a metric like the journal impact factor, scholars continue to use it. The journal impact factor matters more to decide where to publish than to decide what to read or cite—an observation that reflects the influence of tenure, promotion and other university policies [ref. 11]. These influences remain strong to this day, as illustrated by a 2020 study [ref. 3].

Moreover, research assessment exercises are typically done in very constrained environments: hyper-competitive situations in which many qualified candidates compete for limited funding and

positions, with limited time and resources for assessors to achieve informed decisions. In such environments, priority is given to impact as measured through easily accessible metrics, with a central role for the journal impact factor [ref. 27,3].

The relationship between perception of impact and metrics such as journal impact factor, which is common in assessment instructions [ref. 2, 3], creates an assessment system which assigns a premium value to publication in a small number of highly selective journals with high impact factors—a system in which the journal name or impact factor becomes an easily accessible proxy for some intrinsic characteristics of the research [ref. 27]. Multi-stakeholder organizations like DORA [ref. 8] have stressed the negative consequences of this system and the need for reform and realignment of research incentives [ref. 5].

In 2014, Tenopir, Nicholas and colleagues concluded that attitudes towards trust don't evolve quickly and that despite the transformation of scholarly communication from print to digital, scholars continue to use similar traditional and social clues to decide what to read, to cite and where to publish. However, they did report differences in how younger researchers assess trustworthiness. Utility and pragmatism, which were important influences for all demographics, are central for young researchers who spend less effort to obtain information. Younger researchers also viewed Open Access publishing much more positively and used more outlets to improve visibility of their work. These findings suggested that change might still be forthcoming [ref. 13].

Recent developments of the scholarly communication landscape provide an opportunity to probe this further. In particular, the new momentum that preprints have been gathering in the biological and biomedical sciences has created new challenges for researchers and other stakeholders in evaluating this type of publication without the traditional framework of journal peer review [ref. 14]. Interestingly, a recent survey of more than 3,700 researchers across a wide range of disciplines examined how they assess the credibility of preprints and indicated that cues related to information about Open Science content and independent verification of authors' claims are highly important for judging preprint credibility [ref. 15]. It is possible that what appears as an evolution of behaviors towards trustworthiness with regards to Open Science is consistent with a previously identified influence of personal inspection, which requires more access [ref. 11].

Similarly, the notion of independent verification of authors' claims may be related to the previously identified relationship between peer review and perceived quality [ref. 9]. As another example of change in the scholarly communication landscape, publishers are now increasingly providing access to peer review reports alongside published papers, which offers new ways of considering peer review and as a qualifier of quality and trustworthiness [ref. 16]. Furthermore, PLOS, ASCB and others have developed initiatives to encourage the adoption of preprints as the mechanism for more rapid, author-controlled dissemination of research [ref. 17, 18, 19], and to facilitate review of preprints in a journal-agnostic way [ref. 20,21].

The present study extends existing research in attempting to define the overlapping needs of the Discovery and Committee contexts by applying the same methodology for each context. Our methodology was adapted from a "jobs to be done" framework, specifically Outcome-Driven Innovation [ref. 6,7]. Following this approach, we conceptualize the assessment of research outputs as a "job" that the researcher is trying to complete. Through a series of interviews with researchers,

we create a detailed map of the various steps of this “job” and determine the specific desired outcomes that researchers are trying to achieve when assessing research outputs.

In contrast to earlier studies, this approach clearly distinguishes the goals the researchers are trying to achieve, i.e., their “desired outcomes,” from the solutions they are presently using to try to achieve those outcomes. By focusing on researchers’ goals as opposed to current practices, we can better understand how we might transform those practices — offering better solutions to achieve the same goals. Moreover, by identifying which goals are regarded by researchers as underserved, our subsequent quantitative research will provide insight into the kinds of solutions researchers will be intrinsically motivated to adopt in place of current solutions.

Methods

We conducted two separate rounds of interviews. The first explored the Committee context, and included a group of active cell biology researchers who have served on research assessment committees within the past year, and focused on the steps they went through while evaluating impact and credibility as part of these committees. This context was broken into two behavioral cohorts, to account for potential contextual and cultural differences between grant committees and hiring committees. Participants were recruited from the US, UK, and EU, and to the extent possible, the behavioral segments were balanced geographically. A second round of interviews explored the Discovery context, and included a single cohort of active cell biology researchers from the US, UK, and EU. These interviews interrogated researchers’ goals when evaluating impact and credibility in the course of their own research.
Given the need to compare the contexts, it was important to ensure that the cohorts had similar characteristics in terms of discipline, geography, and career stage. As it was deemed unlikely we would find a sufficient number of early career researchers on hiring and grant committees, early career researchers (defined as those with < 11 years experience as an active researcher) were excluded from all cohorts using a screening survey.

Participant Recruitment

Our recruitment strategy leveraged a range of methods. We engaged in direct outreach to the ASCB membership, ASCB outreach to non-member cell biologists, utilized social media posts and promoted posts, direct email via PLOS and partner email lists, and distribution of recruitment materials by a variety of partners. Participants were offered a $100 incentive for their participation, available as cash, gift card, or charitable donation.

All respondents completed a brief screening survey to ensure they met our inclusion criteria, including discipline, geography, career stage, and recent experience. The screener survey also gathered additional demographic data including gender and ethnicity that was used to balance the cohorts.

We did not seek approval from a research ethics committee as the research was considered to be low risk and we did not collect sensitive information about the participants. All participants were informed that their anonymized responses will be made available as part of an aggregated dataset, and might be included in a public report. Participants completed a consent form and were informed

that their participation was completely voluntary, and that they were free to withdraw from the study at any time. Answers will never be associated with identifiable individuals and the results will only be published in aggregate. The data collection and storage procedures were compliant with the General Data Protection Regulation 2016/679.

Participant demographics

All participants were cell biologists from the US, UK, or EU (ex-UK) with 11+ years experience as active researchers, who had recent experience in one of three contexts of interest. Respondents who had recent experience in more than one of these contexts were asked to discuss their needs with respect to one assigned experience. Given our aim of comparing these contexts, we attempted to match the cohorts in terms of geography and career stage. Participant demographics are summarized in Table 1 below.

		Discovery context	Committee context	
Cohort	**All**	**Discovery**	**Hiring**	**Grants**
Total	**52**	**22**	**15**	**15**
Male	31	13	8	10
Female	21	9	7	5
US	20	9	6	5
UK	16	6	6	4
EU	16	7	3	6
16+ years	40	14	12	14
11-15 years	12	8	3	1
Prefer not to say	10	4	1	5
White/Caucasian	30	12	9	9
Other ethnicity	12	6	5	1

Table 1: Participants demographics by cohort

Attention to Diversity

All respondents to the screener survey were shown a diversity statement and asked to provide gender and ethnicity information. Respondents were able to select ‘prefer not to say’, and were not excluded from participation based on this response. In order to gather the broadest perspective possible, we prioritized diversity when selecting participants from the group of candidates who met inclusion criteria.

Nearly 66% of respondents to our screening survey who met our inclusion criteria were male. This may be partly due to our interviews being conducted during the COVID-19 pandemic, which may have disproportionately impacted the ability of women to commit to this kind of research, given the disproportionate impact of the pandemic on women [ref. 30]. In constructing our cohorts, we targeted a 60/40 male/female distribution in order to give us the flexibility to balance our cohorts geographically, as well as make some allowance for ethnic diversity.

