U.S. National Science Foundation
Evaluation and Assessment Capability

A Review of Funder Instructions and Grant Reviewer Practices for Assessing the Intellectual Merit and Other Impacts of Research

A report from the Evaluation and Assessment Capability Section of the U.S. National Science Foundation.

About the Evaluation and Assessment Capability Section

The Evaluation and Assessment Capability (EAC) Section bolsters NSF efforts to make informed decisions and promote a culture of evidence. Located in the Office of Integrative Activities of the Office of the Director, EAC provides centralized technical support, tools, and resources to conduct evidence-building activities and to build capacity for evidence generation and use across the agency. EAC is led by NSF's Chief Evaluation Officer.

About this report

This report was prepared for EAC under contract number 49100421D0011, order number 49100423F0104. The views expressed are those of the authors and should not be attributed to NSF, nor does mention of trade names, commercial products, or organizations imply endorsement of same by the U.S. Government.

Preferred citation

Chandler, Jesse, Emily Rosen, Kimberley Raue, Katlyn Lee Milless and Danielle Rockman. 2024. *A Review of Funder Instructions and Grant Reviewer Practices for Assessing the Intellectual Merit and Other Impacts of Research*. Alexandria, VA: U.S. National Science Foundation.

NSF Quality Certification: 2

Quality Certifications

Level 1 — The author(s)/contractor(s) are responsible for the quality and conclusions presented in this report

Level 2 — NSF verified that this report underwent quality assurance procedures and contributed to assessing its content

Level 3 — NSF independently reproduced the analysis presented in this report

A Review of Funder Instructions and Grant Reviewer Practices for Assessing the Intellectual Merit and Other Impacts of Research

February 28, 2024

Jesse Chandler, Emily Rosen, Kimberley Raue, Katlyn Lee Milless, and Danielle Rockman

Submitted to:

U.S. National Science Foundation
2415 Eisenhower Ave.
Alexandria, VA 22314

Submitted by:

Mathematica
1100 First Street, NE, 12th Floor
Washington, DC 20002-4221
Phone: (202) 484-9220
Fax: (202) 863-1763

Abstract

Researchers and policymakers have noted the challenges associated with assessing the broader impacts of scientific research. This report reviews the existing literature on how reviewers at institutions outside of the U.S. National Science Foundation (NSF) assess nonmedical basic and use-inspired research and science, technology, engineering, and mathematics (STEM) education. In doing so, it provides evidence about the criteria used to assess grants and factors having important, unimportant, or unstudied impacts on the evaluation of the intellectual merit and broader impacts of grant proposals. We found that when evaluating intellectual merit, funders and grant reviewers seem to consider both the potential for scientific discovery and the plan for scientific inquiry. In contrast, when evaluating broader impacts, funders and reviewers seem to focus on outcomes, paying less attention to the methods by which these outcomes might be realized. The empirical literature shows that reviewers often do not agree about the merit of grant proposals, though they seem to become more consistent with experience. Several ideas on how to support, modify, or replace peer review have been proposed in the literature, but evidence about the efficacy of these ideas is limited. The findings of this literature review will inform the design of a process evaluation that will assess how NSF applies its Broader Impacts review criterion across its work.

Contents

Exhibits

Executive Summary

The U.S. National Science Foundation (NSF) is a federal executive branch agency with the directive "to initiate and support basic scientific research and programs to strengthen scientific research potential and science education programs at all levels" (The National Science Act 1950). In fiscal year 2022, NSF advanced this mission by evaluating almost 40,000 proposals for research and education activities, and making nearly 11,000 new awards totaling more than $8.5 billion (U.S. National Science Foundation 2022).

NSF evaluates proposals using two review criteria approved by the National Science Board: Intellectual Merit and Broader Impacts (U.S. National Science Foundation 2023). The Intellectual Merit review criterion encompasses the potential for a project to advance knowledge. The Broader Impacts review criterion encompasses the potential for a project to benefit society and achieve specific, desired societal outcomes. NSF relies on the expertise of program directors and the input of expert peer reviewers to assess the Intellectual Merit and Broader Impacts of proposals.

Most funders of basic scientific research use processes similar to those NSF uses. Proposals are generally assessed according to how they will advance knowledge in their field (analogous to Intellectual Merit) and what benefits this knowledge might have to others (analogous to Broader Impacts). However, funders have very different definitions of the benefits and beneficiaries that should be considered. Most funders rely on review by subject matter experts to evaluate proposals, with variations in the frequency and processes by which feedback is collected from peers, the instructions provided to reviewers, and the factors beyond peer review that can be considered as a part of the decision process.

This literature review is part of a congressionally mandated evaluation to "assess how [NSF's] Broader Impacts review criterion is applied across the Foundation and make recommendations for improving the effectiveness for meeting the goals established in section 526 of the America Creating Opportunities to Meaningfully Promote Excellence in Technology, Education, and Science Reauthorization Act of 2010 (42 U.S.C. 1862 p-14)" (Creating Helpful Incentives to Produce Semiconductors [CHIPS] for America Fund Act 2022). This review will provide context for the evaluation, which includes a review of NSF documents; interviews and focus groups with NSF staff, principal investigators (PIs), and reviewers; and an analysis of survey and review data.

Mathematica examined the grant proposal evaluation processes of seven funders and reviewed 104 studies about practices for assessing concepts related to Intellectual Merit and Broader Impacts by agencies outside of NSF. This review will provide context for the evaluation, which includes a review of NSF documents; interviews and focus groups with NSF staff, principal investigators (PIs), and reviewers; and an analysis of survey and review data. The literature review

highlights practices across funders with similar review criteria and explains the evidence for their efficacy. It also highlights effective strategies for mitigating concerns about reviewer bias and subjectivity. The literature review will inform the data collected through interviews and focus groups, and guide the analysis of survey and review data.

Key findings indicate the following:

- Funders generally defer to reviewers about how to assess the merit of grant proposals. Except for the Netherlands Organisation for Scientific Research (NWO), the funders we reviewed do not specify how much weight reviewers should assign to the intellectual merit and to the broader impacts of research. Although funders provide criteria by which intellectual merit and broader impacts should be assessed, the criteria are not defined precisely and the relative emphasis placed on them is left to the reviewer's discretion.
- Among the seven funders we reviewed, criteria for intellectual merit usually direct reviewers to consider research methods and potential outcomes, contextualized by the applicant's professional experience. In contrast, funders' criteria for broader impacts usually direct reviewers to consider potential outcomes and are less likely to require input on the proposed methods to achieve outcomes or any relevant previous experience of the applicant.
- Across six large studies of research grant proposals, the overall reliability of reviewer scores is low. At the same time, the differences in proposal scores between funded and unfunded proposals are quite small, suggesting that the reviewers assigned to a proposal will play an important role in these scores. No studies examined how much a reviewer's score might vary based on contextual factors.
- Research has suggested many sources of variability across reviewers. Some variability is caused by genuine differences in opinion and contributes to a complete understanding of the strengths and weaknesses of a grant proposal. Other sources of variability are caused by errors and oversights in the review process, and decrease the reliability of reviewer scores without providing information relevant to the decision process. For example, reviewers unfamiliar with the review process may not apply criteria correctly, strategies to manage workload may lead reviewers to overlook information that might change their score, and reviewers tend to differ in their overall harshness or leniency when evaluating grant proposals.
- It is unclear whether reviewers are more likely to agree about certain criteria or elements of proposals. For example, the empirical literature does not provide any evidence on whether consensus is higher for assessments of intellectual merit or broader impacts, or whether it is higher for the rigor of a proposal's technical approach or the importance of the potential outputs.

- We did not identify studies concerned with the predictive validity of proposal scores. Very little information is available about the validity of the review process (that is, whether reviewers can identify projects especially likely to have impact). This question is difficult to answer for both methodological reasons and because the concept of impact is difficult to define regarding the relative value of contributions and how long it might take for these impacts to occur.
- Some funders have adopted review practices that allow for more interaction between applicants and reviewers. NWO allows applicants to respond in writing to comments after the first round of review and requires an oral presentation as a part of the funding process. The United Kingdom Research and Innovation Engineering and Physical Research Council sometimes brings proposers, research users, and other experts together in interactive "sandpits," where proposals can be developed with real-time feedback. Program officers at philanthropic organizations often work with applicants to develop the plans and desired outcomes of proposals.
- Researchers have suggested many tools to support the proposal review process, as well as alternative review methods that might supplement peer review. The Hungarian Scientific Research Fund uses an automated bibliometric scoring system that adjusts scores to account for disciplinary differences and career level. Several researchers have proposed decision aids to help integrate sub scores into an overall proposal score. One study found that program officers saw promise in this tool, but its impact on proposal decisions has not been evaluated.
- Alternative review methods like partial lotteries and crowdfunding are either in the early phases of conceptual development or being piloted by funders.

1. Introduction

In 1950, the National Science Foundation Act created an independent executive branch agency with the directive "to initiate and support basic scientific research and programs to strengthen scientific research potential and science education programs at all levels" (The National Science Act 1950). The U.S. National Science Foundation (NSF) is the only U.S. federal agency whose mission is to invest in fundamental, basic research and education across the full spectrum of science, technology, engineering, and mathematics (STEM) disciplines, with the exception of the medical sciences. NSF achieves its unique mission by making merit-based awards to around 1,900 colleges, universities, businesses, informal science organizations, and other research organizations throughout the United States. In fiscal year (FY) 2022, NSF evaluated almost 40,000 proposals for research and education activities, making nearly 11,000 new awards totaling more than $8.5 billion (U.S. National Science Foundation 2022).

Organizations submit proposals for new projects to NSF, which are then evaluated using two review criteria approved by the National Science Board (NSB): Intellectual Merit and Broader Impacts[1] (U.S. National Science Foundation 2023). The Intellectual Merit review criterion encompasses the potential for a project to advance knowledge. The Broader Impacts review criterion encompasses the potential for a project to benefit society and achieve specific, desired societal outcomes. Solicitations for proposals might contain additional NSF-specified review criteria particular to the goals and objectives of the program.

NSF program directors, knowledgeable experts in both technical and programmatic areas, lead the review of submitted proposals and recommend which projects NSF should fund. They share most proposals with three to five external reviewers chosen for their relevant disciplinary expertise. Reviewers are asked to describe the strengths and weaknesses of proposals related to NSF's merit review criteria. Reviewers provide feedback either through written comments sent directly to NSF or a combination of individual written comments, and a summary and recommendation made collectively with a panel of other reviewers. Some proposals receive feedback through both methods. NSF program directors consider external reviews along with other factors related to portfolio composition and NSF's strategic goals. Based on the results of this analysis, they make award recommendations for final approval by their division director.

NSF's directorates, divisions, and programs operationalize the merit review process in varied ways based on disciplinary conventions, solicitation and programmatic requirements, availability of staff and reviewers, and other factors. This process may result in differences in how research

[1] In this report, we follow NSF terminology, using the term "criterion" to refer to the Intellectual Merit criterion and Broader Impacts criterion, and the term "element" to refer to a combination of an object of evaluation and a dimension along which it is evaluated. Most of the literature refers to both criteria and elements as "criteria."

communities, external reviewers, program directors, and others involved in the merit review process understand the merit review criteria and apply them when writing proposals, reviewing proposals, and making award decisions. Principal investigators (PIs) and reviewers have expressed confusion and concerns about NSF's Broader Impacts review criterion in particular, despite NSF's efforts to provide additional guidance.

Like NSF, most funders of basic research use peer review to assess the potential impacts of research projects in ways roughly analogous to NSF's Intellectual Merit and Broader Impacts review criteria. However, some concerns have surfaced related to the merit review process and peer review more specifically. For example, concerns have been raised about assigning value to the intellectual merit of research (Xu et al. 2021; Kuhn 1962), predicting the potential impact of funded projects (Avin 2015), and the potential for reviewer bias (Chen et al. 2022). The assessment of broader impacts is further complicated because broader impacts are difficult to quantify and can shift over time (Holbrook and Frodeman 2011; Ramos-Vielba, Thomas, and Aagaard 2022). Reviewers might also have strong prior beliefs that influence their perceptions of broader impacts, especially about technologies with widespread and varied impacts such as mining with hydraulic fracturing, conducting stem cell research, or growing genetically modified crops (Gunn and Mintrom 2017).

This literature review is part of a congressionally mandated evaluation to "assess how [NSF's] Broader Impacts review criterion is applied across the Foundation and make recommendations for improving the effectiveness for meeting the goals established in section 526 of the America Creating Opportunities to Meaningfully Promote Excellence in Technology, Education, and Science Reauthorization Act of 2010 (42 U.S.C. 1862p-14)" (Creating Helpful Incentives to Produce Semiconductors [CHIPS] for America Fund Act 2022). This request comes at a time when funding agencies like NSF are becoming aware of a lack of evidence on the efficacy of their merit review processes: funders from 25 countries recently concluded that "agencies around the world use very different [elements], and very few of these [elements] are evidence-based" (NWO [2017], as cited by Hug and Aeschbach 2020).

The literature review examines evidence-based practices for assessing the intellectual merit, broader impacts, and related concepts of sponsored research, and provides context for the evaluation. The evaluation includes a document review; interviews and focus groups with NSF staff, PIs, and reviewers; and an analysis of survey and review data. The literature review highlights practices across funders with similar review criteria and explains the evidence for their efficacy. It also highlights effective strategies for mitigating concerns about reviewer bias and subjectivity.

The literature addresses the following research questions:

RQ1: Do funders specify or define the relative emphasis reviewers should place on intellectual merit and broader impacts? If so, how much emphasis is placed on each of these facets?

RQ2: What practices and elements do funders use to evaluate intellectual merit?

RQ3: What evidence exists to support the reliability, validity, or efficacy of the processes used to evaluate intellectual merit?

RQ4: What practices and elements do funders use when evaluating broader impacts?

RQ5: What evidence exists to support the reliability, validity, or efficacy of the processes used to evaluate broader impacts?

RQ6: Are there important gaps in the literature on the assessment of the merit of sponsored research that are not otherwise addressed by research questions 3 or 5?

2. Data and Methods

A. Overview of methods

The literature review sought to identify research that might guide NSF's use of peer review of Intellectual Merit and Broader Impacts. Our examination comprised identifying, screening, reviewing, and classifying the existing literature.

B. Identifying literature

To better understand the current practices of funders of scientific research and address research questions 1, 2, and 4, we scanned documents and web pages created by a set of funding agencies developed in consultation with NSF and the project's technical working group (TWG). The search set consisted of seven funding agencies, including three government STEM funding bodies and four U.S.-based foundations. We selected two funding bodies because they have large funding budgets and extensive English language documentation of their grant review process; the TWG recommended the third one. We selected the four U.S.-based foundations from a list of major philanthropic foundations maintained by NSF. We first excluded foundations that did not have substantial nonmedical STEM funding lines and then selected four that provided relatively more public information about their process for evaluating proposals.

We searched the website of each funding agency using the search string shown in Exhibit 1 and reviewed the first 20 returned results. We also examined each funder's home page and identified the funding lines that most closely matched our focus on nonmedical basic and use-inspired research. We navigated their websites and searched for additional documentation related to their decision processes.