Interviews

The interviews themselves were “semi-structured”, and used a series of open-ended prompts, rather than a formal list of questions. These prompts were designed to encourage participants to talk through, in some detail, the various steps they go through when evaluating the impact and credibility of research artifacts. The prompts were structured around the generic ‘job stages’ from the ODI framework to ensure we captured contextual steps around the core assessment tasks (see Figure 1). When needed, the interviewer would prompt the participants to provide additional detail and context, or expand on a particular step. The interviewer would also query the participants around various open science practices (preprints, data sharing, etc.) where these did not arise organically in the interview.

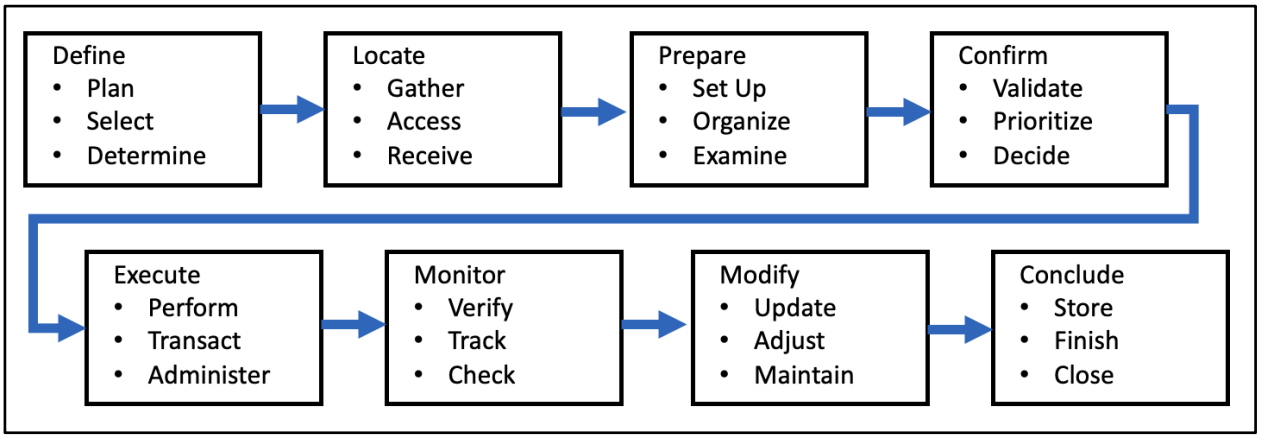

Figure 1: Generic Job Stages, as defined in the ODI framework. See Appendix A for additional detail on how these were implemented in this project. [ref. 7]

Interviews were scheduled for between 60-90 minutes and were conducted using Zoom video conferencing by a lead and secondary researcher. Both video and audio of interviews were recorded for transcription purposes after obtaining the participant’s verbal consent. Recordings of interviews were initially transcribed automatically using Zoom transcription functionality. If an initial quality check revealed issues with this transcription, interviews were subsequently transcribed either manually or using Otter.ai. The interview prompts and screener questionnaire are available on OSF at https://osf.io/sphfj.

Data Analysis

The data from our interviews was analyzed to create a “job map”, which describes the steps a researcher goes through in completing the job of assessment, and the desired outcomes associated with each. For purposes of generating our job map, we defined our core job-to-be-done as: “Assess scientific research as part of a) my own research, b) a hiring committee, or c) a grant committee”. Desired outcomes were structured with reference to this specific core job, not, for example, the broader job of “hiring the correct candidate”. Thus, the outcome statements focus on identifying and understanding various attributes of research outputs.

In order to build this map, transcribed interviews were reviewed by an interviewer, and relevant extracts were compiled into a spreadsheet. These extracts were grouped first into the same generic

job stages that structured our interviews (Figure 1), and then based on the similarity of the goals being described. These goals were captured in desired outcome statements, which express, in a standardized format, the researcher's goal, abstracted from any solutions they might currently use to achieve it. These desired outcome statements were crafted so as to serve as the basis for one or more future quantitative phases of research which will measure the relative importance of the goals uncovered during qualitative research, and how well researchers feel they are addressed by existing solutions.

We added two additional layers of hierarchy to the standard ODI job map. Similar outcomes were compiled into "groups" to better highlight trends in the data, and "tasks" were identified to document how researchers currently attempt to achieve their desired outcomes. Some tasks were identified which could not clearly be attributed to a specific desired outcome. These were included in the job map so as to give as full a view of the researchers' process as possible. The full job map hierarchy is illustrated in Figure 2. The grouped job map data and a data visualization for each cohort are available on OSF at https://osf.io/sphfj.

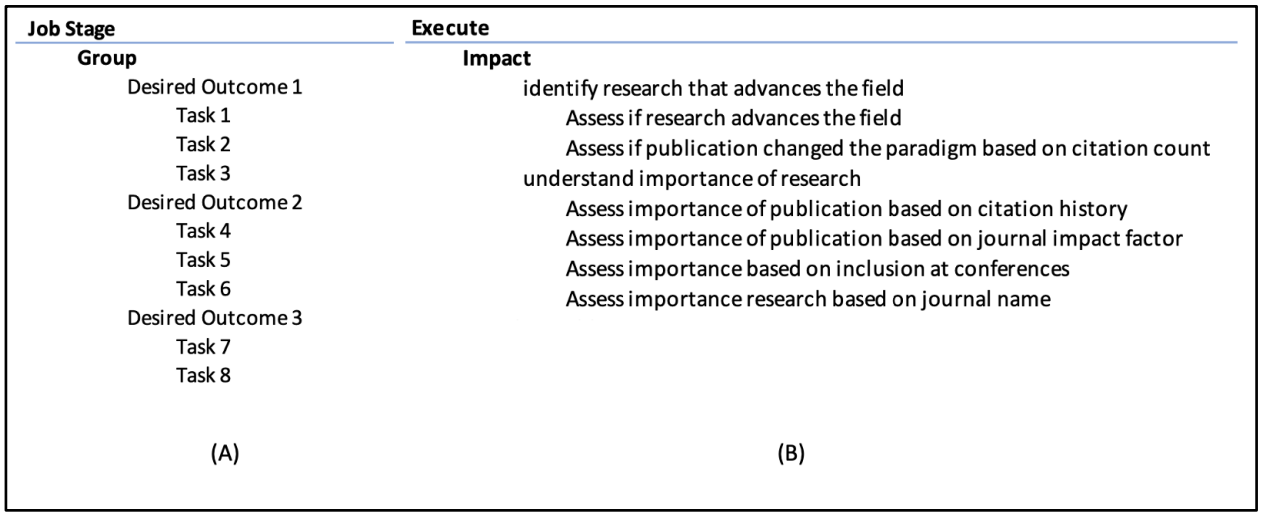

Figure 2: Structure of a single stage and group of the job map, with both abstract example (A), and the Impact group from the Discovery cohort (B). Each job stage is composed of one or more "groups". Each group represents a set of related "desired outcomes", and a given desired outcome is associated with one or more "tasks". The full job map is composed of 8 of these stages, as shown in Figure 1.

Finally, we compared the overlap in the individual job maps of these three cohorts to identify the intersection of desired outcomes between them. Each cohort (Discovery, Hiring, Grants) received a separate job map, which allowed us to overlay the mappings and identify the intersection. Though the semi-structured nature of these qualitative interviews does not allow us to determine the relative importance of various desired outcomes, the job map does allow provisional insight into areas where researchers have a greater concentration of desired outcomes. If a particular cohort of researchers identified more desired outcomes in certain areas, this suggests these areas may be more important to that cohort. Examining the distribution of desired outcomes lets us identify potential trends, though these trends should be validated by future quantitative research.

Results

Intersection between Discovery and Committee contexts

Comparison of the cohorts does reveal a substantial overlap in the desired outcomes reported in the Committee and Discovery contexts, especially around outcomes related to “credibility” and “quality”. Our analysis here focuses on the “Execute” stage of the job map as it encompasses the core tasks of assessing research outputs. The other stages are contextual with respect to the Execute stage (see Appendix A for a summary of the stages applied). There was minimal overlap in the non-Execute stages we mapped, which is unsurprising, as the contexts in which Discovery and Committee assessments occur are quite different. The intersection of desired outcomes in the Execute stage is summarized in Table 2 and Figure 3, and described in more detail below. A full list of desired outcome statements across all job stages is available on OSF at https://osf.io/sphfj.

Group	Outcomes unique to Discovery	Intersecting Outcomes	Outcomes unique to Committees
Credibility	identify potential issues with publication understand the limitations of research identify researchers who produce reliable research understand the crediblity of conclusions understand the crediblity of results minimize time to understand the soundness of methods	understand if the data supports conclusions understand if research was done honestly understand experimental design understand the level of peer review received understand if proper controls are in place understand the soundness of methods understand the validity of statistical analysis understand data credibility understand preprint credibility	minimize the time it takes to understand publication credibility
Quality	understand the strength of data	understand publication quality understand data quality understand preprint quailty minimize the time it takes to understand publication quality	identify exceptional research questions identify publications that tell a complete story understand the quality of preliminary data minimize the likelihood of bias when assessing publications minimize the time it takes to understand preprint quality
Impact		identify research that advances the field understand the importance of research	identify highly visibile publications identify potentially impactful publications identify seminal research understand the impact of available code understand the impact of research understand the incremental value of research identify research with social impact understand the impact of dataset understand the impact of publication minimize the time it takes to understand importance of research maximize the likelihood of identifying impactful research
Novelty	identify novel publications identify innovative publications	identify publications with novel methodologies identify publications that demonstrate original thinking	identify publications with innovative methodologies
Reproducibility	ensure results can be reproduced identify research that is reproducible identify work that has been reproduced understand if results are supported elsewhere		understand reproducibility of a publication understand the reproducibility of proposed research

Table 2: Execute-stage desired outcomes that are unique to the Discovery context, shared between contexts, and unique to the Committee context. Only groups that occur in both the Discovery and Committee contexts are included.

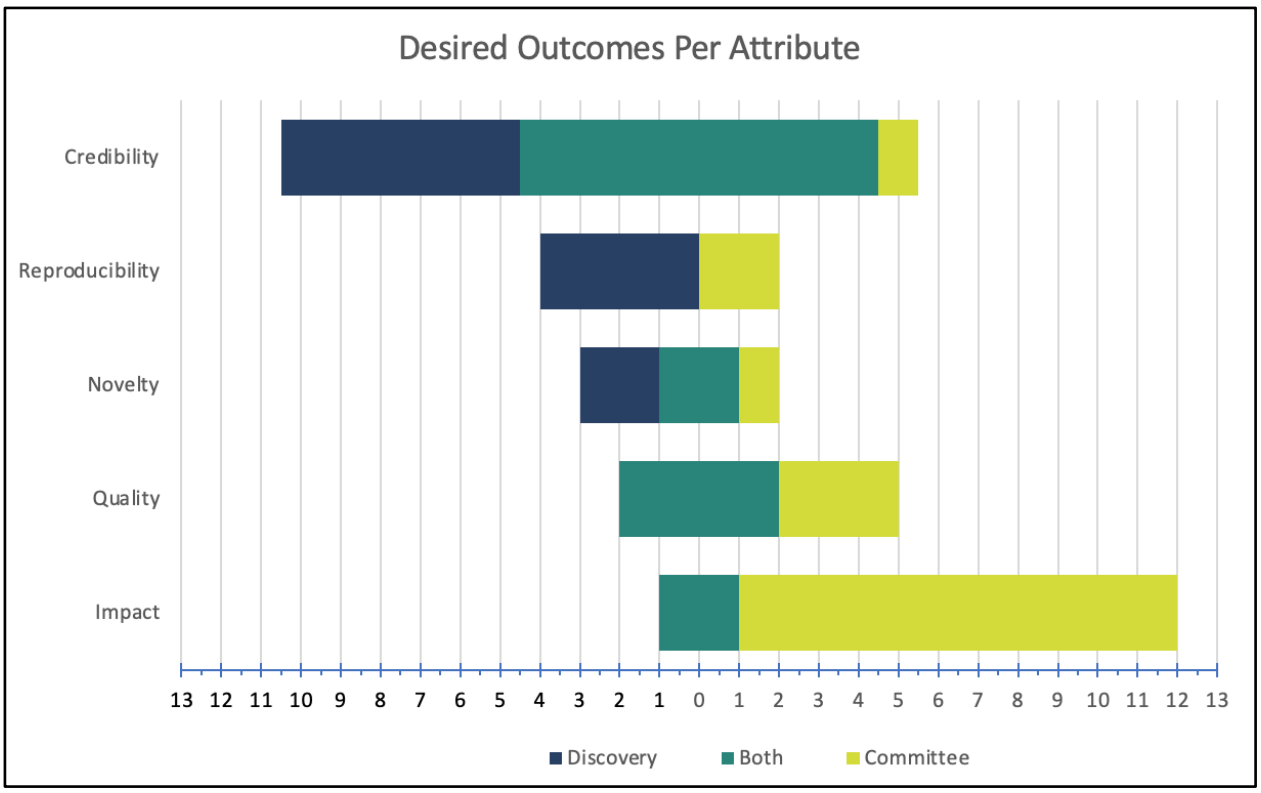

Figure 3: Count of identified desired outcomes unique to the Discovery context (blue), shared between contexts (green), and unique to the Committee context (yellow).

Reviewing Research Outputs

We also identified a Review Research Outputs group (not represented in Table 2 and Figure 3) that captures many of the basic tasks related to reading and reviewing publications, but relatively few of these tasks were associated with specific desired outcomes. Instead they were often precursors to assessments of credibility, quality, novelty, or even importance.

There was, however, a desired outcome of note here, which is shared across all the contexts we investigated, namely to "minimize time spent reviewing publications". Researchers consistently utilized strategies to minimize the amount of time they needed to spend reading and understanding publications, including skimming abstracts, selective reading of various sections of a manuscript, and various forms of screening. Given its importance at the level of these basic tasks, we might extrapolate that researchers generally desire to minimize time spent in the enterprise of assessment as a whole, even though they cite time savings less frequently when discussing specific assessment criteria.

Credibility

Members of all three cohorts described a substantial number of desired outcomes related to assessing the credibility of research outputs. Credibility-related outcomes were identified based on references to “credibility” “trust”, “reliability”, or proper execution and methodology. As expected, the credibility of research outputs was a central concern for researchers in the Discovery context,

especially when they planned to reuse a given research output. However, credibility-related desired outcomes were common in the Committee contexts as well.