To identify literature for the study that might address research questions 3, 5, and 6, we considered the following sources: literature that NSF and the TWG provided, articles identified through hand searches of journals likely to publish articles on grant evaluation, database searches of EBSCO Academic Premier and Google Scholar, and documents frequently cited by relevant documents about broader impacts obtained from the other sources. NSF provided 31 citations in an initial memo that discussed grant review. Most of these articles focused on assessing broader impacts of research. The TWG also recommended two articles.

To supplement the articles provided by NSF and the TWG and develop a broad set of keywords for the database search, we hand searched three journals for articles related to merit and impact evaluation in grant review. We selected these journals based on their declared scope, lack of representation in the list or within the articles shared by NSF, and journal impact factor (JIF). We used JIF as a proxy for the likelihood of finding articles that were highly influential on the field.

We excluded only a single journal (JIF = 0.2) based on this factor. The journals we searched were as follows:

- Research Evaluation: an information science journal focusing on the evaluation of activities concerned with scientific research, technological development, and innovation.
- Research Policy: a management journal focused on the interaction between innovation, technology, or research, and economic, social, political, and organizational processes.
- Scientometrics: an information science journal focused on the quantitative features and characteristics of science and scientific research.

We used the articles provided by NSF or identified through a hand search to develop a list of keywords designed to identify studies of the assessment of intellectual merit or broader impacts of grant proposals. We then searched EBSCO Academic Premier and Google Scholar for English language articles published after January 1, 2000, that contained these keywords.

Exhibit 1. Databases searched, access dates, search terms, and number of results

Source	Date accessed	Search string	Number of results
Academic Search Premier: Research funding string	10/10/23	(TI ((grants OR grant OR grant-making OR grantmaking OR funding OR funder OR funded) N3 (review* OR process* OR award OR opportunit* OR impact* OR research OR science OR scientific)) OR AB ((grants OR grant OR grant-making OR grantmaking OR funding OR funder OR funded) N3 (review* OR process* OR award OR opportunit* OR impact* OR research OR science OR scientific)))	38,671
Academic Search Premier: STEM concepts string	10/10/23	AND ((TI ((science OR scientific OR technolog* OR engineering OR mathematics OR research)) OR AB ((science OR scientific OR technolog* OR engineering OR mathematics OR research)))	33,488

Source	Date accessed	Search string	Number of results
Academic Search Premier: Intellectual merit and broader impacts string	10/10/23	AND (TI (((Broad* N2 (impact* OR benefit*)) OR (overall N2 (impact* OR benefit*)) OR (societal N2 (impact* OR benefit*)) OR (social N2 (impact* OR benefit*)) OR (public N2 (impact* OR benefit*)) OR ((science OR scientific) N2 (impact* OR benefit*)) OR (merit N2 review) OR (intellectual N2 merit) OR (grant* OR funding OR funder OR review*) N3 ((broad* N2 impact*) OR merit OR evaluat* OR norms OR consideration* OR review* OR criter* OR process* OR practice* OR alignment OR align OR measur* OR accountab* OR metrics OR data OR quantitative OR mapping OR strateg* OR framework* OR Bibliometric* OR Scientometric* OR "science of science" OR "Peer review" OR altmetrics) OR (merit N2 review) OR (intellectual N2 merit) OR ((science OR scientific) N2 merit)) OR AB (((Broad* N2 (impact* OR benefit*)) OR (overall N2 (impact* OR benefit*)) OR (societal N2 (impact* OR benefit*)) OR (social N2 (impact* OR benefit*)) OR (public N2 (impact* OR benefit*)) OR ((science OR scientific) N2 (impact* OR benefit*)) OR (merit N2 review) OR (intellectual N2 merit) OR (grant* OR funding OR funder OR review*) N3 ((broad* N2 impact*) OR merit OR evaluat* OR norms OR consideration* OR review* OR criter* OR process* OR practice* OR alignment OR align OR measur* OR accountab* OR metrics OR data OR quantitative OR mapping OR strateg* OR framework* OR Bibliometric* OR Scientometric* OR "science of science" OR "Peer review" OR altmetrics) OR (merit N2 review) OR (intellectual N2 merit) OR ((science OR scientific) N2 merit)) OR (TI predict* N2 impact OR AB predict* N2 impact) OR (TI impact N2 assessment OR AB impact N2 assessment) OR (TI selection N2 procedures OR AB selection N2 procedures) OR (TI research N2 quality OR AB research N2 quality) OR (TI "research impact" OR AB "research impact")	3,535
Academic Search Premier: Research funding string	10/10/23	OR (TI predict* N2 impact OR AB predict* N2 impact) OR (TI impact N2 assessment OR AB impact N2 assessment) OR (TI selection N2 procedures OR AB selection N2 procedures) OR (TI research N2 quality OR AB research N2 quality) OR (TI "research impact" OR AB "research impact")	3,535
Google Scholar	10/21/23	(intitle:grant|fund|funding|award)(science|technology|engineering|math| research|STEM|basic)(broader|overall|societal|social|public|merit|intellectual| scientific|"research quality") (predict|~evaluate|criteria|framework| review|select)	1,000

Source	Date accessed	Search string	Number of results
Google custom website search	10/21/23	Site: [*domain*] AND (grant OR fund OR award) AND (predict OR evaluate OR criteria OR Framework OR review OR select OR panel) Where [*domain*] = the website domain of the corresponding funding entity: Netherlands Organisation for Scientific Research (NWO): https://www.nwo.nl/en United Kingdom Research and Innovation (UKRI): https://www.ukri.org/ German Research Foundation (DFG): https://www.dfg.de/en/ Alfred P. Sloan Foundation: https://sloan.org/ Bill & Melinda Gates Foundation: https://www.gatesfoundation.org/ MacArthur Foundation: https://www.macfound.org/ Gordon and Betty Moore Foundation: https://www.moore.org/	140

Notes: The asterisk (*) is a Boolean operator and allows the truncation of the term so the search returns any word that begins with the specified letters. Our search returned records based on keywords appearing in titles, abstracts, subject headings, and author-supplied keywords.

Google Scholar allows the pipe character ("|") to function as a Boolean "OR" operator. Terms that are not separated by a Boolean operator are treated as "AND" operators. Including a tilde ("~)"" in front of a term will search for synonyms of the term.

These searches are limited to English language articles published after January 1, 2000.

The Academic Search Premier string in Exhibit 1 is separated into three components to make the search logic easier to understand. The first row constrains the search to articles with titles or abstracts (including keywords) that refer to research funding. The second row adds the requirement that titles or abstracts also refer to STEM-related concepts. The third row further restricts the search to articles that include keywords related to the assessment of intellectual merit or broader impacts.

The Google Scholar search is based on the Academic Search Premier string. The string was modified to fit within the 256-character Google Scholar search limit (Google drops characters beyond the 256th character before starting the search). Unlike the database searches that returned records based on keywords appearing in titles, abstracts, and author-supplied keywords, Google Scholar searches the full text of documents and returns many more records. It uses an algorithm that returns the 1,000 most relevant articles, ranked in order of their likely relevance.

C. Screening literature

In screening the literature, we eliminated duplicate search results and reviewed titles and abstracts to screen records for possible inclusion in the literature review. We then created a set of screening criteria and iteratively improved the criteria while reviewing a set of 200 articles.

The study team discussed points of disagreement and revised definitions through consensus. The final screening protocol excluded any document that did not meet the following six criteria:

- The full text of the document must be available in English.
- The document must be about evaluating the merit of research.
- The document must not be exclusively about NSF. Articles about NSF will be reviewed under a separate task. (We included for review those articles that referenced NSF along with other funding agencies.)
- The document must concern the review of nonmedical basic and use-inspired research and/or STEM education.
- The document must discuss ex ante funding decisions. Ex post data can be included if the data is used to evaluate the quality of ex ante funding decisions.
- The study must address at least one of the following: funders' descriptions of how grant review is supposed to happen, descriptions of current practices, evidence of best practices, or gaps in evidence for best practices in the review of intellectual merit or broader impacts.

We used ASReview to screen articles retrieved from EBSCO Academic Review and Google Scholar. ASReview is a computer application that uses artificial intelligence (AI) to streamline the literature screening process. It uses active learning to predict which documents are likely to be relevant based on an initial training set of relevant and irrelevant articles and sorts the documents in order of likely relevance. A person reviews each document in sequency and decides whether it is relevant. ASReview uses each screening decision to dynamically update the predicted relevance of each document and re-rank them.

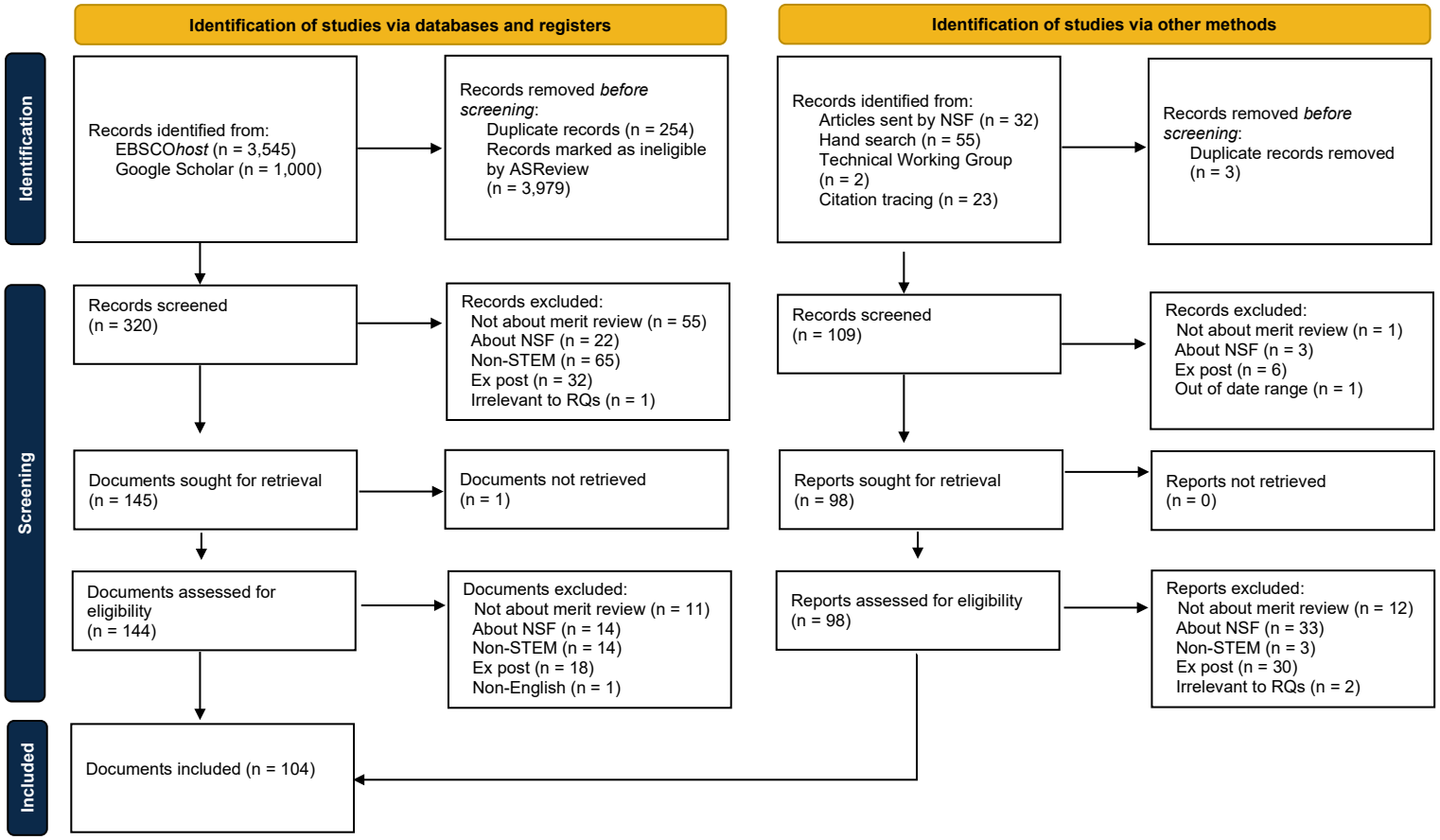
We used ASReview's recommended setup for an active learning model. This setup uses a term frequency-inverse document frequency (TF-IDF) feature extraction technique to assign weights to individual terms and a Naïve Bayes classifier to predict document relevance based on the presence of each term. For each article, TF-IDF assigns values to words that represent how frequently they appear in the article relative to how frequently they appear across all articles. A Naïve Bayes classifier estimates the likelihood of each term given its classification (as relevant or irrelevant) and then uses this information to predict the category that the article belongs to, given the terms that appear in it.

To screen documents retrieved from Academic Search Premiere, we trained the initial model with eight articles; four articles were relevant and four were irrelevant. The article titles and abstract text served as inputs to the ASReview algorithm. A trained analyst continued screening documents until we encountered 10 irrelevant articles in a row. We tested the sensitivity of the model by examining a random sample of 200 automatically screened out articles. None of them

were relevant. To screen documents from Google Scholar, we trained the model on the titles and abstracts of all screened articles from Academic Search Premiere. The article titles and text snippets returned by Google Scholar served as inputs to the ASReview algorithm.

As described later, we discovered very few documents that specifically addressed the evaluation of broader impacts. We used citation tracing to identify additional articles deemed relevant to experts in assessing broader impacts (the authors who published relevant articles) regardless of whether they include relevant keywords. Because there is often little consensus about terminology among researchers, citation tracing reliably detects papers missed by keyword search strategies (Hirt et al. 2022; Horsley et al. 2011). We found an additional 23 documents that cited or were cited by at least three documents identified through other means, screened into the study and classified as relevant to broader impacts (as described below). We screened each of these articles using the screening protocol described above.

Exhibit 2. PRISMA diagram with results of the screening process



Note: RQ = Research Question. Diagram adapted from: Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021;372:n71. doi: 10.1136/bmj.n71. For more information, visit: http://www.prisma-statement.org/

D. Reviewing studies

The study team reviewed each study by using a standardized template. This template was designed to capture information that directly addressed the research questions and collected additional contextual information, such as features of the study (like study design, data, outcomes, sample, and methods), limitations of the study, and considerations for interpreting study results. A senior team member checked the reviewed articles to ensure that relevant details were not overlooked. Once this review was completed, we identified 104 articles for inclusion in the literature review (Exhibit 2).