While researchers in the Discovery cohort identified the greatest number of outcomes and tasks related to assessing the credibility of publications, the majority of credibility-related outcomes were shared across contexts, suggesting a significant overlap of researcher goals with respect to credibility. The desired outcome, “understand publication credibility,” was identified for each cohort, and was often, though not always, related to tasks that implemented shortcuts or proxies for understanding credibility or trustworthiness. Many of these tasks mirror the kinds of proxies e.g. author reputation, journal reputation, or impact factor and metrics identified by Nicholas, Tenopir and colleagues [ref. 9]. These proxies represent judgments made about a research output based on criteria that are fundamentally extrinsic to the output. However, researchers also utilized a number of shortcut methods of personal inspection, which focused on factors intrinsic to the research output, including evaluating *presentation* of the data or figures, or whether data was made available.

Interestingly, participants in the Discovery cohort were as likely to report relying on proxies or shortcuts as researchers in the Committee cohorts. Participants from both contexts reported a variety of tasks that assessed publication credibility based on the reputation, prestige, or impact factor of the publishing journal. Sometimes this judgment was attributed solely to the publishing journal, but it was often mediated through the researcher's perception of the quality or rigor of the journal’s peer review process, such that they felt they could trust outputs that were published in trusted journals. This later method is captured in the desired outcome “understand level of peer review received”. Some participants reported that it was enough that the research *was* peer reviewed for them to regard it as trustworthy.

All three cohorts also reported assessing credibility via personal inspection of the research output, for example by examining the soundness of the methods or research design, the quality of the statistics, or whether the conclusions were well supported. Finally, participants from all three cohorts also shared a pronounced tendency to assess publications by assessing the figures, and sometimes *only* the figures.

A few modes of assessing credibility were unique to the Discovery context. Here, researchers reported using external commentary to evaluate credibility in a few different ways, including analyzing published peer review reports and checking for negative post-publication commentary. The Discovery cohort was also unique in its use of author and laboratory reputation to assess credibility. In the Grant cohort the author’s reputation was used when applications were triaged in the Confirm stage, but did not figure in the assessment of credibility.

Researchers identified significantly fewer outcomes around preprints as compared to publications, despite being prompted to address them during the interviews. Nonetheless, all three cohorts identified a high-level desired outcome to understand the credibility of preprints. Researchers in the Committee context were more likely to say that they carefully review preprints or conduct their own peer review of a preprint than researchers in the Discovery group.

There are a couple of factors in the Committee context which may explain this difference in engagement. First, insofar as committee members were concerned to understand a candidate or

applicant's overall productivity and record, and saw preprints as a part of this, they may have been more inclined to engage with them more deeply. Further, committee members, in particular on grant committees, were concerned to identify the preliminary data related to a proposal, and acknowledged that this data was sometimes reported in preprints.

Along the similar lines as preprints, members of all three cohorts identified the desired outcome "understand data credibility". Within the Discovery and Grant cohorts, researchers indicated they were likely to assess the quality of the controls and analysis, and to infer the credibility of data from figures. Only researchers in the Discovery cohort reported engaging with the raw data.

Quality

Participants in both contexts identified a significant number of outcomes related to "quality," though these were somewhat more prominent in the Committee context. Like outcomes related to credibility, "quality" outcomes represent assessments of the research output as such, as opposed to its external impact on the field or world at large. However, unlike "credibility" outcomes, these tasks were not explicitly associated by researchers with "trust", "relying on", or potential for reuse, and represent more general assessments of whether an artifact is "good", "quality", or otherwise worth engaging with. Given these fuzzy boundaries, it is likely that "quality" is also sometimes encapsulating elements of other criteria, e.g. credibility, impact, importance, novelty, or innovation. However, as used by the participants in our interviews, it was distinct enough to merit separate treatment here.

All of our cohorts identified a desired outcome to "understand publication quality," which serves to describe a broad assessment of whether a publication is "good". While criteria like "good" or "quality" sometimes stood alone, unelaborated, they were sometimes associated with research that was well-executed, "excellent", "substantial", "complete", or "compelling".

Further, all three cohorts shared an explicit goal of minimizing the time spent on assessments of quality. This may explain why, despite "quality" representing an assessment of a specific research output, it was (like credibility) often proxied in both the Discovery and Committee context using either the publishing journal or journal level metrics, i.e. impact factor. In many cases these inferences were bound up with assumptions about peer review and selectivity, and the journal or journal impact factor served as an indication that others had already judged this work to be worthwhile. Article-level metrics, specifically citation count were noted as a means of assessing quality in the Committee contexts, but not in Discovery.

At the same time, all three cohorts also described using personal inspection of a research output to assess its quality, though the depth of this engagement varied from superficial judgments about the writing quality and presentation of the publication, to more detailed assessments of the research question and methodology.

All three cohorts shared a desired outcome of understanding preprint quality, though, as with preprint credibility, the data here was limited. In the Committee context, preprints were important as a source of preliminary data. Otherwise, their quality tended to be assessed not based on the preprint itself, but rather on whether, and how quickly, a preprint was ultimately published in a peer reviewed journal. Committee members expressed suspicion of preprints that were not ultimately published, and of candidates who had too many as-yet unpublished preprints.

Participants in all three cohorts identified tasks around understanding the quality of the data, either associated with a publication, or in the context of a publicly available data set.
As distinguished from data credibility, these outcomes focused on whether the data is presented well, or whether it is compelling, as opposed to whether it is trustworthy. In comparison to other outputs, proxies were less commonly used to evaluate the quality of data. Only the Grants cohort reported the use of proxies like journal name and impact factor. The Discovery and Hiring cohorts were more likely to engage directly with the data, often via figures, which emerge throughout our analysis as a dominant assessment shortcut.

Novelty

Desired outcomes in the Novelty group were adjacent to but distinct from those in the Impact group. Assessments of novelty didn't carry the burden of having advanced the field, changed scientific practice, or being "important". Rather, in these cases, researchers were concerned to identify research that had something new or innovative to offer, regardless of impact. Researchers in both the Discovery and Committee contexts cited goals related to identifying publications with new methodologies and which demonstrated original thinking. Overall, the Discovery cohort demonstrated a greater variety of approaches to assessing novelty, and utilized both proxy solutions like journal and impact factor to judge whether a publication was likely to be novel, as well as personal inspection of the output. The Committee cohorts tended to assess novelty based on their own judgment without relying on proxies.

Areas of Marginal Overlap

Impact

While there was some overlap in desired outcomes within the Impact group, the vast majority of outcomes identified were unique to the Committee context. We did not identify any impact-related outcomes unique to the Discovery context.

The overlapping outcomes related to identifying publications that were "important" or which "advanced the field". Surprisingly, when it came to impact, the Discovery cohort was relatively *more* likely to make judgments based on proxies like journal, journal metrics, or citations. Committee members were relatively more likely to report engaging with the research, if only the abstract, to make a judgment of its importance, often couched in terms of the significance of the research question, the potential impact of the work, and whether it had changed subsequent scientific practice. That said, committee members did also report using journal and journal metrics to make assessments of importance.

Overall, researchers in the Committee context had a more varied and nuanced view of impact, with a greater variety of desired outcomes as well as a greater variety of current solutions for achieving these. A few points of interest:

- Hiring committees reported a unique desired outcome to identify "visible" publications, typically by identifying publications in venues where they are likely to have been read. These venues were often, but not always, determined using journal impact factor.

- Grant committee members reported a unique desired outcome around identifying work with clinical and societal impact.
- Committee members were more likely to assess impact or importance based on whether and how research had been reused. Tasks evaluating reuse were associated not only with publication impact, but also the impact of publicly shared code and data set.

There was some evidence that importance also serves as a pre-assessment filter in the Discovery context. The outcomes included in the Confirm job stage suggest that researchers are sometimes screening outputs for importance or impact, typically using citation counts, before deciding which outputs to assess more carefully. There are also indications that researchers in the Discovery context may utilize search strategies intended to identify important research outputs. The role of impact and importance in searching and screening should be more fully explored in subsequent survey work.

Reproducibility

While we identified outcomes related to reproducibility in both contexts, none of them overlapped. The Committee context included some tangential references to reproducibility, however, these were qualitatively different, and less numerous than those in the Discovery cohort.

In the Discovery context, participants reported directly assessing reproducibility, often by attempting to confirm the results themselves, or by attempting to discern if someone else had already reproduced or successfully built upon the work. These assessments of reproducibility were largely pragmatic, and related to assessments of credibility. Some researchers reported wanting to ensure that they, or someone, could reproduce the results of research before incurring the cost of building upon that research. One researcher noted that it was a time investment but “that will always outweigh jumping in straight away without checking, building upon something just to find out months, or in the worst case years later, that all you've done, which you've done in a as good as you could way, was based on something that wasn't right. And then you wasted even more. So that to me is always the scenario that I try to avoid.”

Despite direct queries from interviewers on the topic, reproducibility was only occasionally acknowledged in the Committee context, and then only as a formality. Relatively few tasks were identified in which participants acknowledged actively evaluating reproducibility in the course of assessment. Those that were documented focused on the question of whether a publication or proposal *appeared* reproducible, for example based on the sample size and statistical methods, or the use of multiple, synergistic methods. Sometimes, the question of reproducibility in the Committee context was reputational, as expressed by a Hiring participant who noted that a reputation for non-reproducible results would “kill your ability to get grants”.

Outcomes Unique to the Committee Context

The Committee context included a number of groups of outcomes that did not occur in the Discovery context. This is unsurprising, as most of these groups centered around either research proposals, which we might expect to be of minimal interest to researchers in the Discovery context, or else

productivity and the publication record, which represent aggregate assessment at the level of the individual rather than the output.

Proposal-related Outcomes

In our analysis, we treated the research proposal as a distinct output, just as with preprints or data, and classified desired outcomes and tasks related to the proposal separately. There was no direct intersection between the Committee and Discovery contexts, as this research artifact is not routinely available in the Discovery context.

It is worth noting here that proposals were evaluated along many of the same dimensions as publications, preprints, or other outputs from completed research. Our analysis identified groups of outcomes around credibility, quality, novelty, and impact, paralleling the groups identified for other research outputs across the Committee and Discovery contexts. Within these groups, we saw similar assessment criteria to those identified for publications, e.g. identifying proposals with sound methodologies, good research design, and which were well supported by the existing literature. Obviously many of the proxies used to assess publications are not applicable to proposals, though there was some evidence that committee members would triage applications based on author and lab reputation.

Publication Record and Productivity Outcomes

The Committee context reported a relatively large number of outcomes and tasks centered on understanding the quality, credibility, or impact of an applicant's publication record as a whole, as well as the magnitude and consistency of output implied by that publication record. These aggregate measures seemed to be central to the Committee context, and included the largest number of unique tasks and outcomes, edging out the Impact group. There was no analogue to these outcomes in the Discovery context.

The Productivity group includes tasks that assess the magnitude and consistency of a candidate's research output, typically measured in terms of publication counts. This count is sometimes limited to certain kinds of publications, e.g. first author, or those in reputable journals. Consistency was a common theme among participants, and gaps in published output are often considered to be suspect. Preprints and other non-publication outputs were cited infrequently in terms of counts.

The Publication Record group captures a second large group of outcomes which correspond to the overall quality, credibility, or impact of a candidate or group's research output. Much as with the quality group, it is not always explicit what, precisely, is being assessed when participants report looking for a "good" publication record. These assessments were often described strictly in terms of the "proxy" measure used, e.g. "are these people publishing in the best journals in their subfield?". In these indeterminate cases an outcome of "understand publication record holistically" has been attributed.

Though a number of participants reported including preprints or other non-publication outputs in these aggregate assessments, our qualitative data suggests that the publication record is most often assessed based on the publishing journal and journal metrics. Aggregate citation metrics like h-index were also common in the Grants cohort.

Fundable Group

A group of outcomes centered around whether research was likely to secure funding was unique to the Hiring cohort. These outcomes were frequently concerned with ensuring research was compatible with the institution's funding goals, including the UK's REF assessment. Arguably the question of fundability is inherent to the entire grant assessment enterprise, and therefore not acknowledged explicitly as a task. Regardless, there is no corresponding group of outcomes in the Discovery context.

Discussion

Overlap between Contexts

We found significant overlap in how cell biology researchers assess credibility and impact in the Discovery and Committee contexts. It is important to keep in mind that our interview cohorts were comprised of researchers from a single field, and consequently one should use caution in generalizing to other fields in advance of work to test these findings more broadly.

We hypothesized that both credibility and impact matter in principle in both contexts but that the specific circumstances, motivations and practicalities of each context confer different relative importance to the two concepts. Credibility was expected to be relatively more important in the context of discovery, and impact to dominate in the context of grant and hiring assessments. Based on our qualitative data, we can make some provisional statements about these hypotheses.

Credibility matters in both contexts. There was significant overlap between contexts in the desired outcomes relating to the credibility of research outputs.

Impact matters in both contexts. We also identified impact-related desired outcomes in both contexts, though the distribution of these outcomes were heavily weighted toward the Committee context. Nonetheless, researchers in the Discovery context are considering impact and importance as part of their assessments. Further, there is some overlap between contexts in the impact-adjacent group of novelty outcomes.

Credibility is more important than impact in Discovery. This seems provisionally to be the case. There were far more credibility related outcomes identified for the Discovery context, and a far greater variety as well. It seems relatively unlikely that additional quantitative work will show that the importance of the limited number of impact outcomes identified outweighs the aggregate importance of the credibility outcomes.

Based on the existing literature [ref. 27, 3] we had expected to find that in the Committee context priority would be given to impact as measured by easily accessible metrics, such as journal impact factor. However, while it is true that the preponderance of impact-related outcomes are associated with the Committee context, the distribution of credibility and impact *within* the Committee context is not nearly as unbalanced as it was in Discovery.

Relevance of Credibility in Committees

Previous studies of what we have referred to as the Committee context, for example, those conducted as part of ScholCommLab's "Review, Tenure, and Promotion" project have tended to focus on institutional policy, whereas our work here has focused on researcher's practices when serving on committees. Our findings based on the researcher's perspective were in many ways aligned with earlier findings [ref. 4]. We confirmed that committee members continue to view the publication as the key unit of assessment, though there is some evidence that preprints are slowly gaining ground, especially when they are directly related to a research proposal. Further, it was clear from our interviews that impact is important, and often addressed via a variety of metrics including impact factor, citations, etc.