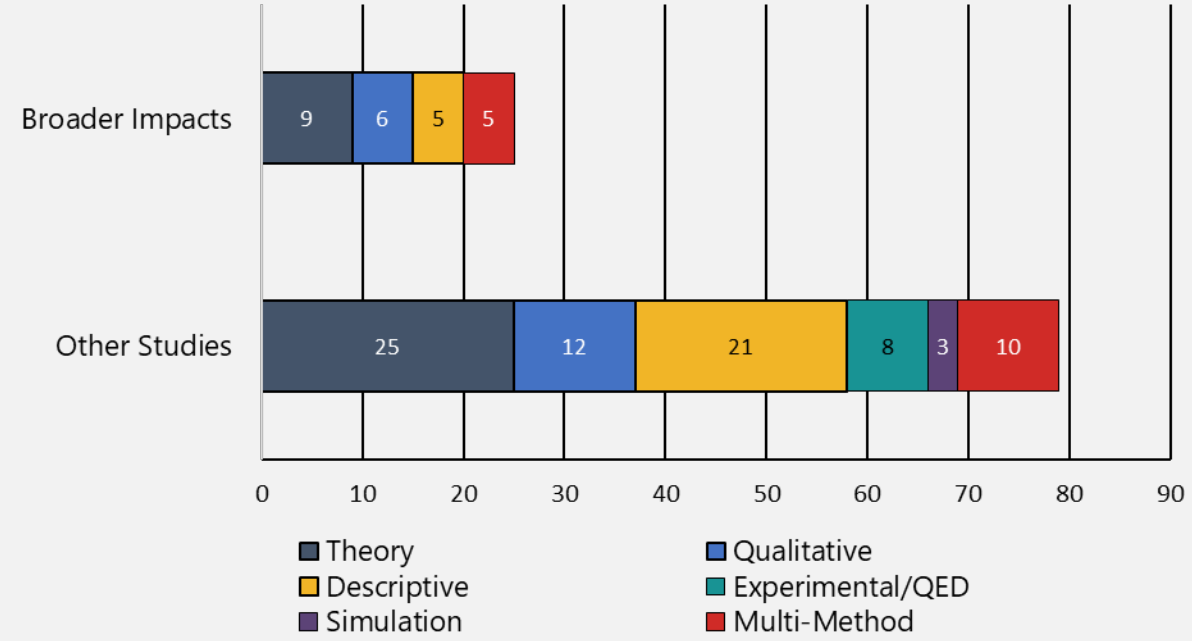
E. Classifying literature

Once the study team screened in a publication for review, they classified documents according to whether they focused on broader impacts or not. Articles that did not focus on broader impacts discussed grant review in general terms or more rarely focused on intellectual merit. We also classified documents into one of six categories according to the type of evidence they contained:

- **Theory/literature review:** Studies that review existing literature and describe concepts and theories but do not report original data.
- **Qualitative:** Studies that use quotations from interviews, document review, or observations as evidence.
- **Simulations:** Studies that model potential outcomes based on mathematical representations of inputs and decision rules.
- **Descriptive/correlational:** Studies that present quantitative or qualitative descriptive statistics, such as frequencies, patterns, or trends, which may include associations between variables.
- **Experimental/QED:** Studies that report primary data and claim causal relationships between variables.
- **Multi-method:** Studies that report using more than one of the above methods.

See Exhibit 3 for the number of studies in each category.

Exhibit 3. Classification of literature



Note: Theory includes all studies that do not report data. Qualitative studies use quotations from interviews, document review, or observations as evidence without counts. Descriptive studies present descriptive statistics, such as counts, patterns, or trends. Experimental/QED studies report primary data and claim causal relationships between variables. Simulation studies model potential outcomes based on mathematical representations of inputs and decision rules. Multi-method studies report using more than one of the above methods.

3. Limitations

Our approach to the literature review has two main limitations. First, the literature search might have missed relevant articles. The search posed calibration challenges because many of the terms (such as "funding" or "science") appear frequently in journal articles.

Another challenge is that we limited the scope of research to focus on ex ante nonmedical research. There is a large literature base on grant review in medicine that has explored similar topics and has reached conclusions about potential best practices that might generalize to other disciplines.

Similarly, there is also a large literature base on the ex post assessment of intellectual merit and broader impacts. We excluded this literature because ex post assessments are not faced with the same uncertainty as ex ante assessments. These studies differ too much from ex ante assessments to provide information about potential best practices. However, they could provide insight into the difficulties in assessing broader impacts under the best possible circumstances, where the impacts have already occurred and presumably are known.

4. Results

A. What processes do funders use when evaluating grants?

Funders use broadly similar processes to evaluate grant proposals, with some differences in terminology and implementation of processes. Funders post calls for grant proposals that describe proposal requirements and the funding terms. Applicants then submit grant proposals to funders for consideration. For calls with a deadline, funders usually make funding decisions for all applicants at the same time. For calls without a deadline, funders evaluate applications on a rolling basis, either alone or in batches.

Grant proposal evaluation is a multistage process that can be thought of as a kind of "aggregation machine" that collects and synthesizes information to make an informed funding decision (Rip 2000). Applications are usually evaluated first for completeness and compliance with eligibility and proposal requirements, either by program staff or through automated checks. Most funders then gather input from both internal and external reviewers, with differences in which group is asked to provide general scientific expertise and which group provides specialist expertise about the proposal topic (Biegelbauer, Palfinger, and Mayer 2020; Heinze 2008; Rip, 2000).

Reviewers evaluate grants against proposal requirements and usually provide a score or summative evaluation of proposals and written comments that highlight their strengths and weaknesses. Some funders ask review panels to reach consensus about a proposal score, whereas others only ask for individual reviewer scores.

Funders use different approaches to make funding decisions. Funding recommendations often are made by a program officer[2] based on the scores or comments provided by reviewers. In some cases, review panels may themselves make funding recommendations. In other cases, the scores or rankings produced by reviewers are used to determine funding: applications are either ranked by score and funding is disbursed in order of rank until all funding is allocated, or funds are disbursed to applicants with scores above a defined threshold (called a "payline").

B. What do funders ask reviewers to consider when evaluating grants?

Reviewers assess the substantive content of proposals against defined criteria and elements. Criteria are the broad standards against which the proposal is evaluated. Criteria are composed of elements that define those parts of the proposal relevant for assessing criteria. Elements specify an entity to be evaluated, like the research idea or the applicant's prior publications. They may also specify a dimension by which to evaluate the entity. For example, an applicant's

[2] We use program officer as a general term that refers to individuals with a role analogous to that of a program director at NSF.

publications could be evaluated according to their impact, rigor, or relevance to the proposed work. Different stages of the review process may emphasize different criteria or elements.

Most funding elements can be classified as supporting one of NSF's two criteria: Intellectual Merit and Broader Impacts).[3] All science funders care about the intellectual impact of proposals—the degree to which proposals may advance a field of study through theory building, methodological advances, and their associated academic outputs (sometimes referred to as "Mode 1"; Bornmann 2013). Most funders are also concerned about the impacts of proposals outside of the applicants' academic community, which include social, ethical, and environmental consequences of research ("Mode 2"; Bornmann 2013). These distinctions follow the Stokes (1997) conceptualization of scientific activities along two orthogonal dimensions: "quest for fundamental understanding" and "considerations of use" (Veletanlić and Creso 2020). Sometimes funders will define specific strategic objectives that a proposal must support, whereas in other cases researchers are free to define the potential consequences for society.

Experts in funding policy also sometimes distinguish between elements related to outcomes and elements related to the plans and resources to achieve these outcomes (Ramzgir et al. 2021). Research can be described both as what might be learned through exploring a topic, and an approach by which the applicant will explore the topic. Likewise, the broader impacts of a grant could be described both in the user-oriented outputs it expects to produce (like new technologies, processes, or policies) and the methods by which it will ensure that the research has value to society (the inclusion of non-academics in developing ideas, defining the priorities or approaches of a research project and disseminating research findings; Arnott et al., 2020; Bennewoth and Olmos-Peñuela 2022; Ramos-Vielba et al. 2022).

Theorists have noted that it is harder to evaluate research outcomes than research plans for at least three reasons. First, the soundness of the research plan is a smaller part of the evaluation of research outcomes because such outcomes are conditional on both whether something interesting exists to be found and the soundness of the research plan (Franzoni and Stephan 2022). Second, projects can differ in their potential to make or capitalize on unanticipated discoveries that are secondary to their main objectives. Third, even when scientific results are known, the value of a discovery is inherently more subjective than the rigor of the method (Tennant and Ross-Hellauer 2020).

[3] In this report, we use the term "criteria" to refer to intellectual merit and broader impacts. Other funders use this term in different ways, either listing more granular criteria that can be subsumed into the NSF criteria or defining criteria the way we define elements and using dimensions that assess intellectual merit or broader impacts to evaluate entities.

Comparisons of published funding criteria and elements

We reviewed four documents that described, classified, and compared grant funding criteria used by different funders. They are a comparison of review criteria used by U.S. federal agencies, a comparison of the funding criteria used by a purposive sample of funders focused on promoting innovation, a comparison of five frameworks used to assess broader impacts, and a comparison of NSF's Broader Impacts criterion to the Responsible Research and Innovation (RRI) approach used by the European Union's (EU's) Framework programs.

U.S. federal agencies generally ask reviewers to use the same funding criteria and elements to evaluate grant proposals (Falk-Krzesinski and Tobin 2015). The study by Falk-Krzesinski and Tobin compared descriptions of the research grant review process available on the websites of 10 U.S. federal agencies (National Institutes of Health [NIH], NSF, Department of Veterans Affairs [VA], Department of Education [ED], Department of Defense [DOD], National Aeronautics and Space Administration [NASA], Department of Energy [DOE], U.S. Department of Agriculture [USDA], National Endowment for the Humanities [NEH], and National Endowment for the Arts [NEA]). All agencies evaluated whether the proposed research mattered (analogous to intellectual merit), in what way it was new, how the work would be done, and whether the proposal team had the appropriate capabilities. Nine agencies asked about the potential impacts of research on society: DOE did not, but in this case, impacts might be closely tied to why the research matters. Many agencies shared several other elements: nine agencies asked about facilities or resources, and seven asked about the value of the work (cost or budget). Occasionally, agencies required grantees to explain how they would know if their project was successful (n = 4) or explicitly evaluated the clarity of proposal writing (n = 1). The authors mentioned that agencies varied in the weighting assigned to each criterion or element but did not systematically explore these differences.

A study of nine programs, intended to encourage groundbreaking research, identified two strategies that funders used that had distinct goals and criteria elements (Heinze 2008). Five of nine programs emphasized the applicant's track record and leadership qualifications, based on the assumption that providing stable, unrestricted funding would allow highly qualified applicants to take risks. These programs included the Hughes Investigator Program, Krupp Förderspries, McDonnell 21st Century Science Initiative, European Science Foundation Young Investigator Award, and European Research Council (ERC) Independent Researcher Grant. Four programs emphasized the originality and intellectual merit of the proposed research and sought to fund highly speculative, risky, or (to a lesser extent) interdisciplinary projects. In this approach, the applicant's academic track record and leadership qualifications were secondary considerations. These programs included the Wellcome Commemorative Award, the Volkswagen Off the Beaten Track Scheme, Israel Science Foundation Focal Initiatives in Research and Technology, and UK Engineering and Physical Sciences Ideas Factory.

Existing frameworks that focus on how to evaluate proposals for their potential to achieve broader impacts vary widely in their emphasis on methods and outcomes. Pederson, Grønvad, and Hvidtfeldt (2020) examined five frameworks that could be used to evaluate ex ante impacts: the Health Economics Research Group (HERG) Payback Framework; Social Impact Assessment Methods for research and funding instruments through the study of Productive Interactions (SIAMPI); Contribution Mapping; the Research Contribution Framework; and the Research and Policy in Development (RAPID) Outcome Mapping Approach. HERG examines impacts on teaching, policy, and practice, and is the simplest model in that it assumes that broader impacts of research are linear (that is, inputs feed into research processes that output broader impacts).

Other models make the more complex assumptions that steps in the research process can feed back to earlier steps and that broader impacts can occur at any stage of the research process. SIAMPI focuses on productive interactions between researchers and non-academics. Contribution Mapping focuses on the activities of different actors toward research methods rather than final impacts. The Research Contribution Framework focuses on research uptake by non-academics and is intended to assess impacts within defined organizations rather than society at large. The RAPID Outcome Mapping Approach focuses exclusively on outcomes, but these include changes in the attitudes and behaviors of non-academics that are assumed to be important precursors to change.

Davis and Kelly (2013) also highlighted differences in funders' emphasis on methods when contrasting NSF's outcomes-based approach with the RRI approach used by the EU's Framework programs. They found that RRI adopts a much more expansive definition of broader impacts than NSF, referring to dimensions like sustainability, ethical acceptability, and societal desirability that are abstract and applicable to nearly any project. RRI also explicitly defined practices that might support broader impacts, such as increased research transparency, interactive processes, and a movement to expand who may participate in research.

Descriptions of criteria used by grant proposal reviewers

Hug and Aeschbach (2020) developed a comprehensive taxonomy of the criteria and elements that reviewers use to evaluate proposals. To do so, they screened more than 16,000 articles and identified 12 studies on grant review criteria and elements in one or more disciplines. Most studies (n = 8) included biomedicine, but several also included domains relevant to NSF: two included natural science, two included engineering, and four included social science.

The authors content coded grant elements according to combinations of an "entity" (an element of the proposal) to be evaluated and the dimension used to evaluate it. Across the 12 studies they analyzed, the authors identified 373 entity-dimension combinations. Within these combinations, they identified 27 distinct entities that belonged to three superordinate

categories: those related to what applicants say they will do with the funds (activity elements); the applicant's abilities, usually demonstrated through their prior achievements (applicant elements); and resources present in the environment in which the applicant will complete the work (access to specialized equipment, partnerships, etc. — resources elements). They also identified 15 dimensions (such as quality, completeness, or originality) to apply when evaluating entities.

The frequency with which different elements are used indicates how reviewers think about proposals. First, reviewers place most of their emphasis on the project: 72 percent of mentioned entities were related to the project, 21 percent of elements were related to the applicant, and the remainder concerned the research environment (Hug and Aeschbach 2020). Interestingly, the authors also noted that 13 percent of the identified elements specified the entities to be evaluated but not which dimensions should be considered in the evaluation.

The dimensions used to evaluate each element suggest that reviewers are mostly concerned with research methods. Hug and Aeschbach (2020) identified five main clusters or elements and dimensions through a weighted network analysis.[4] In order of the frequency of their occurrence, these are rigor of research, clarity, and completeness of proposals; potential project results (which combines academic and broader impacts); project feasibility; global quality assessment; and applicants' psychosocial characteristics (such as motivation or traits). Finally, the extra academic relevance dimension (similar to the broader impacts criterion) was almost always applied to the project in general or the research question or results, but rarely to the methods used to conduct the work.

The available data also suggest that intellectual merit is of primary importance to reviewers. Using a small sample of proposals submitted to the Russian Foundation for Basic Research, Devyatkin et al. (2016) demonstrated that reviewers' assessments of intellectual impact and project feasibility were the strongest predictors of the overall score they assigned to the proposal. Consistent with this finding, a nonprobability survey of agency officials, reviewers and applicants from NSF, NIH, the Natural Sciences and Engineering Research Council of Canada (NSERC), and the European Commission's Framework Program were more likely to agree that the benefits of scientific research arise mostly through its contribution to the scientific body of knowledge, than with statements about the economic and social impacts of research (Holbrook and Hrotic, 2013).

[4] Network analysis evaluates the structure of networks of entities that are related to each other—in this case, proposal elements and evaluation criteria. The strength of the relationship between entities is defined by the number of times they occur together. Network analysis can be used to group entities together into communities that maximize the strength of within-community connections and minimize the strength of between-community connections.

A combined sample of proposal reviewers in physics, economics, and cardiology were most likely to rate the research question (94 percent) and methods (84 percent) as most important, followed by aspects of the research team and facilities (41–45 percent), the team's prior research productivity (23 percent), and the applicant's experience with risk-taking research (19 percent; the order of the ranking did not depend on discipline). Only 6 percent of academics rated the team's communication plan for addressing non-academic users—the only item that approached broader impacts—as highly important (Langfeldt, Reymert, and Aksnes2021). Similarly, Holbrook and Hrotic (2013) reported that program directors and proposal reviewers rated the intrinsic value of research (intellectual merit) as more important (by a full scale point on a 7 point scale) than the instrumental value of research (broader impacts) for making funding decisions.

One small study suggested that reviewers prefer to discuss concrete details when evaluating broader impacts. Ma et al. (2020) content coded the impacts section of individual reviews for 261 proposals submitted to Science Foundation Ireland. Reviewers commented more about economic impacts and interactions with others (students, partners, scientists) than about social impacts. The authors observed that comments about process-oriented impacts (that is, methods or practices that might lead to outcomes) were more specific than those about outcome-oriented impacts. Reviewers also tended to express more concerns about long-term than short-term impacts. This study did not analyze proposal text, so the authors could not determine the extent to which these observations reflected reviewer behavior or proposal content.

Anecdotally, researchers will sometimes deviate from the instructions that funders provide about which elements to use to evaluate proposals. In some cases, they might consider irrelevant dimensions, such as applicant character, whereas in others they might overlook funding agency-identified criteria or dimensions, such as strategic importance or social impact (Bulathsinhala 2015; Hug and Aeschbach 2020; Langfeldt 2001). Ma et al. (2020) observed that reviewers often discussed research outputs in their comments about impact, even though the proposal call and reviewer guidelines stated that this section was reserved for economic and social impact. These tendencies can be difficult to overcome even when funders require reviewers to be transparent about how they arrive at their final proposal scores (Reale and Zinilli 2017).

Comparisons of current practices and funding criteria of scientific research funders

To better understand the elements that funders use to evaluate intellectual merit and broader impacts and the weight they place on them, we reviewed the processes and published funding criteria used by seven funding agencies to assess the merit of proposals, including three government funding bodies and four U.S.-based foundations. The funders we examined included the German Research Foundation, NWO, United Kingdom Research and Innovation, Alfred P. Sloan Foundation, Bill & Melinda Gates Foundation, the MacArthur Foundation, and the Gordon and Betty Moore Foundation.

We reviewed the elements and element definitions used by funders and classified them according to whether they assessed the intellectual merit or broader impacts of the proposed work. We further classified these elements according to whether they focused on outcomes, the plan to achieve these outcomes, or applicant qualities and resources that will ensure the plan's success. This taxonomy is based on the distinction made by Hug and Aeschbach (2020) between project-, applicant-, and resource-related proposal elements. We collapsed the applicant and resource elements into a single category because resources are mentioned infrequently and (when referring to the research team's human capital) overlap with applicant characteristics. We further distinguished project elements related to outcomes from those related to processes, following the distinction between plans and outcomes made by Ramzgir et al. (2021). The elements of proposals that government funding agencies assess are found in Exhibit 4, and the elements of proposals that foundations assess are found in Exhibit 5. The sources we consulted for this review are listed in Appendix A.2.

The German Research Foundation (DFG) is a federal agency that funds research projects in the sciences and humanities. The DFG receives most of its money from the German federal government, though it also receives funding from state governments. We examined the elements assessed by two funding lines:

1. **Investigator Funding Focus** funds researchers at different career levels to pursue research that is not thematically constrained. This funding line includes the Walter Benjamin Programme, Emmy Noether Programme, Heisenberg Programme, and Research Fellowships.
2. **Themes Funding Focus** includes two relevant funding lines. Individual Research Grants fund research on specifically defined topics, and the Reinhart Koselleck Projects fund exceptional researchers to pursue high-risk or innovative research.

The DFG identifies reviewers to assess proposals based on their subject-specific qualifications. Next, the proposal and reviews are shared with a review board of elected scientists. The agency has ratified the Declaration on Research Assessment (DORA),[5] but its reviewer instructions do not describe how to review applicants' academic output. The review board performs a comparative review of all proposals within a given subject area. It then develops a recommendation to fund or not fund a proposal, which it then shares with the joint committee of the DFG, which makes the final decision.

The Netherlands Organisation for Scientific Research (NWO) is the Dutch national funding body dedicated to advancing scientific research. The agency operates under the Ministry of Education, Culture and Science. We examined the elements assessed by two funding lines:

[5] DORA recommends that JIFs not be used to compare the scientific output of individuals. It further recommends that funders prioritize the scientific content of papers over publication metrics or their publication outlet, and consider the value of all research outputs using a broad range of measures, rather than focusing only on the publication impact.

1. **Open Competition** funds researchers to perform research in a domain of their choosing without being limited by thematic restraints. Funded domains include science, social science and humanities, and applied and engineering sciences.
2. **Talent Programme** funds individual researchers at varying levels in their research careers. It includes the Veni, Vidi, Vici program and the Rubicon program.

NWO funding lines use different evaluation procedures, but usually applicants follow a multistage process that minimizes the effort applicants and reviewers spend on noncompetitive proposals. Applicants usually start by submitting a pre-proposal that is first reviewed for eligibility and then reviewed against the proposal criteria and elements. Pre-proposals are ranked on their performance, and promising applicants are invited to submit a full proposal. Full proposals are evaluated by external reviewers based on the corresponding program's assessment criteria and elements. Reviewers provide their initial scores and comments to a committee, which scores and ranks proposals. An interview might follow, during which the applicant can describe their proposal and answer questions. The committee provides proposal scores for interviewees, which the agency then standardizes; the highest scoring applicants qualify for the grant.

The formal criteria and elements used at each stage and across funding streams can differ, but the weight assigned to each major criterion is clearly specified. Applicants can specify whether their proposal should be assessed for scientific impact, social impact, or both. When assessing social impact, reviewers have broad latitude about how to use criteria and elements to evaluate applicants, but must follow DORA guidance to avoid considering JIFs when evaluating applicants' academic output.

United Kingdom Research and Innovation (UKRI) is a national funding body that funds research within five strategic themes: building a green future; building a secure and resilient world; creating opportunities and improving outcomes; securing better health, aging, and well-being; and tackling infections. It is funded by the Department for Science, Innovation, and Technology. We examined the elements assessed by four funding lines:

1. **Biotechnology and Biological Sciences Research Council (BBSRC)** supports work aimed at furthering the field of biology to improve future societal and scientific outcomes.
2. **Economic and Social Research Council (ESRC)** funds work in the fields of economic, social, behavioral, and human data science.
3. **Engineering and Physical Sciences Research Council (EPSRC)** funds engineering and physical sciences as opportunities to support the UK's societal and economic success.
4. **Natural Environment Research Council (NERC)** funds environmental science.

UKRI's review process usually begins with a peer review conducted by independent reviewers. Proposals are scored on a 6-point scale. When reviewing applicants, reviewers must adhere to DORA guidance to avoid considering JIFs when evaluating applicants. When possible, UKRI tries to provide applicants with an opportunity to respond to reviewer comments. Next, the proposals and peer-review comments might be provided to a review panel that compares all the proposals to identify those that will be funded.

EPSRC occasionally uses other processes to provide applicants with rapid feedback, including reviewing outlines and expressions of interest before a full proposal, and inviting potential grantees to participate in "sandpits." A sandpit is an interactive discussion forum where researchers can engage with experts and receive review feedback in real time. Though currently paused, UKRI until recently had a policy that restricted historically unsuccessful funding applicants to a single submission every 12 months, which might help limit submission volume (Roebber and Schultz 2011).

Exhibit 4. Comparison of elements used to assess funding criteria across selected government research funders

German Research Foundation

Program	Element	IM outcome	IM method	IM applicant/ resources	BI outcome	BI method	BI applicant/ resources
Investigator/ Themes Funding	Quality of the project		X				
	Objectives and work program	X					
	Applicant's qualifications			X			

Netherlands Organisation for Scientific Research

Program	Element	IM outcome	IM method	IM applicant/ resources	BI outcome	BI method	BI applicant/ resources
Open Competition	Scientific quality		X				
	Scientific/social impact	X			X		
Talent Programme	Pre-approval	X		X	X		
	Scientific quality	X	X	X			
	Scientific/social impact[a]	X			X		

United Kingdom Research and Innovation

Program	Element	IM outcome	IM method	IM applicant/ resources	BI outcome	BI method	BI applicant/ resources
Biotechnology and Biological Sciences Research Council (BBSRC)	Aims and objectives	X					
	Strengths and weaknesses	X					
	Feasibility		X	X			
	Timeliness and promise	X	X				
	Strategic relevance: Industry				X		
	Strategic relevance: BBSRC				X		
	Economic and social impact				X		
	Value for money	X					
	Training potential				X		
Economic and Social Research Council (ESRC)	Originality	X					
	Design and methods		X				
	Data management plan		X				
	Research ethics		X				
	Outputs, dissemination, and impact	X			X		
	Value for money	X					
Engineering and Physical Sciences Research Council (EPSRC)	Novelty, context, timeliness, relevance to stakeholders	X			X		
	Ambition, adventure, transformative aspects, or potential outcomes	X			X		
	Suitability of the methodology approach to achieving impact		X			X	
	Importance	X			X		
	Applicant and partnerships			X			?
	Resources and management		X			X	

Program	Element	IM outcome	IM method	IM applicant/ resources	BI outcome	BI method	BI applicant/ resources
Natural Environment Research Council (NERC)	Research excellence	X					
	Fit to scheme		X				
	Resources		X				
	Risks[b]	X					
	Capability to deliver[b]			X			

Notes: IM = intellectual merit. BI = broader impacts. Element names are those used by the funders. Elements are classified according to criteria and the object of evaluation based on the definitions provided by funders: "X" indicates that a review element can be classified as measuring a combination of a criterion and an entity; "?" indicates that it was unclear whether the element maps to a combination of a criterion and an entity.

[a] Applicant may focus on either academic impact, social impact, or both.

[b] Does not apply to all programs within the funding stream.

The **Alfred P. Sloan Foundation** funds research by nontenured faculty in STEM fields and economics. We examined the elements used by three funding lines (information about selection procedures were unavailable for two other funding lines: Energy and Environment and Small-Scale Fundamental Physics):

1. **Economics** funds basic research that promotes equity, protects consumers, strengthens institutions, incentivizes innovation, tests technologies, or improves the value of scientific research.
2. **Energy and Environment** funds research that informs transitioning to low-carbon energy systems in the United States.
3. **Matter-to-Life** funds research that advances an understanding of building blocks of life.

The Foundation instructs interested applicants to identify a program that aligns with their goals and interests and submit a letter of inquiry to a program director. If the program director decides the letter of inquiry is promising, they will invite the applicant to submit a full proposal. The Foundation then evaluates the full proposal through a review process described as like the peer-review process used for high-quality academic journals. Depending on the proposal content and requested funding, the Foundation might also request feedback from independent experts to supplement the existing peer review process.

The **Bill & Melinda Gates Foundation** funds work through programs in one of six divisions: Gender Equality, Global Development, Global Growth and Opportunity, Global Health, Global Policy and Advocacy, and the U.S. Program. We examined the elements assessed by two funding lines from divisions which were most aligned with NSF's funding streams:

1. **Strengthening African National Regulatory Authorities Data Systems to Enhance and Track Performance** funds innovation in tackling global health and development challenges. This grant aims to address problems with access to quality medical products in Africa, but the central challenges relate to measurement, database development, and systems interoperability.
2. **Accelerating Catalyzing Solutions for Climate Change's Impact on Health, Agriculture, and Gender** funds efforts to mitigate the effects of climate change, with a focus on health, nutrition, agriculture, and knowledge management.

The Gates Foundation has program officers who identify applicants in one of three ways: through direct solicitation, by receiving a letter of inquiry, or through a response to a request for proposals. The program officer works closely with the applicant to develop the full proposal and corresponding funding recommendation. When the proposal is ready, the program officer reviews it and makes a funding recommendation for review by a foundation executive, who will then make a funding decision.

The **MacArthur Foundation** provides grants aimed at addressing significant social challenges. We examined the elements assessed by two funding lines:

1. **Big Bets** funds potentially transformative research in different topic areas. We examined funding for climate solutions and nuclear challenges.
2. **MacArthur Fellows** funds individuals identified as being exceptionally creative and does not place any restrictions on awarded fellows' grant use.

The assessment process for the MacArthur Foundation varies by program. For the Big Bets funding opportunities, a program officer invites an organization to draft a proposal and works with the applicant to identify a deadline. Once the applicant prepares their proposal, the program officer will analyze it to share with Foundation leadership. If questions arise, the program officer may contact the applicant to provide answers. MacArthur Fellows are first nominated by a pool of external nominators. Program staff prepare a file of letters of evaluation and samples of the nominee's work. These are reviewed by an independent selection committee.

The **Gordon and Betty Moore Foundation** provides grants through programs aimed at producing lasting benefits for science, conservation, and medicine. We examined the elements assessed by two funding lines:

1. **Science** provides grants to advance scientific progress in terms of technological development and supporting work in new and existing disciplines.
2. **Environmental Conservation** provides grants to projects aimed at building healthy ecosystems.

The Gordon and Betty Moore Foundation has an interactive grant proposal process, which starts with program officers developing an internal grant team. The grant team works with the applicant to identify shared goals and a general plan for the grant. The applicant then develops the specific plan for the grant, including the activities and resources needed. Once the applicant has submitted the proposal, an "appropriate authority" reviews it and decides whether to fund the project (Grant Development Overview, n.d.).

Exhibit 5. Comparison of elements used to assess funding criteria across selected philanthropic research funders

Alfred P. Sloan Foundation

Program	Element	IM outcome	IM method	IM applicant/ resources	BI outcome	BI method	BI applicant/ resources
Economics	Policy relevant				X		
	Motivated by questions	X					
	Engaged with fundamental puzzles	X					
	Unbiased and replicable		X				
	Careful about methods		X				
	Contributes to resources like research infrastructure				X		
	Concerned with U.S. quality of life				X		
	Applicant curriculum vitae (CV)			X			
Other programs	Scientific importance/goals	X					
	Methodology		X				
	Outputs	X			X		
	Applicant CV			X			

Bill & Melinda Gates Foundation

Program	Element	IM outcome	IM method	IM applicant/ resources	BI outcome	BI method	BI applicant/ resources
All programs	Potential to lead to solutions with substantial impact				X		
	Technical excellence and innovation	?	?		?	?	
	Project plan			?			?

MacArthur Foundation

Program	Element	IM outcome	IM method	IM applicant/ resources	BI outcome	BI method	BI applicant/ resources
Big Bets	Problem or opportunity: ambition	?			?		
	Goal: boldness, creativity, and strategy	?			?		
	Challenge: feasibility, confidence, progress, learnings, and durability		?	?		?	?
	Moment in time: timeliness		?			?	
MacArthur Fellows	Exceptional creativity, demonstrated through track record of significant achievement			?			?
	Promise for important future advances	?			?		
	On the precipice of great discovery or a game-changing idea	?			?		