However, despite the centrality of impact in institutional guidelines [ref. 2,3], committee members in our study appear to spend a significant amount of time considering questions of credibility. While the importance of credibility should be further specified via survey work, it appears to be the case that committee members are spending more time making assessments of credibility than we would have expected, and are often explicitly concerned to ensure that whoever is hired or funded is doing their work "properly". This suggests the possibility that committee members are considering factors, i.e. credibility, which are not explicitly prioritized by guidelines.

Follow-up survey work would help to illuminate any relationship between the importance of credibility and the practical constraints of assessing it. If credibility is deemphasized in the Committee context due to practical constraints on its assessment, we would expect credibility-related outcomes overall to be rated as important but poorly satisfied. This would suggest that researchers value credibility, as such, but are spending less time on it because they are not satisfied with the available options for assessing it. In that case, better solutions for evaluating credibility could affect how often it is utilized in grant and hiring decisions.

Role of Proxy Criteria

Our results confirm those of previous studies, in particular the extensive work of Nicholas, Tenopir and colleagues, in that proxies based on factors extrinsic to the research outputs -- journal-level metrics, journal reputation, and researcher or laboratory reputation -- are prevalent in the evaluation of credibility as well as quality and impact [ref. 3]. The present study extends Nicholas and Tenopir's findings insofar as all of these criteria, with the exception of author reputation, were found to be relevant in the Committee context as well, which was not in scope for their study.

Moreover, we identified a similar number of references to proxy use in each of our cohorts. If we exclude proposal-related outcomes, which are rarely proxied and have no equivalent in the Discovery context, 20-25% of interview extracts from each cohort describe the use of proxy methods. Additional survey work is needed to understand if this actually means proxy methods are equally important in both contexts, or if those tasks are significantly more *important* in one context than the other despite their similarity in number.

Journal name and reputation were especially common proxies. Participants in all contexts reported using journal name and reputation as proxies to assess multiple attributes, including credibility,

quality, impact and to a lesser extent, novelty of publications. They regarded these proxies as signals of reliable or rigorous peer review, echoing the centrality of peer review reported by Nicholas and Tenopir [ref. 10]. Many participants reported trying to avoid using the journal impact factor, however the mechanisms invoked suggest their judgments were still related to a competitive publication landscape dominated by highly selective and prestigious journals. For example, participants suggested that surviving a competitive submission process for a prestigious journal, which rejects most submissions, suggested a work was of high quality. Many participants who reported avoiding the impact factor were relying instead on a personal list of quality or reliable journals.

Furthermore, participants mentioned journals as a way to infer authors' own assessment of their work--expressing a belief that researchers tend to publish 'quality' work in 'good' journals. The inference being that not publishing in a good journal suggested the authors didn't think it was quality work or else it had been rejected by more discerning journals and was therefore suspect. This tracks with the previous findings that journal impact factor matters more when deciding where to publish than when deciding what to read or cite [ref. 9].

In the Committee context, researchers frequently reported using journal-based proxies, especially when examining the publication record as a whole. However, in some cases, for example assessments of the impact of a specific publication, they were more likely to report relying on personal inspection than researchers in the Discovery context. This was often framed in terms of the significance of the research question, the potential impact of the work, or whether it had changed subsequent scientific practice.

As was the case in the studies of Nicholas and Tenopir, participants also reported relying on personal inspection to evaluate credibility (and in our study, reproducibility). Researchers relied on examination of specific elements of a publication, for example the soundness of the methods or research design, the sample size, the quality of the statistics, whether the conclusions were well supported, or the use of multiple, synergistic methods. Researchers also evaluated presentation of the data or figures, or whether data was made available.

Open Science and Assessment

Overall, our participants continued to rely on traditional markers of trust and quality, much as those in Nicholas and Tenopir's studies had. We were particularly interested in understanding if open science practices, such as the sharing of more diverse research artifacts and transparency in the publishing process, which have been increasing since that work was completed, were becoming part of evaluations. While open science practices rarely came up in our interviews as desired outcomes or formal criteria, they did come up as a solution--a way of accomplishing another goal:

- As found in previous research [ref. 15], transparency, for example sharing of data or code, was generally associated with credibility (for the dataset and/or publication). Having made data open, or failing that being willing to share that data, was seen as a signal that the data and associated publication were trustworthy in both the Discovery and Committee contexts.
- In the Discovery context, transparency was sometimes connected to facets of reproducibility, with one participant noting "you can really believe it, if everything is presented in the right

way providing the data...sometimes it's just so impossible to repeat what they were having published because there are no details at all."

- In the Committee context, researchers reported assessing impact or importance based on whether and how research had been reused. Tasks evaluating reuse were associated not only with outcomes around publication impact, but also the impact of publicly shared code and datasets.
- Discovery participants reported utilizing open peer review reports to better understand publications and their credibility. Participants noted that the availability of peer review reports gives "more credibility to the paper" and helped to fill in gaps in their own knowledge.

While the publication (along with the proposal) remained the dominant unit of assessment, some participants assumed preprints were as valid as any other output and ought to be included in assessments of productivity. Some participants felt strongly that they should be included in assessments of candidates, including one who reported "trying to convince other people on panels that that's a sensible, viable way to disseminate work early on."

Participants in all contexts expressed concern that preprints had not been peer-reviewed, which a number of participants addressed by applying additional levels of scrutiny to preprints, sometimes framed in terms of doing one's own peer review on a preprint. In addition, preprint quality tended to be assessed based on whether, and how quickly, a preprint was ultimately published in a peer reviewed journal. This criteria appears somewhat more important in our study than in a recent survey around preprint credibility [ref. 4] perhaps because that work did not focus specifically on the committee context.

However, none of our participants in either context reported being opposed, in principle, to using, citing, or considering preprints. In many cases, there was an implication that they might use a preprint, but simply hadn't yet. The landscape around preprints is changing rapidly, and our work may not fully account for the impact, either positive or negative, of changing practices during the COVID-19 pandemic [ref. 31]. Consequently, it is plausible that follow up work may reveal greater utilization of preprints than our qualitative findings suggest.

The fact that elements like data sharing and reuse, peer review reports, and personal scrutiny of preprints are mentioned as means of assessing credibility and impact suggests the potential for new signals which might reduce time as compared to personal inspection methods, yet be better tailored for credibility and impact judgements. The critical importance of the publication record in the Committee context suggests that new signals would be more effective if they can be aggregated across the publication record, and expanded to preprints and other outputs.

Limitations

Our study has a number of potential limitations. Given our sample size, we can be more confident that we have identified most of the relevant desired outcomes than that we have identified the full variety of *tasks* related to these outcomes. There is likely to be less variation in researchers' goals than the solutions that they use to reach these goals. Therefore, this work cannot be regarded as a full accounting of tasks, and any comparisons at the task level are directional only. As noted above, additional survey work is needed to validate the relative importance of the outcomes described, and

the number of occurrences can't be taken as indicative of importance. We had a high proportion of PLOS-affiliated participants, given that most of our participants were recruited via PLOS email lists. This may contribute to a more general risk of responder bias wherein researchers who are motivated to participate in interviews about assessment might not be representative of researchers as a whole. Finally, it is possible that recruiting challenges related to the COVID pandemic may have also affected the diversity of our sample.