Gordon and Betty Moore Foundation

Program	Element	IM outcome	IM method	IM applicant/ resources	BI outcome	BI method	BI applicant/ resources
All programs	Important	?			?		
	Enduring difference	?			?		
	Measurable		?			?	
	Contribute to portfolio effect				X		

Notes: IM = intellectual merit. BI = broader impacts. Element names are those used by the funders. Elements are classified according to criteria and the object of evaluation based on the definitions provided by funders: "X" indicates that a review element can be classified as measuring a combination of a criterion and an entity; "?" indicates that it was unclear whether the element maps to a combination of a criterion and an entity.

C. How reliable are reviewer assessments of grant proposals?

The merit review process relies heavily on peer review, in which experts in a relevant research area provide input to decision makers about the strength of a proposal, usually through written comments accompanied by a qualitative rating or numeric score. To ensure a fair and accurate process, the measures used to make funding decisions must be reliable. That is, when the proposal scores a reviewer provides are used to determine funding (as is the case with many funding agencies), they should be consistent across time (test-retest reliability), and scores by

different reviewers should be consistent with one another (interrater reliability).[6] Analogously, if the written comments reviewers provide are important, there should be some overlap between reviewers about the critical strengths and weaknesses of a proposal. If the grant review process is less reliable, chance will play a greater part in the scores assigned to proposals, which in turn will affect the reliability of funding decisions.

We did not identify studies examining the test-retest reliability of proposal scores or the degree of overlap between reviewer comments. However, empirical studies of reviewer interrater reliability found consistently low levels of agreement. We identified six studies that examined the interrater reliability of proposal scores:

- Baimpos, Dittel, and Borissov (2020) examined reviews of proposals submitted to the Future and Emerging Technologies Program (FET-Open) at the Research Executive Agency (REA) of the European Commission. The FET program funds collaborative research on "radically new, high-risk ideas" in science and technology (External Funds Service 2023). FET-Open focuses on funding early-stage research related to new technologies. External reviewers assessed FET-Open proposals on three dimensions: excellence, impact, and implementation.
- Jerrim and Vries (2020) examined reviews of proposals submitted to the ESRC between 2013 and 2018 across different types of grant programs—individual fellowships, large research center grants, and open-call grants.
- Langfeldt (2001) examined reviews of proposals submitted to the Research Council of Norway, including proposals reviewed by its Science and Technology division. The evaluation elements included (1) applicants' prior merits; (2) project descriptions, including methods; (3) expected value of the project; (4) distributional policy, such as diversity in field, institution, geography, and demographics; (5) research policy objectives; and (6) other considerations, such as budget.
- Marsh, Jayasinghe, and Bond (2008) examined reviews of proposals submitted to the Australian Research Council (ARC). The ARC is Australia's main funder of basic research in science, social science, and the humanities.
- Mutz, Bornmann, and Daniel (2012) examined reviews of proposals submitted to the Austrian Science Fund (FWF) over a 10-year period, from 1999 to 2009. Their data reflected the universe of individual research proposals submitted to FWF across all disciplines and represented approximately 60 percent of all of FWF's grants.

[6] Interrater reliability is important for one of two reasons. First, interrater reliability establishes that individual ratings accurately measure a single unidimensional latent construct, such as the value of a proposal. If reviewers attend to and value proposal elements differently, low interrater reliability is expected and does not provide information about the accuracy of the individual reviews. Second, regardless of why interrater reliability is low, it provides information about the consistency of the outcome of expert evaluations. High interrater reliability implies that only a few experts need to be consulted, because their opinions are likely to be shared by others.

- Pina and colleagues (2021) examined reviews of proposals submitted to the EU Marie Skłodowska-Curie Actions (MSCA; named the Marie Curie actions before 2014) between 2007 and 2018. Their data included proposals submitted to the Individual Fellowships, Innovative Training Networks, and Research and Innovation Staff Exchange programs.

Each study measured reliability in one of two ways. The first is to examine how close proposal scores are to each other by reporting either the proportion of proposals with unanimous scores (Baimpos et al. 2020; Langfeldt 2001) or the average absolute deviation of individual proposal scores from the average score for a proposal (Baimpos et al. 2020; Pina et al. 2021). The second way is to examine the consistency of proposal scores, represented as an intraclass correlation (ICC; Jerrim and Vries 2020; Marsh et al. 2008; Mutz et al. 2012). An ICC value of zero represents no agreement, and a value of one represents perfect agreement.

The observed reliability of proposal scores was modest across all studies. Reviewers agreed with each other less than a third of the time, and reported ICCs were always below 0.54 (Exhibit 6). Unanimous agreement among proposal reviewers is uncommon, and the levels of absolute deviation are high relative to the likely differences in average proposal scores between competitive proposals. Total disagreement between reviewers — where none of them provided a score that was close to the median — happened about half as often as total agreement (Baimpos et al. 2020; Pina et al. 2021). Baimpos and colleagues (2020) even found that for 9 percent of proposals, none of the reviewers provided a score that was close to the median, indicating that they were divided about the merits of the proposal.

The reported ICCs are inadequate to reliably measure proposal quality (Cousens 2019; Portney and Watkins 2000), especially considering the relatively small differences in scores between proposals (Mohan and Brakaspathy 2018). By convention, an ICC above 0.8 is regarded as a sign of good reliability. An ICC below 0.5 means that more variability in grant scores can be attributed to differences between raters than to true differences between grant scores (Liljequist, Elfving, and Roaldsen 2019). Using regression modeling, Seeber and colleagues (2021) made a similar observation about the variance explained by rater and grant-level effects. An implication of low reliability (and large reviewer effects) is that among samples of raters, ranges of possible scores will be large relative to the differences between the scores of proposals that are above and below the funding cutoff (see Kaplan, Lacetera, and Kaplan 2008, as cited by Heyard et al. 2022).

Exhibit 6. Studies of reliability of merit review

Study	Sample disciplines	Proposals	Reviews	Rating scale	Reliability measure	Reliability score
Baimpos et al. (2020)	Science and technology	3,764	15,056[a]	1–5[b]	Median AD All agree[c] None agree	0.45 18% 9%

Study	Sample disciplines	Proposals	Reviews	Rating scale	Reliability measure	Reliability score
Jerrim and Vries (2020)	Social science	4,144	15,047	1–6	ICC	0.18
Langfeldt (2001)[d]	Science and technology	128	256[a]	1–5	All agree	30%
Marsh et al. (2008)	Science, social science, and humanities	2,331	10,023	1–100	ICC (project) ICC (researcher)	0.44 0.53
Mutz et al. (2012)	Basic science research	8,496	23,977	1–100	ICC	0.26
Pina et al. (2021)	All disciplines	75,624	226,872[a]	0–100	All agree None agree Mean AD	25% 12% 7.02 (SD = 4.56)

[a] This value is estimated from the number of proposals and the reported number of reviewers per proposal.

[b] Scored in half-point increments.

[c] Agreement means all proposal scores are within 0.5 scale points of the median score.

[d] Science and technology grants only.

AD = average deviation index; ICC = intraclass correlation; SD = standard deviation.

Comparisons of the reliability of proposal scores across disciplines have reported inconsistent results. Seeber and colleagues (2021) and van Arsenbergen, van der Weijden, and van den Besselaar (2013) reported much lower reliability across reviewers in the social sciences and humanities than those in chemistry, using data from the EU's MSCA program. However, both Marsh and colleagues (2008) and Mutz and colleagues (2012) observed marginally higher reliability across reviewers in the social sciences and humanities than those in the physical sciences.

Despite some researchers' suggestion that peer review is effective at identifying the top 20 percent of applicants (Fang and Casadevall 2016a; Mow 2011; van den Besselaar and van Arensbergen 2013; van Arensbergen, van der Weijden, and van den Besselaar 2013), we did not find compelling evidence that reliability is a concern only for weaker proposals. Some of the studies we reviewed did observe that funded grants tended to have more consistent proposal scores (Baimpos et al. 2021; Pina et al. 2021). However, for funders that rely on proposal scores to make funding decisions, any disagreement will pull proposal scores toward the midpoint of the evaluation scale and reduce the likelihood that a proposal is funded (Jerrim and Vries 2020; see also Roumbanis 2022; Baimpos et al. 2020; Pina et al. 2021).

D. Why do reviewers disagree with one another?

Reviewers might disagree with one another about the score of a proposal for several reasons. Some of these differences might be a good-faith difference in opinion. Others might be a result of errors that reflect the complexity of the task and limits on time, attention, or motivation. Finally, some disagreements about scores might reflect differences between reviewers in the

strategies they use to manage workload, different inclinations to be lenient or harsh, or different levels of experience with the peer-review process. We address each of these potential explanations in turn. We address potential reviewer biases for or against specific types of proposals or applicants in the following section.

Good-faith differences of opinion

Reviewers might assign different scores to a proposal because they have different assessments of some or all elements of the proposal. Scientific subdisciplines and topic areas differ in the questions they find interesting, and their priorities and trade-offs when selecting a technical approach to answer these questions. Further, as Lee and colleagues (2012:6) note, "review [dimensions]—such as novelty, soundness, and significance—may be open to different, normatively appropriate interpretations" that will also depend on the values and priorities of the proposal reviewer. For at least some scholars of peer review, disagreement is a feature of the peer-review process rather than a result of error, and one of its purposes is to invite divergent perspectives, stimulate debate, and redefine the meaning of quality (Langfeldt 2001).

We did not find studies of individual differences in the understanding of review dimensions, but a study by Neufeld, Huber, and Wegner (2013) hints at interdisciplinary differences. The authors used bibliometric data to predict proposal scores of 480 Life Sciences and Physical Sciences and Engineering grant proposals submitted to the European Research Council 2009 Starting Grants program. Reviewers were provided with identical instructions to consider significant publications in major international peer-reviewed scientific journals. Life Sciences reviewers seemed to focus on the instruction to consider publications in major outlets: success among Life Sciences proposal funding was predicted by the mean JIF of an applicant's publications but not the citation impact of their published papers. Conversely, Physical Sciences seemed to focus on the significance of the work: success among Physical Sciences proposal funding was predicted by the citation impact of published papers but not the impact factor of the outlets in which they published.

Disagreements in proposal scores can also result from proposal reviewers assigning different weights to the various elements and criteria that contribute to the final score (Roumbanis 2022). Consequently, reviewers who are in total agreement about the merits of the components of a proposal can still arrive at very different scores (Lee 2015; van den Besselaar, Sandström, and Schiffbaenker 2018).

Task complexity and limits on time, attention, or motivation

Reviewers encounter significant time pressure due to the volume of proposals and competing demands on their schedules (Brunet and Müller 2022; van den Besselaar et al. 2018). We were unable to find studies that reported how much time reviewers spend evaluating each proposal,

but observations of panel review sessions suggest that reviews are often incomplete at the beginning of panel review sessions and that very short amounts of time are dedicated to discussing each proposal (Langfeldt 2001). Notably, one study observed a U-shaped function in reviewer reliability, with initial improvement over the first seven reviews followed by a subsequent decline, suggesting the need to limit reviewers' workloads (Seeber et al. 2021).

Reviewers deliberately adopt strategies to efficiently review proposals. Instead of evaluating each proposal fully, they triage "noncompetitive" proposals using easy-to-evaluate elements before more carefully assessing the quality of the remaining proposals in the pool. Reviewers report that they begin the review process by assessing the curricula vitae (CVs) of applicants because CVs are well structured, familiar, and easy to evaluate (Brunet and Müller 2022; van Arensbergen et al. 2014b). Reviewers will examine prior outputs, including publication counts, citation impact, and co-authorship patterns, even though they recognize the limitations of these metrics for identifying good research (Langfeldt et al. 2021; van Arensbergen et al. 2014b).

Top-down review processes might be simpler than bottom-up processes that require a careful review of each element, but they can also lead reviewers to incorporate extraneous information into their assessment. Reviewers tasked with scoring multiple proposals reported ranking the remaining proposals rather than scoring each one individually (Brunet and Müller 2022). However, a consequence of this strategy is that proposal scores are influenced by the other proposals in the pool of applicants. Elhorst and Faems (2021) demonstrated that proposal scores are affected by the quality of other proposals in the same pool. Proposals received lower scores when grouped with higher-scoring proposals, and higher scores when grouped with lower-scoring proposals. Providing additional evidence about potential spillover effects across judgments, after observing an unusually high correlation between PI and project scores, van den Besselaar and colleagues (2018) speculated that reviewers focus on one dimension and adapt their score for the other dimensions rather than assess each component independently.

Reviewers might also avoid evaluating proposal elements according to complex dimensions entirely, substituting them with more easily evaluable dimensions. For example, ERC reviewers are expected to assess whether research is groundbreaking. For the ERC, this is defined as research that addresses important challenges and has ambitious objectives that are beyond state of the art (ERC 2023). Reviewers often overlooked these directions and instead applied simpler dimensions, such as novelty or social contribution of the project (Brunet and Müller 2022). In some cases, reviewers may even start with an assessment of a proposal and then work backward from this assessment to generate proposal scores consistent with this assessment (Lee 2015). These affective responses are motivated both by the quality of the research idea but also how the idea is sold by the proposal team and the quality of the writing (Porter 2005; Mow 2011; Lamont 2009, as cited in Roumbanis 2022).

Differences in reviewer disposition

Materia, Pascucci, and Kolympiris (2015) illustrated the impact of individual preferences by examining funding decisions made for 1,221 proposals submitted to an Italian agricultural research funding agency. Panel decisions were supposed to reflect the combined input of scientific merit scores from academic peer reviewers and the review panel's assessment of the suitability of the project for regional development. Although these factors were the strongest predictors of whether a grant was funded, the third strongest predictor of grant success was the number of proposals the review panel had previously approved, indicating that some review panels were inclined to be more lenient in how they defined fundable research.

Reviewer experience

Although each reviewer might begin their career with different practices in scoring proposals, over time their scores tend to converge. Marsh and colleagues (2008) observed that reviewers who evaluated three or more proposals produced lower proposal scores, but these scores were more reliable and closer to the final panel score. Some evidence suggests that this convergence represents reviewers learning about the practices and norms of a specific program, rather than becoming better reviewers. An examination of more than 50,000 proposals evaluated under the Horizon 2020, MSCA, and European Cooperation in Science funding programs supported this view. Within these panels, past reviewing experience in a specific funding program reduced disagreement with other reviewers, but general reviewing experience did not (Seeber et al. 2021). This study did not disentangle the separate effects of experience with a panel's instructions and procedures from experience working with the same panel members over time.

Interestingly, reviewers seem to be unaware that their scores become more convergent over time and feel that reviewer expertise is rooted in subject matter knowledge rather than review practices. Steiner Davis et al. (2020) sought to develop a list of skills thought to be important to peer review through a literature review, interviews with panelists and program officers, and a small survey of panelists. Interviewees generally felt that the skills necessary to succeed as a panelist could not be acquired over time. They identified the most important skills as subject matter expertise and the reviewer's scholarly track record. Communication skills, interpersonal skills, critical thinking, and willingness to change one's viewpoint appeared to be of secondary importance. Familiarity with the review process was seen as relatively less important, and skills related to broader impacts were not mentioned.

E. Are reviewers biased for or against certain proposals?

Studies of reviewer behavior have identified several reviewer characteristics that are likely to bias the process of reviewing proposals. We define bias as the systematic overstating or understating of the true value of a proposal's merit. Potential sources of bias include unwarranted skepticism

about novel or interdisciplinary ideas, conflicts of interest, the consideration of applicants' prior funding success, and bias against reviewers with certain demographic characteristics.

Bias against novel ideas

First, and most generally, reviewers have a reputation for "conservatism"—favoring incremental advances within established areas of research rather than innovative research (for a review, see Lee 2015). Funders often design review processes to reduce risk and so criteria and practices often encourage these tendencies (Rip 2000). As a result of this preference, researchers who take academic risks might be less likely to receive grant funding than their more conservative peers. Consistent with this hypothesis, Zoller, Zimmerling, and Boutellier (2014) found that researchers with less third-party funding tended to publish articles with a greater distribution of citation impacts than researchers with more third-party funding. That is, underfunded researchers' most impactful publication had a larger impact relative to the impact of their median publication. Zoller and colleagues interpreted this citation pattern as consistent with a high-risk, high-reward pattern of research, with a few projects that produced highly influential findings and many projects that did not work out.

Other researchers have pointed out that some aspects of the proposal review process encourage conservatism. Review elements that emphasize the plausibility or feasibility of planned activities are likely (by definition) to discourage scientific breakthroughs (Heinze 2008). Similarly, elements related to potential impact usually do not distinguish between the potential impact of a proposal and the probability that the impact will occur, forcing reviewers to try and capture both considerations in a single score (van den Besselaar et al. 2018). People's assessments of risk deviate from rational choice in systematic ways, leading to a preference for more certain outcomes. This may be especially true for proposals, which are characterized by many different sources of risk, not all of which can be addressed through a sound technical approach (Franzoni and Stephan 2022).

Bias against interdisciplinary research

Aside from their novelty, interdisciplinary research proposals might face unique challenges in the grant review process. Seeber and colleagues (2021) noted that it is hard to find reviewers familiar with all the disciplines relevant to interdisciplinary proposals. When grant funding is competitive, a single negative review can be sufficient to disqualify a grant from being funded, so reviewer misunderstandings about the methods or research questions important to other disciplines are a serious risk. Further, these risks might compound over time, as the challenges of publishing interdisciplinary work make applicants appear less productive than their monodisciplinary peers (Seeber et al. 2022).

The evidence for bias against interdisciplinary research is mixed. Seeber and colleagues (2022) examined 1,928 interdisciplinary grant proposals in the European Cooperation in Science and Technology (COST) research funding framework. They found that interdisciplinary grants were not penalized, but only 17% of COST grants are monodisciplinary, suggesting that reviewers had more experience reviewing interdisciplinary grants and that monodisciplinary grants posed less competition. An examination of funding patterns of 255 Sinergia grants evaluated by the Swiss National Science Foundation in 2008–2012 did suggest potential bias (Ayoubi, Pezzoni, and Visentin 2021). The authors constructed a novelty index for each applicant by identifying their published research and counting the number of papers that cited unique combinations of referenced journals. They observed that proposals from researchers with a high novelty index were graded 0.7 points (out of 6) lower, and they were 31%less likely to be funded than proposals from less novel researchers.

Conflicts of interest

Proposal reviewers are often expected to recuse themselves from evaluations of colleagues, research collaborators, or projects in which they might have a financial interest, and for good reason. Reviewers nominated by grant applicants (who are presumably colleagues or friends) provide substantially higher proposal scores than reviewers selected by the funder (Jerrim and Vries 2020; Marsh et al. 2008). These scores were inconsistent with the scores provided by other reviewers, leading inter-reviewer reliability to fall to nearly zero (Jerrim and Vries 2020). Despite this, a positive review from a nominated reviewer was strongly associated with a proposal being funded (Marsh et al. 2008).

Conflicts of interest are especially concerning because they are usually undetected. Gallo and colleagues (2016) examined the frequency of declared and undeclared conflicts of interest within eight panels of 14 or 15 reviewers responsible for reviewing 282 molecular and cellular biology grant proposals on behalf of the American Institute of Biological Sciences (AIBS). Reviewers were asked to review the institute's conflict of interest policies and declare any potential conflicts of interest. AIBS staff then reviewed the CVs of panelists to identify any undisclosed conflicts. Potential conflicts of interest were rare (only 66 conflicts were identified among more than 4,000 reviewer-proposal pairings), but CVs enable identification of only some types of conflict of interest. Importantly, only a third of these conflicts were self-reported by panelists.

Applicant's prior funding success

Researchers have observed that prior grant funding has been shown to relate to future success in securing funding, and these differences cannot be accounted for by increased academic outputs of funded researchers. Bol, de Vaan, and van de Rijt (2018) examined applicants to the NWO's Veni program who fell just above or just below the funding cutoff (applicants within two

points of the cutoff). Although both groups were equally productive, 26% of those who won an early-career grant went on to win a mid-career grant, whereas only 10% of those who applied for but did not win an early-career grant went on to win a mid-career grant. About half of this difference could not be explained by the increased likelihood that grant winners will apply for subsequent funding. Other researchers have also observed that successful grant applicants are not necessarily productive researchers. Instead, the relationship is curvilinear, with output plateauing or even dropping among highly funded researchers (van Leeuwen and Moed 2012; Mongeon et al. 2016).

Demographic characteristics of applicants

Researchers often raise concern about potential discrimination against applicants with specific demographic characteristics, especially if reviewers do not share those characteristics (Gallo, Sullivan, and Croslan 2022). Potential gender bias has received the most attention. A meta-analysis investigated the comparative grant funding rates of women and men for 66 different peer-review procedures (including 11 from NSF) from 21 studies (Marsh et al. 2009, using data collected by Bornman, Mutz, and Daniel 2007). For project-based proposals, the study found no evidence of gender bias in funding rates, and effect size estimates were homogenous. However, Marsh and colleagues (2009) did observe large and heterogeneous biases within fellowship applications, an observation supported by other studies of person-focused funding programs (Brouns 2000; van der Lee and Ellemers 2015).

It should be noted that these findings are focused narrowly on gender bias within funding rates and do not consider other potential systematic biases that might influence proposal patterns (Ranga, Gupta, and Etkowitz 2012), social barriers (such as more family obligations; Sato et al. 2020), or gender dynamics within the review team (Matiera et al. 2015).

Some studies have uncovered potential bias against applicants from less well-resourced institutions. Marsh et al. (2008) found that higher-prestige Australian universities also had higher grant funding rates — even for early-career scholars. Similarly, Piro et al. (2020) found that institution size and prestige (as measured by inclusion in the Shanghai Ranking, also known as the Academic Ranking of World Universities) predict funding outcomes for grant proposals to the ERC, even after accounting for the institution's citation impact. Murray et al. (2016) also found that institution size was associated with proposal scores. They examined 13,526 proposal scores for the NSERC Discovery Grant and found that funding success was 20% and 42% lower for established researchers from medium and small institutions, respectively, compared to their counterparts at large institutions.

Although there are clear differences in funding success rates across institutions, research has not established why this is the case, including whether this represents reviewer bias. Marsh and

colleagues (2008) speculated that a preference for proposals from prestigious institutions might be an indicator of validity rather than bias. Piro et al. (2020) suggested that differences in funding rates could be accounted for by the increased volume of proposal submissions. None of these studies controlled for differences in the quality of applicants hired or the institutional supports available to them.

F. How does panel discussion affect proposal evaluation?

Some funding streams require that reviewers meet to discuss proposal scores to resolve score disagreements or even to decide which proposals will be funded. Panel discussions tend to be unstructured and unpredictable compared to individual reviews. After reviewing individual reviews and sitting in on a series of panel review sessions intended to reconcile these scores, Langfeldt (2001:835) commented that "panelists do what they like, whereas individual reviewers do as they are told." Decades of research on group dynamics has identified distortions to judgment that occur during group discussion. In particular, when deliberating groups tend to discuss shared information and overlook any unique information that group members might have (Olbrecht and Bornmann 2010). The existing literature on grant review discusses three main aspects of panel discussions that might have an impact on proposal evaluations, including the content panelists bring forth for discussion, who among panelists speaks out during this discussion, and whether the conversation focuses on the positives or negatives of the proposal.

Unfortunately, there is usually insufficient time to thoroughly discuss proposals, and the success of specific proposals often depends on what panelists say or do not say during discussion. According to interviews conducted by van Arensbergen et al. (2014a), panelists who speak first heavily influence the resulting discussion. During panel review sessions, panelists often refer to their intuition about projects or grantees (which Roumbanis [2022] calls "epistemic-aesthetic feelings"). Reviewers also tend to spend more time disagreeing about the relative importance of proposal characteristics (such as innovation or rigor) than discussing the actual content of proposals. Panelists sometimes refer to other grants or the overall composition of awardees, and decisions often are reached through compromise rather than consensus.

Disagreements about the merits of a proposal often result in its rejection. Baimpos et al. (2020) evaluated 3,764 proposals submitted over seven grant cycles to the REA of the EC and found that panel discussion generally resulted in proposal scores becoming more negative; scores also flattened (became more uniformly distributed) and the number of proposals with the top score increased slightly. In qualitative interviews, Porter (2005) found that nine of 16 experienced panelists agreed that review panels discourage creative work in favor of incremental work.

G. Are assessments of grant proposals valid?

Researchers' and funders' concerns about the reliability of proposal scores and potential sources of bias are likely to be motivated by broader concerns about the validity of review panel or program director decisions, which is a far more challenging question to answer. It would be unsurprising if the validity of assessments of factual statements (such as whether a lab has sufficient resources or technical training to conduct a line of research) were quite high. Evaluations of grantee characteristics are also factual in a sense. There might be disagreement about the value of the number of publications, their outlets, or the scientific contribution of specific publications, but the facts that inform these decisions are certain and shared among reviewers.

Even when reviewing applicants' past achievements, validity might be difficult to achieve. Concurrent validity assesses the extent to which a measure can predict scores on other measures of the same characteristic. Santos (2022) examined grant evaluations in Portugal's Fundac 2020 and 2021 Individual Call to Scientific Employment Stimulus. This program is intended to fund promising scholars (rather than specific projects) and awards grants based on the applicant's CV, a CV synopsis, and a motivation letter. The overall sentiment of the written evaluation as assessed through Natural Language Processing was related (correlation = 0.44) to the final proposal score. Scores did not correspond to any scientometric measure of past activity except for sole authorship. The lack of a relationship between scientometric scores and funding decisions might speak to the importance of reviewer assessments of academic contributions, but the moderate relationship between comments and proposal scores provided by the same reviewer raises concerns about whether the scores reflect the merit of the applicant.

We did not locate studies that investigated the validity of planned activities, intellectual impact, or broader impacts, all of which have a degree of uncertainty. These assessments are difficult to evaluate for validity because an appropriate comparison group (for example, a proposal that reviewers rejected but had the same funding resources as accepted projects; Harangel 2019) is usually unavailable. The evaluation of potential impacts is especially challenging because the existence or magnitude of an impact is highly dependent on the time frame examined (Shaw 2023): some lines of research are initially very impactful before essentially dying out, whereas other findings lie dormant for years before they are recognized as a precursor to breakthrough research.

Researchers are cautious about the potential validity of proposal evaluations. Studies have revealed that other experts in information-rich domains—political scientists predicting future events, investment advisors predicting financial returns, and firms predicting the value of research and development efforts—have only modest success at predicting future outcomes,

despite a high degree of confidence in their ability to do so (Cunha et al. 2012; Fang and Casadevall 2016a).

H. Evidence-based practices that improve processes for reviewing grant proposals

Even amid the recognized limitations of the grant peer-review process, few empirical studies investigate methods that might improve the reliability and validity of reviews of nonmedical STEM research proposals. Though the empirical evidence base is not yet robust, researchers and funders have put forth several ideas hypothesized to improve aspects of the grant review process.

One solution for improving the reliability of peer review is to increase the number of reviewers. However, this is likely to be impractical, given that reviewers already have significant demands on their time. Marsh and colleagues (2008) calculated that it would take six reviewers per proposal to achieve an ICC of more than 0.7 for project quality and 0.8 for team quality.

Other researchers have proposed that increasing the granularity of the scoring system might improve reliability. A simulation that examined the reliability of proposal scores found support for this claim. However, the simulation assumed that reviewers form a proposal score more granular than the available response options, which they must then map onto the available response alternatives (Feliciani et al. 2022). Langfeldt (2001) suggested that less granular scores result in more tied scores, and these scores allow decision makers the flexibility to construct a portfolio of funded research that addresses multiple policy objectives without passing over more meritorious proposals.

If reviewers tend to overlook or underweight some proposal elements, review processes that require scores for each element could be more reliable because they encourage reviewers to consider each element. Individual criteria or element scores might also be more transparent to program officers if they provide another means of understanding how the reviewer integrates information about different criteria or elements. We did not find studies that compare this kind of bottom-up scoring approach with an approach that asks reviewers to assign a global score. However, Pina and colleagues (2021) reported that standardizing and reducing the number of elements (from four or five to three) had no impact on consensus among reviewers in the MSCA program.

Several researchers have suggested using Multiple Attribute Decision Making tools to support the evaluation process.[7] As a basic example of how these tools could be used, program officers can define proposal attributes (criteria and elements) and assign each attribute a weight (how important it is). Reviewer scores for each attribute are aggregated according to their weight

[7] Multiple Attribute Decision Making is a field of study that concerns optimizing choice between alternatives that vary simultaneously on several attributes. For an overview, see Wallenius et al. 2008.

(Dwitayanti and Amin 2023). This approach can be extended by assigning each attribute a value function that allows score differences to have a smaller or larger effect depending on where they fall on the evaluation scale. The value function is either chosen by the program officer or derived from scoring data (Parreiras et al. 2019). The model can also be extended to incorporate reviewers' certainty about their judgments (Öztayşi et al. 2017). An appealing feature of these models is they enable funders to examine how changing attribute weights might affect scores on other attributes within the grant portfolio. In a field test of this approach, program officers reported that it encouraged reviewers to carefully consider the importance of attributes and facilitated strategic portfolio selection (Parreiras et al. 2019).

Others have suggested using bibliometric tools to support the review of applicants' research track records. Even though bibliometrics provide an incomplete picture of applicants, reviewers often provide feedback about quantitative aspects of applicants (Cousens 2019). Győrffy, Weltz, and Szabó (2023) described a bibliometric tool the Hungarian Scientific Research Fund used that extracts bibliometric statistics for applicants, adjusts scores by discipline, and ranks them within strata determined by career stage. Similarly, Cañibano, Otamendi, and Andújar (2009) developed several models that score applicants based on their CVs. Using seven variables, the tool produced funding decisions that agreed with reviewers' funding decisions at least 82% of the time. To avoid unfairly rejecting applicants, the authors proposed running a set of predictive models and rejecting applicants only when all models recommended rejection.

To address differences in reviewer harshness or leniency, Kuhlisch and colleagues (2015) proposed statistically adjusting reviewers' proposal scores. They demonstrated that doing so resulted in substantial changes to the set of papers accepted in a data set of conference abstract submissions but did not establish the preferability of one set of funded applicants over another.