We have generated a list of standardized statements representing the desired outcomes we identified. This can serve as a basis for survey work in which survey participants are asked to rank desired outcomes in terms of importance and satisfaction. By identifying areas regarded as important but not well satisfied, we can identify areas where researchers are most likely to be intrinsically motivated to adopt novel solutions. Because we have generated standardized outcomes statements *across* the two contexts, we will be able to easily identify solutions that will be successful in both. At the same time, expanding the disciplinary scope to include other areas of life and medical sciences, and surveying across all career stages will allow us to determine if these findings apply more broadly. The larger sample size of such a survey would also ensure a more diverse set of participants and mitigate the risk of responder bias noted above.

Conclusions

We have identified substantial overlap between the goals of researchers when assessing research outputs in the context of researchers' own work (Discovery context) and when researchers participate in research assessment committees for grant review and hiring (Committee context).

The general trend emerging from this qualitative research, to be confirmed by quantitative research, is that both impact and credibility matter in each of the contexts we examined, with credibility being the dominant factor in the Discovery context and somewhat less represented but still important in the Committee context. In addition, researchers also assess attributes related to quality, novelty and reproducibility. Furthermore, researchers across all of the contexts utilized strategies to minimize the amount of time they needed to spend reading and understanding publications, including skimming abstracts, selective reading of various sections of a manuscript, and various forms of screening. While publications remain the dominant unit of research assessment, researchers in our sample also evaluate research data, code and preprints, in both contexts.

Our findings confirm previous studies in that researchers use proxies to evaluate research outputs. In our study, the use of proxies occurred with similar frequency in the Discovery and Committee context, and journal-based proxies were particularly prevalent to evaluate all key attributes of research outputs that we identified. Considering the documented flaws of these proxies, our work reinforces the opportunity to develop more reliable signals to improve evaluation. Our qualitative data suggests areas of further inquiry to identify more reliable signals, and also suggests that applying these signals to preprints as well as journal publication may be effective in encouraging their use.

The prevalence of assessments of credibility in current committee practice also suggests an opportunity for funders and institutions to better align their guidelines with the practice and motivations of committee members.

The existence of a significant overlap in the goals of researchers in the Discovery and Committee contexts, and in the attributes of publications that they are seeking to assess, is a valuable insight as we expect that solutions addressing a frequent and important need in the Discovery context, where researchers spend most of their time, are more likely to be broadly adopted and to influence behavior when researchers are in Committee contexts as well.

Data availability

The following outputs are available on Open Science Framework at https://osf.io/sphfj to support interpretation and reuse of our results:

- List of structured "desired outcome" statements for validation in further, quantitative phase of research (txt)
- Interview prompts (pdf)
- Screener Questionnaire (pdf)
- Underlying data for job map, including extracts from interviews (xlsx)
- Visualizations of the job map model for each cohort (pdf)

Interview transcripts and participant demographics are not publicly available to respect research participant privacy. Questions about research data availability should be sent to PLOS (research@plos.org) or jharney@plos.org.

Funding

This research was supported by a grant from the Alfred P. Sloan Foundation. (grant number G-2020-14053)

Acknowledgments

The authors would like to thank the American Society for Cell Biology (ASCB), and in particular Erika Shugart, Mark Leader, Brian Theil, and Mary Spiro, who consulted on messaging to the cell biology community, and assisted in recruiting through outreach and social media. We would also like to thank PLOS colleagues Helen McDonald, Susan Hagen, and Dan Morgan for their support of our recruiting efforts.

References

1. Jamieson, Kathleen Hall, Marcia McNutt, Veronique Kiermer, and Richard Sever. 2019. "Signaling the Trustworthiness of Science." Proceedings of the National Academy of Sciences 116 (39): 19231. https://doi.org/10.1073/pnas.1913039116.

2. Cassidy Sugimoto and Vincent Lariviere. 2018. *Measuring Research: What Everyone Needs to Know*®. What Everyone Needs To Know®. Oxford, New York: Oxford University Press.

3. Niles, Meredith T., Lesley A. Schimanski, Erin C. McKiernan, and Juan Pablo Alperin. 2020. “Why We Publish Where We Do: Faculty Publishing Values and Their Relationship to Review, Promotion and Tenure Expectations.” *PLOS ONE* 15 (3): e0228914. https://doi.org/10.1371/journal.pone.0228914.

4. Schimanski LA and Alperin JP. (2018) The evaluation of scholarship in academic promotion and tenure processes: Past, present, and future [version 1; peer review: 2 approved]. F1000Research 7:1605. https://doi.org/10.12688/f1000research.16493.1

5. National Academies of Sciences, Engineering. 2020. *Advancing Open Science Practices: Stakeholder Perspectives on Incentives and Disincentives: Proceedings of a Workshop–in Brief*. https://doi.org/10.17226/25725.

6. Christensen, Clayton, et al. 2016. Competing Against Luck: The Story of Innovation and Customer Choice. New York, NY: Harper Collins.

7. Ulwick, Anthony W. 2008. Jobs to be Done: Theory to Practice. Idea Bite Press.

8. Rieh, Soo Young. 2002. “Judgment of Information Quality and Cognitive Authority in the Web.” Journal of the American Society for Information Science and Technology 53 (2): 145–61. https://doi.org/10.1002/asi.10017.

9. Nicholas, David, Anthony Watkinson, Rachel Volentine, Suzie Allard, Kenneth Levine, Carol Tenopir, and Eti Herman. 2014. “Trust and Authority in Scholarly Communications in the Light of the Digital Transition: Setting the Scene for a Major Study.” Learned Publishing 27 (2): 121–34. https://doi.org/10.1087/20140206.

10. Tenopir, Carol, Suzie Allard, Ben Bates, Kenneth Levine, Donald King, Ben Birch, Regina Mays, and Chris Caldwell. 2010. “Research Publication Characteristics and Their Relative Values: A Report for the Publishing Research Consortium.” Center for Information and Communication Studies, University of Tennessee. Available at: https://trace.tennessee.edu/utk_infosciepubs/20.

11. Tenopir, Carol. 2014. “Trust in Reading, Citing and Publishing.” Information Services & Use 34 (1–2): 39–48. https://doi.org/10.3233/ISU-140725.

12. Thornley, C., Watkinson, A., Nicholas, D., Volentine, R., Jamali, H. R., Herman, E., et al. (2015). The role of trust and authority in the citation behaviour of researchers. Information Research, 20 (3) paper 677, http://InformationR.net/ir/20-3/paper677.html

13. Nicholas, David, Hamid R. Jamali, Anthony Watkinson, Eti Herman, Carol Tenopir, Rachel Volentine, Suzie Allard, and Kenneth Levine. 2015. “Do Younger Researchers Assess Trustworthiness Differently When Deciding What to Read and Cite and Where to Publish?” International Journal of Knowledge Content Development & Technology 5 (2): 45–63. https://doi.org/10.5865/IJKCT.2015.5.2.045.

14. https://www.nytimes.com/2020/04/14/science/coronavirus-disinformation.html

15. Soderberg, Courtney K., Timothy M. Errington, and Brian A. Nosek. 2020. “Credibility of Preprints: An Interdisciplinary Survey of Researchers.” R. soc. Open sci. 7 (10): 201520 https://doi.org/10.1098/rsos.201520

16. Polka, Jessica K., Robert Kiley, Boyana Konforti, Bodo Stern, and Ronald D. Vale. 2018. “Publish Peer Reviews.” Nature 560 (7720): 545–47. https://doi.org/10.1038/d41586-018-06032-w.