Finally, turning to panel review sessions, several researchers have examined how panels perform in person versus virtually. In interviews, reviewers and program officers reported virtual panel reviews generally require more sustained attention and interpersonal skills than in-person reviews; however, the increased structure of virtual panel conversations might discourage panelists from dominating the conversation or talking over their colleagues (Pederson and Husu 2022; Steiner Davis et al. 2020). There is little evidence that virtual panels affect panelists' decisions. Pina et al. (2021) found that consensus scores were unaffected when the MSCA grant program panel switched from in-person to virtual panel discussions.

I. Proposed alternatives to traditional peer review

Although peer review is the gold standard for funding decisions and widely used by funding agencies (Biegelbauer, Palfinger, and Mayer 2020), its lack of reliability, potential for bias, and uncertain validity have led some researchers to propose other methods of funding grants (Ayoubi et al. 2021; Franzoni and Stephan 2022; Holbrook and Frodeman 2011). Some

researchers have made this argument from a pragmatic perspective, noting that the amount of effort required to make the fine-grained distinctions required by funders is impractical (Bedessem 2020). Others have noted that the review process encourages "grantsmanship" — that is, scientists focusing on optimizing proposals to score well, rather than optimizing the merit of the work the grants would support (Gross and Bergstrom 2019; Smaldino, Turner, and Kallens 2019).

Expand the definition of "peer reviewer" and include non-academic reviewers

Several funding agencies have expanded their definition of peer reviewer to include people other than academics working in the field of study relevant to the proposal. Some funders have sought to include industrial experts, education experts, public outreach professionals, "user evaluators" (medical patients), or "research users" (policymakers) as reviewers. These reviewers provide feedback about impacts and feasibility for the user community (most commonly within applied medical research; Luo, Ma, and Shankar 2021).

Researchers argue that including non-academics in the review process could help ensure that the review process pays appropriate attention to broader impacts, increases confidence in the evaluation process, increases the epistemic diversity of science, and counterbalances any special interests of scientists (Gunn and Mintrom 2017; Santana 2022). Some evidence shows that including non-academic reviewers in proposal reviews helps ensure the careful consideration of potential broader impacts. Luo et al. (2021) examined panel review reports from Science Foundation Ireland and found that panel comments about the broader impacts of a proposal were four times longer when non-academic reviewers were included in a panel than when they were not included. Panels with non-academic reviewers also tended to describe impacts, beneficiaries, and potential risks in more specific terms.

Dramatically expand the number of reviewers through crowdsourcing

Some researchers propose solving the low reliability of proposal funding decisions by radically expanding the number of scientists who provide input about funding decisions. Bedessem (2020) proposed that individuals (scientists or members of the public) vote on specific projects to determine which ones to fund. Similarly, Bollen (2018) and Bollen et al. (2014, 2019) proposed that scientists each receive a fixed budget that includes funds they must distribute to other researchers. This method would allow all scientists to provide input into how funding is distributed, but because this process is iterative, scientists who receive significant funding from other researchers would in turn be more influential. NWO was reportedly considering a pilot of this approach (Bollen 2018), though findings from any implementation of it were not available.

Make peer review interactive

In most models, the only interactions between grant applicants and reviewers are the funding decision and reviewer comments, but some models seek to make this process more interactive. A notable example is EPSRC's IDEAS Factory, which uses a sandpit selection process in which applicants, active researchers, and research users participate in a five-day residential workshop and collaboratively allocate funding (Heinze 2008). The National Natural Science Foundation of China proposed a similar method (Wang, Li, and Zheng2011), but it is unclear whether it was implemented. These approaches allow for real-time feedback and amendments to the proposed work and encourage deeper discussions of it.

Replace or supplement peer review with a lottery system

Lottery systems are the most widely discussed alternative to peer review, mentioned by at least 14 articles. Some of the interest in lotteries is that they form a useful comparison to measure the effectiveness of peer review. They are also appealing because of their potential to increase efficiency (Roumbanis 2019; Phillips 2021) without substantially reducing the accuracy of funding decisions (Harnagel 2019). Others have suggested that lotteries might motivate lower-quality proposals (Cousens 2019; Horbach, Tijdink, and Bouter 2022).

Most proposed lottery designs include peer review as a first step to ensure that all proposals entered in the lottery meet a minimum standard of quality. The lottery is then either applied to all qualified proposals or to those close enough to the funding cutoff that the role of chance within the peer-review process is likely to be high (Fang and Casadevall 2016b; Heyard et al. 2022). Shaw (2023) described the various kinds of lotteries in detail, including restricting the lottery to specific types of proposals, weighting the probability of selection based on proposal scores, or stratifying applicants into categories before selection.

Currently, some funders are exploring whether lotteries can play a role in funding disbursement. The Health Research Council of New Zealand (HRC) distributes Explorer Grants and Science for Technological Innovation Challenge Seed Grants through a lottery system. Also, the Volkswagen Foundation awards both panel-selected and lottery-based grants (Shaw 2023; Roumbanis 2019).

Distribute grant funds evenly among all proposals

A few have argued that funding should be allocated evenly among all of those who apply in the service of equity. However, given the sheer number of proposals that funders receive for any given solicitation, such an approach would not be feasible. In most cases, the funding amounts allocated would not match research needs and so would likely be insufficient on their own. Ironically, this strategy could possibly benefit more established researchers with secure positions more than junior researchers or those without much additional funding (Roumbanis 2019; Vaesen and Katzav 2017).

Researchers have also suggested increasing the number of funded applicants by reducing the size of grant awards. Dresler (2022) proposed a variation of the equal funding mechanism wherein an initial set of applicants are funded through a process requiring low reviewer effort (such as base funds, lottery, or light-touch peer review) and receive a portion of the allocated funds. Researchers could then access the remaining funds through a demonstration of methodological rigor and merit, such as a peer-reviewed, pre-registered study plan. Others have proposed a continuous funding system that allows applicants with high, though not outstanding, proposal scores to adjust their project scope to align with a budget proportional to their proposal score (De Los Reyes and Wang 2012; Mutz, Bornmann, and Daniel 2016).

Use quantitative methods of predicting researcher output

Last, some have argued for the use of predictive models to forecast an applicant's output, which could be used to make funding decisions. Previous publication outputs explained 87.5% of the variance of near-term scientific impact of a researcher, as measured by their h-index.[8] Using this modeling approach, nearly a quarter (23.5%) of those researchers classified as future high performers in fact became high performers. Almost all of this predictive power for future scientific impact came from the baseline measure of impact (Kuppler 2022). Győrffy, Herman, and Szabó (2020) found similar results: in an analysis of 13,303 Hungarian basic research proposals, past bibliometric measures of academic performance (publication in top quintile journals and citation rates, excluding self-citations) predicted future publications in top quintile journals (correlations between .46–.79), whereas reviewer evaluations did not (correlations between .08–.11). This approach might be sufficient if a funder's only outcome of interest from a grant award is to support researcher output.

[8] The h-index is a metric of scientific output and impact, calculated as the largest number h, such that h articles have at least h citations each.

5. Discussion by Research Question

A. RQ1: Do funders specify or define the relative emphasis reviewers should place on intellectual merit and broader impacts? If so, how much emphasis is placed on each of these facets?

For the government funders we reviewed, the importance of broader impacts varies across funding streams but is generally secondary to intellectual merit. NWO is the only organization to assign quantitative weights to the importance of different project criteria or elements for reviewers. In NWO funding lines, intellectual merit usually carries more weight than broader impacts. The DFG does not require reviewers to comment on the broader impacts of proposed research, nor does the agency appear to consider these impacts as a part of funding decisions.

UKRI funding streams consider potential broader impacts, but in a way that cuts across themes. For example, EPSRC uses excellence as a primary criterion but evaluated excellence according to novelty, ambition, and suitability of the approach (which NSF would consider as relevant for intellectual merit), as well as social impact elements. However, UKRI programs evaluate proposals using more intellectual merit elements than broader impacts elements and the elements refer to more entities.

Foundations generally provide less information about their decision-making processes, making it challenging to determine the relative importance of broader impacts and intellectual merit. However, for the Bill & Melinda Gates Foundation and the Gordon and Betty Moore Foundation, broader impacts could be characterized as necessary for successful funding. The Bill & Melinda Gates Foundation defines funding streams as challenges to achieve specific broader impacts. Although these challenges do require basic scientific questions to be answered, the Foundation's funding materials explicitly state that, for the Foundation, these are a means to an end. The Gordon and Betty Moore Foundation also focuses on social benefits and develops projects that can attain goals they share with applicants.

The Alfred P. Sloan and MacArthur Foundations appear to leave more room for grant proposals that emphasize intellectual merit to succeed. For example, Sloan provides funding streams defined by basic scientific topics and does not explicitly require grantees to address broader impacts. The MacArthur foundation evaluates proposals using elements that could apply to either intellectual merit or broader impacts and does not explicitly require that successful proposals address both criteria.

Scholarly literature suggests that when reviewing grant proposals, reviewers tend to care more about the intellectual merit of the proposal than its broader impacts. The taxonomies developed and applied by grant reviewers contain more elaborate representations of elements related to

intellectual merit than of elements related to broader impacts, suggesting a stronger emphasis on this criterion. In two surveys, reviewers and program officers said they consider intellectual merit to be more important. In fact, the emphasis reviewers place on intellectual merit may go beyond what funders intended. One study even found that grant reviewers use space intended to discuss broader impacts to discuss intellectual merit instead.

B. RQ2: What practices and elements do funders use to evaluate intellectual merit?

The government agencies we examined (GRF, NWO, UKRI) rely primarily on peer review to evaluate intellectual merit. In all cases, a separate review panel reconciled reviewers' proposal scores through discussion. NWO has the most elaborate process, using expressions of interest to filter out applicants who are unlikely to be competitive before sending proposals to peer review and allowing applicants to respond to written comments before their application moves on to panel review. Foundations generally rely on staff members to review projects, but occasionally use external reviewers to provide technical input. The review processes some foundations use could also be described as more collaborative, with program officers providing substantive input during the final proposal activities.

Almost all funding lines supported by these agencies and foundations consider the potential intellectual outcomes of the proposed research and whether the proposed research activities are sound. The sole exception is the MacArthur Fellows program, which evaluates the merits of ideas, but appears to trust the applicant's track record rather than a defined technical approach. Most of the funding lines also consider whether the applicant is qualified or has sufficient resources to pursue the project. The publicly available guidelines that funding agencies — especially philanthropic foundations — provide to reviewers about these elements are quite broad, and there is no guidance about how much emphasis to place on these different elements when assessing the merit of the proposal.

C. RQ3: What evidence exists to support the reliability, validity, or efficacy of the processes used to evaluate intellectual merit?

All available evidence suggests that individual proposal reviewers cannot reliably assess the intellectual merit or overall merit of proposals. The variability in reviewers' proposal scores implies that the credible interval (that is, the range of possible average scores a large group of reviewers would provide) around reviewer scores is quite wide. At the same time, the merit scores of funded and many (but not all) unfunded proposals are often close together, leading to a high degree of unpredictability about whether a proposal might be funded.

Many factors contribute to a lack of reliability. The literature we reviewed did not report the consistency with which reviewers assign the same score to a proposal evaluated at different

times, which places an upper bound on the consistency expected between reviewers. Further, reviewers may disagree with each other for many reasons. These include healthy disagreement between reviewers when evaluating elements of a proposal or defining the dimensions used to evaluate them (the extent of which might vary by topic or discipline), differences in the weight placed on various elements or criteria when forming a composite score, consideration of elements not specified within the review guidance, inexperience with the review process, conflicts of interest, and bounds on time and attention that make it difficult to extract and properly integrate all relevant information from a proposal. One implicit assumption behind panel discussion is that it is easier to achieve consensus among reviewers when scores are based on more complete information. We were unable to locate evidence about how downstream decisions by program officers or other decision makers are made, including the consistency of their funding decisions, whether they calibrate their decisions based on information about the reviewers, and if they influence reliability of grantee evaluations.

The concept of validity itself is difficult to define without clearly specifying the timescale and nature of the impact, but very little evidence supports the validity of intellectual merit scores by any measure. Several observations suggest that validity of proposal scores is likely to be low. In general, experts' ability to predict the future impact of events is modest at best. The reliability of a measure usually places an upper bound on its validity, and the reliability of merit scores also are modest at best. Finally, reviewer scores (when evaluating applicants) correspond only weakly with the sentiment of their own comments and not at all with scientometric indicators (though in this case, it is possible the score is valid and the other measures are not).

The efficacy of the review process depends on the goals that funders are trying to achieve. The competitiveness of funding likely means that only meritorious proposals are funded (that is, few if any of the funded projects are of poor quality). Most of the concerns raised about the grant peer-review process by those proposing alternative processes for evaluation are about whether the *most* meritorious proposals are funded, which is critical for ensuring a fair process. It is less clear that peer review achieves this goal because of the low reliability and unknown validity of proposal scores. However, scholars have suggested numerous alternatives to peer review intended to make the process more efficacious. Some of these alternatives are being tested on a small scale.

We identified very few articles that examined the impact of interventions on nonmedical STEM proposals. The literature suggests that within funding programs, differences in reviewer characteristics like expertise, disposition, and evaluation strategies play a bigger role in proposal score variability than differences in grant characteristics. This suggests that protocols to reduce bias and ensure equal treatment of applicants, although important to ensure fairness and equity, are unlikely to address substantially the aggregate validity of funding decisions.

There are several plausible ways to reduce reviewer effects. Increasing the number of reviewers would improve reliability by reducing the impact of drawing a reviewer with an opinion that differs from consensus, but the number of needed reviewers is likely to be impractical. Experience with a funder's review procedures was associated with more reliable evaluations, suggesting consistency can be learned. We did not find studies on the impact of training on reviewer performance for nonmedical STEM proposals.

We found more evidence in support of computational tools to sustain the various complex information integration tasks a review might entail. They include the following:

- Calculating normalized bibliometrics, which are often discouraged as a tool to evaluate an applicant's merit but might be helpful in certain situations, such as when evaluating applicants from different disciplines, or when behavioral studies reveal that reviewers are trying to calculate these scores on their own as a part of their review process.
- Integrating the value of outcomes and the probability these outcomes will be realized.
- Applying a consistent method when weighting and aggregating elements and criteria into a single proposal score.
- Consistently adjusting the scores provided by historically harsh or lenient reviewers to reduce the impact of their predisposition on a proposal's final score.

Finally, some proposed but unproven alternatives to peer review seek to address the perceived shortcomings in reliability or validity by either expanding the number of decision makers dramatically (through crowdsourcing) or removing them entirely (through using quantitative metrics).

D. RQ4: What practices and elements do funders use when evaluating broader impacts?

According to their published review criteria, NWO and UKRI ask reviewers to comment on the broader impacts of research, but the GRF does not. Except for the UKRI Engineering and Physical Sciences Research Council, funding lines that require reviewers to comment on broader impacts seem to focus on the potential outcomes of the research project rather than the planned activities, the applicant's capabilities and track record for achieving broader impacts, or available resources that might support these outputs. The guidelines provided to reviewers about these elements are quite broad, except for UKRI's EPSRC and BBSRC, both of which specify impacts relevant to UKRI strategy and national interest more generally.