17. Stern, Bodo M., and Erin K. O’Shea. 2019. “A Proposal for the Future of Scientific Publishing in the Life Sciences.” PLOS Biology 17 (2): e3000116. https://doi.org/10.1371/journal.pbio.3000116.

18. “Power to the Preprint.” 2018. The Official PLOS Blog (blog). May 1, 2018. https://theplosblog.plos.org/2018/05/power-to-the-preprint/.

19. “PLOS Authors Say ‘Yes’ to Preprints.” 2018. The Official PLOS Blog (blog). December 6, 2018. https://theplosblog.plos.org/2018/12/plos-authors-say-yes-to-preprints/.

20. “Peer Review: New Initiatives to Enhance the Value of ELife’s Process.” 2019. ELife. eLife Sciences Publications Limited. November 7, 2019. https://elifesciences.org/inside-elife/e9091cea/peer-review-new-initiatives-to-enhance-the-value-of-elife-s-process.

21. “Partnering to Streamline Review.” 2019. The Official PLOS Blog (blog). September 30, 2019.https://theplosblog.plos.org/2019/09/partnering-to-streamline-review/.

22. Wilsdon, James, Liz Allen, Eleonora Belfiore, Philip Campbell, Stephen Curry, Steven Hill, Richard Jones, et al. 2015. “The Metric Tide: Report of the Independent Review of the Role of Metrics in Research Assessment and Management.” https://doi.org/10.13140/RG.2.1.4929.1363.

23. Xu, F., Li, X. The changing role of metrics in research institute evaluations undertaken by the Chinese Academy of Sciences (CAS). Palgrave Commun 2, 16078 (2016). https://doi.org/10.1057/palcomms.2016.78

24. Konkiel, S. Altmetrics: diversifying the understanding of influential scholarship. Palgrave Commun 2, 16057 (2016). https://doi.org/10.1057/palcomms.2016.57

25. Oancea, A. Research governance and the future(s) of research assessment. Palgrave Commun 5, 27 (2019). https://doi.org/10.1057/s41599-018-0213-6

26. Hicks, Diana, Paul Wouters, Ludo Waltman, Sarah de Rijcke, and Ismael Rafols. 2015. “Bibliometrics: The Leiden Manifesto for Research Metrics.” Nature News 520 (7548): 429. https://doi.org/10.1038/520429a.

27. Moher, David, Florian Naudet, Ioana A. Cristea, Frank Miedema, John P. A. Ioannidis, and Steven N. Goodman. 2018. “Assessing Scientists for Hiring, Promotion, and Tenure.” PLoS Biol 16 (3): e2004089. https://doi.org/10.1371/journal.pbio.2004089.

28. Witteman, Holly O., Michael Hendricks, Sharon Straus, and Cara Tannenbaum. 2019. “Are Gender Gaps Due to Evaluations of the Applicant or the Science? A Natural Experiment at a National Funding Agency.” The Lancet 393 (10171): 531–40. https://doi.org/10.1016/S0140-6736(18)32611-4.

29. Li, Danielle, and Leila Agha. 2015. “Big Names or Big Ideas: Do Peer-Review Panels Select the Best Science Proposals?” Science 348 (6233): 434–38. https://doi.org/10.1126/science.aaa0185.

30. Myers, K.R., Tham, W.Y., Yin, Y. *et al.* 2020. Unequal effects of the COVID-19 pandemic on scientists. *Nat Hum Behav* 4, 880–883. https://doi.org/10.1038/s41562-020-0921-y

31. Fraser N, Brierley L, Dey G, Polka JK, Pálfy M, Nanni F, et al. 2021. The evolving role of preprints in the dissemination of COVID-19 research and their impact on the science communication landscape. PLoS Biol 19(4): e3000959. https://doi.org/10.1371/journal.pbio.3000959

Appendix A: Job Stage Definitions

We adopted our standardized job stages from the Outcome-Driven Innovation framework. Structuring interviews around these stages helps ensure that all relevant steps related to assessment are uncovered in the course of the interview. During the analysis, sorting outcomes into these standardized phases lets us more easily highlight alignment between cohorts, or lack thereof. For purposes of this study, these stages were operationalized as follows:

Define: Define the context that the assessment is occurring in

- Own research: Define the research project that I am going to be assessing research in support of
- Hiring/Grants: Define or receive the expectations for the grant or hiring search

Locate: Gather materials for assessment

- Own research: Search for and gather research outputs
- Hiring/Grants: Receive applications; Search for and gather research outputs

Prepare: Get ready to do the assessment

- Own Research: Assign work within group; Background reading
- Hiring/Grants: Assign work within committee; check for conflicts of interest

Confirm: Prioritize the outputs to be assessed

- Own Research: Screen outputs gathered in Locate phase
- Hiring/Grants: Triage applications; Prioritize applicant's research outputs.

Execute: Assess research outputs

Monitor: Summarize and check my assessment

- Own Research: Check that my assessment is complete
- Hiring/Grants: Assign initial score; Confirm my initial assessment;

Modify: Adjust my initial assessment based on new information

- Own Research: Consult colleagues; Adjust level of scrutiny
- Hiring/Grants: Consult colleagues; Revise assessment based on input from others; Contextualize based on situation of applicant

Conclude: Finalize the assessment

- Own research: Decide what I will actually cite or re-use
- Hiring/Grants: Discuss in committee; Submit scores

Appendix B: Desired Outcome Statements

Typically in the Outcome-Driven Innovation framework, a desired outcome statement is composed of a direction of change, a metric of change, an object of change, and an optional context. For example,

“Minimize time to review publications when serving on a committee”

where “minimize” is the direction, “time” is the metric, “review publications” is the object, and “when on a committee” is the context. However, in many cases the text of our interviews did not clearly imply a direction or metric. This is because the assessment job itself is to some extent about mental states like judgment and success which are not as readily “measured” as tasks relating to physical objects might be.

Rather than impose a generic metric like “likelihood,” we have forgone the direction and metric when it cannot be clearly inferred from the text. While the resulting statements diverge from standard ODI practice, it should not impact how well they can be tested in survey work, or their usefulness in identifying researcher needs.

Further, in most cases, the context is implicit -- when assessing research as part of a committee, or own research, etc. -- and is not explicitly included in the desired outcome statement.

Key terms used in desired outcome statements

To the extent possible, desired outcome statements have been drafted using researchers' own terminology. There are, however, a few key terms that have been introduced for the sake of uniformity, according to the following rubric

“Credibility” includes references to concepts like

- “credible”
- “trust”
- “done properly”
- “soundness”
- “reliable”, “rely on”, etc

“Quality”

- Explicitly to “quality” without reference to credibility, impact, importance, novelty, etc.
- References to research that is “good”, “bad”, etc.
- Like credibility, quality was assessed with reference to the research artifact itself, rather than its external impact.

“Impact”

- Effect of research on the external world, including “the field”, “science”, “society”, “clinical practice”, etc.

“Novelty”

- Explicit reference “novelty”, “novel”, “new”
- “Innovation” was included in the Novelty discussion below, but is treated as a distinct criteria, as it tended to be more bound up with notions of impact

“Understand”

- Used where the participant wishes to form an overall judgement about the extent to which a particular attribute is displayed. A continuous variable. If I “understand publication quality”, there are, in theory, infinite shades of “good, better, best”.

“Identify”

- Determine which of the research outputs being assessed have a given attribute. These are akin to a boolean variable. For example, if I want to “identify credible research”, everything is sorted into either credible, or not credible.