The philanthropic foundations we examined appear to place more emphasis on broader impacts. They might pay more attention to the proposed method and applicant's capacity to realize these impacts, though the elements they describe are too vague to be certain. This

vagueness is likely offset by the more collaborative role that program officers at philanthropic organizations play in shaping proposed outcomes and planned proposal activities.

Broader impacts criteria tend to focus on beneficial outcomes of research. Funding criteria overwhelmingly emphasized elements related to research outcomes. Further, proposal review processes do not instruct reviewers to address the potential harms caused by the process of conducting research or its outcomes. Proposal review processes also do not consider how to address potential inequalities in who might realize the benefits of the proposed activities. Expanding the evaluation of broader impacts to address these proposal elements might require other complementary changes to be effective, such as modifying proposals to require this information and including reviewers with experience in evaluating evidence-based practice.

E. RQ5: What evidence exists to support the reliability, validity, or efficacy of the processes used to evaluate broader impacts?

Less evidence supports the reliability, validity, or efficacy of the processes used to evaluate broader impacts than that supporting intellectual merit. However, there appear to be blind spots in proposal review elements that might make assessing broader impacts especially challenging. Review elements assess the overall potential for a proposal to have a broader impact but do not appear to consider how planned activities, an applicant's track record, or existing resources might support these outcomes. To the extent that reviewers are better able to agree on the soundness of procedures than their eventual outputs, it is likely that reviewers will display less consensus about broader impacts than about intellectual merit.

According to at least one view, interdisciplinary research inherently provides a broader impact, in that it fosters the transmission of information from one field or sector to others. There is mixed evidence that reviewers penalize research with broader impact when defined in this way.

The challenges of establishing the validity of broader impacts are similar to the challenges of establishing the validity of intellectual merit. The literature does not specify on what timescale broader impacts should be assessed or what elements should be used. Although measures of academic impact, such as publications and publication impact, are sometimes fairly criticized as overly simplistic, broader impacts lack analogous concepts (with the exceptions of mentions in social or traditional media or policy documents). Some broader impacts can be quantified, such as the number of community members a research team engages or the number of trainees a grant supports, but it is unclear how these figures would be translated across disciplines or outcomes.

Given the uncertain reliability of broader impact measures, it is not surprising that we did not identify trainings or practices that demonstrably improve the reliability, validity, or efficacy of their evaluation. However, we did identify several proposals that include perspectives from

outside of the academic community in assessments of broader impacts. It is unclear whether this practice would improve the reliability or validity of evaluations of broader impacts beyond the benefits of increasing the number of reviewers. This approach would also be appealing if it enhanced the fairness or transparency of the evaluation process.

F. RQ6: Are there important gaps in the literature on the assessment of the merit of sponsored research that are not otherwise addressed by research questions 3 or 5?

There appear to be many important gaps in the literature regarding assessment of the merit of nonmedical STEM research that are important for understanding how reviewers assess the broader impacts of proposals. These gaps make it difficult to evaluate the difficulties reviewers might face when assessing broader impacts, or how to address them. Insight into some of these questions might be available in adjacent literatures.

The research we reviewed provides very little insight on how proposal reviewers think. There is some data on what they say — either in panel review sessions or in written comments — and more data on their final proposal scores, but almost nothing about how reviewers might arrive at their conclusions. The studies that come closest to shedding light on reviewers' decisions make inferences based on either retrospective self-report or correlations between proposal characteristics and outputs. As a result, several surprisingly basic questions arise about reviewer behavior. For example, consider the following:

- How much time do reviewers dedicate to proposals, and how do they allocate this time across proposal elements?
- To what extent can reviewers understand review elements and align their definitions of certain dimensions, such as importance or rigor, with the funder's definition?
- Do reviewers assess proposal elements or criteria and integrate their assessments into a final proposal score, or do they form an overall impression of a proposal and use the criteria to justify their assessment?

All these questions might be fruitfully addressed through think-aloud methods adopted from usability testing or survey research.

We did not locate studies of reviewer reliability that provided insight into how different review elements contributed to reviewer disagreement. Poor reviewer reliability can be explained by disagreement about the importance of different elements of a proposal or disagreement about the quality of these elements. Reviewers might also disagree more about some elements than others. The most immediately relevant question to NSF might be whether the reliability of broader impact scores differs meaningfully from the reliability of intellectual merit scores. However, reviewers might also reach different levels of consensus about the elements that feed

into these scores, including the importance of entities like the applicant, activities, and scientific outputs, and the dimensions on which these entities should be evaluated. Identifying the areas of reviewers' disagreement might help identify ambiguity in NSF's reviewer instructions, help panelists focus conversation, or help program officers contextualize reviewer comments.

The literature on reviewer reliability also focuses exclusively on the consistency of reviewer scores: we found no evidence of how much overlap exists between the strengths or weaknesses noted in reviewer comments. However, if scores differ because reviewers notice or emphasize different strengths and weaknesses, then written comments about proposals are also likely to differ. The degree of overlap between reviewers' comments about a proposal would provide some insight into how many reviewers are needed to ensure that important considerations are not overlooked.

Our review revealed very little information about how reviewers assess the potential broader impacts of research. One important question is how expertise about broader impacts affects reviewer practices. Reviewers' expertise is likely to vary by both discipline and type of impact or activity. Do reviewers focus on or overweight topics on which they have expertise (such as teaching or dissemination)? How do they evaluate activities with which they are unfamiliar? More generally, how is the potential impact of an activity estimated? It is not clear how reviewers integrate factors such as the number of people a project might benefit, the potentially transformative potential for those who do benefit, and the fairness of the distribution of these inputs.

We also did not find evidence about the correlation of broader impact scores with intellectual merit scores. If these scores are strongly related, it is important to understand why. Although identifying examples of scientists who focus more on basic or applied science is possible, in practice, both criteria usually go together. Broader impacts might be conditional on some of the planned activities of the intellectual merit criterion because improperly conducted research is unlikely to produce outputs with tangible benefits. However, reviewers might score elements carefully on dimensions relevant to intellectual merit (like innovativeness or rigor), which is their area of expertise, and then use this score as a starting point when assessing elements on other dimensions.

The research we reviewed also leaves many unanswered questions about how panel discussions function. Overall, the literature suggests that the time constraints and open structure of panel discussions make the panel review process less predictable than the processes reviewers use to form independent scores. The unpredictability makes it unclear whether panel decisions should replace or supplement individual reviewer decisions. It also suggests there could be ways to add structure to panel discussions to increase fairness if it does not impede thoughtful debate. For example, groups tend to be better at sharing information relevant to a decision when they have

sufficient time to discuss information, when individuals are assigned responsibility, and when information sharing and decision making are separated into discrete tasks (for an overview see Kerr and Tindale, 2004)

We found very little information on the decision processes of program officers, division directors, or other nonreviewers. Reviewers or program officers might apply criteria that are different from those specified by funders. Some descriptive research on the deliberations of panels with the authority to make funding recommendations suggests considerations that matter for constructing grant portfolios, but these data are qualitative and bound to a specific time and context that might not generalize to NSF. It is plausible that general observations about the importance of expertise, how reviewers manage workload, and susceptibility to bias among reviewers could generalize to other decision makers. However, roles and context matter, so these processes might unfold differently.

The literature we reviewed included several examples of tools that could support funding decisions, but few studies of how to use them or whether they are helpful. Grant evaluation is a complex process that not only requires evaluating many different elements of a proposal using many different dimensions, but also understanding how these parts fit together. Decision aids can potentially simplify the review process or make it more consistent. These tools can be used in many ways, potentially replacing elements of peer review or supporting or supplementing reviewer decision making.

To use bibliometric scoring as a concrete example, we identified studies that describe the predictive power of bibliometric analysis but none that compares the different ways this tool could be integrated into the review process. A funder could decide that publication metrics adequately measure researcher quality and replace reviewer scores entirely. Or, if a funder observes that reviewers are determined to use some form of counting as a part of the assessment process, they could support this process by providing consistent and preferred metrics or benchmarks that reviewers adjust or contextualize. As another approach, bibliometric analysis could supplement reviewer scores by alerting decision makers about cases where scores seem unusually high or low. This example is not intended to endorse bibliometric tools, but to show that the value of any tool is closely linked to the practices it supports.

In our literature search, we encountered many studies on the assessment of medical research and the ex post assessment of research activities that might address some of the gaps we identified. Other literature related to ex ante predictions in related fields (such as private sector research and development planning and forecasting) and organizational decision making might tentatively address some of these gaps or suggest the most important research questions within them.

6. Implications for the process evaluation of NSF's Broader Impacts review criterion

The primary implication of findings from the literature review is to inform the process evaluation design (consisting of a document review, interviews and focus groups, and extant data analysis), which is structured around answering three research questions:

1. In what ways do the interpretations of the Broader Impacts review criterion among PIs, reviewers, and NSF program staff vary, and what factors might contribute to these variations?
2. How do external reviewers assess the Broader Impact review criterion, and how do NSF Program Officers and division directors factor these assessments into award recommendations and decisions?
3. In what ways do PIs, reviewers, and NSF program staff perceive that variations in interpretation and assessment can advance or hinder the merit review of proposals and ultimately support NSF in meeting its Broader Impacts review criterion across its programs?

Specifically, the literature will inform benchmarks when assessing the alignment of NSF's Broader Impacts review criterion with the promising policies and practices described in this literature review report. Additionally, these findings will direct us to explore evidence gaps identified in the literature related to how people interpret and apply broader impacts criteria. As we develop the analytic approach and data collection instruments for the process evaluation, we are mindful that existing research on the use of broader impacts as a review criterion is exceptionally thin. The literature review provides a useful starting point for defining some of the questions and areas of concern for further exploration.

The process evaluation's document review will summarize the state of the NSF Broader Impacts criterion and map current NSF policies and practices to promising practices identified in the literature review. The document review will focus on how the NSF Broader Impacts criterion is communicated within the organization, with merit review participants such as PIs, reviewers, and other constituencies, via NSF sponsored documents as well as external assessments of merit review participants outside of NSF via the scholarly literature. The review will reveal areas of alignment and misalignment between how the Broader Impacts criterion is communicated by NSF, interpreted by proposal reviewers, and experienced by grantees. The document review may also discover novel practices unique to NSF that are worthy of further study.

Interviews and focus groups with NSF staff, PIs, and reviewers will delve deeper into merit review participants' experiences with the Broader Impacts criterion and may shed light on a key gap identified in the literature: understanding *how* reviewers arrive at their assessments of a proposal's broader impacts. For example, interview protocols include questions about NSF staff's own interpretations of the Broader Impacts criterion, their process for making funding recommendations, the guidance they provide to reviewers, and their experience with how

closely reviewers apply the elements and guidance to their proposal assessments. The focus groups' protocols include questions about reviewers' interpretations of the Broader Impacts criterion, the emphasis they place on NSF Broader Impacts as they assess proposals, how they weigh the value of different proposed outcomes of Broader Impacts (for example, engaging the public compared to developing the STEM workforce), and whether they embody promising practices during the merit review process.

Finally, the analysis of extant text data from NSF's merit review survey and Review Analysis documents will examine several questions raised in the literature. This includes understanding NSF's merit review process documentation, how PIs and reviewers describe their experiences with the Broader Impacts criterion in their feedback, how much emphasis project directors at NSF give to the Broader Impacts criterion in their Review Analyses, and whether there is a correlation between the sentiment of comments on intellectual merit and the sentiment of comments on broader impacts.

Using the literature review to inform the process evaluation will ground our understanding of NSF's approach to broader impacts in a larger evidence base. Identifying promising practices from the literature and understanding NSF practices in relation to them will help us provide NSF with findings that are actionable and can lead to evidence-informed recommendations.

Appendix A: Methodological Details

A1. Strategies used to adhere to NSF's Evaluation Policy

Exhibit A.1. Key principles of NSF's Evaluation Policy and related features

Principle	Features of this study that align with this principle
Relevance and utility	• We presented literature review processes to NSF to ensure that the team would capture information of interest to NSF. • We will present a final briefing to NSF staff for use as a basis for future conversations and decisions about evaluating the no-deadlines approach.
High quality and rigor	• We used several data sources to address the research questions. • In this report, we include clear communication of findings and limitations. • A senior team member reviewed screening decisions and information from literature reviews to ensure accuracy and completeness.
Independence and objectivity	• An independent reviewer reviewed the written report. • In this report, we include all findings, whether positive, indeterminant, or negative.
Transparency and reproducibility	• We defined study objectives and study design before to starting the study. • We documented the literature search strategy and inclusion criteria before beginning the literature search. • We recorded search strings, databases, access date, and number of results for all searches. • We developed standardized literature review templates before conducting reviews. • In this report, we clearly explain methods and findings.
Ethics	• Discussion of findings includes contextual factors that could influence interpretation of findings.
Equity	• We used rigorous and inclusive screening criteria for published articles and ensured that we reported findings relevant to the equity of grant proposal evaluations.

Source: Adapted from NSF's Evaluation Policy, April 2023.

Note: This exhibit demonstrates the ways the study adhered to NSF's Evaluation Policy by listing the features of the study that contributed to upholding each principle.

A2. Website searches of organizations that make funding decisions

The webpages we consulted while reviewing the funding practices of federal research agencies and philanthropic foundations are listed in Exhibit A.2.

Netherlands Organisation for Scientific Research (NWO)

Netherlands Organisation for Scientific Research. 2022. "SSH Open Competition M Round 2022 Call for proposals." (https://www.nwo.nl/sites/nwo/files/media-files/CfP%20SGW%20Open%20Competitie%20M%20def_Eng.pdf).

Netherlands Organisation for Scientific Research. 2023. "NWO Talent Programme Veni 2023 Call for Proposals." (https://www.nwo.nl/sites/nwo/files/media-files/Veni%202023%20Call%20for%20Proposals%20-%20ENG.pdf).

Netherlands Organisation for Scientific Research. 2023. "NWO Talent Programme Vici 2023 Call for proposals." (https://www.nwo.nl/sites/nwo/files/media-files/Call%20for%20Proposals%20Vici%202023%20EN%20%2812-7%29.pdf).

Netherlands Organisation for Scientific Research. 2023. "Open Competition Domain Science - M 2023/2024." (https://www.nwo.nl/en/calls/open-competition-domain-science-m-2023/2024).

Netherlands Organisation for Scientific Research. 2023. "Open Competition Domain Science-M round 2023-2024 Call for proposals." (https://www.nwo.nl/sites/nwo/files/media-files/Call%20for%20proposals%20OC%20ENW-M%20ronde%2023-24_EN_def.pdf).

Netherlands Organisation for Scientific Research. 2023. "Open Competition Domain Science – XL Round 2023-2024 Call for proposals." (https://www.nwo.nl/sites/nwo/files/media-files/Call%20for%20Proposals%2023-24_UK%20%281%29.pdf).

Netherlands Organisation for Scientific Research. 2023. "Open Technology Programme 2024 Call for proposals." (https://www.nwo.nl/sites/nwo/files/media-files/CfP_Open%20technologieprogramma%202024_EN_def.pdf).

Netherlands Organisation for Scientific Research. 2023. "Rubicon Call for proposals." (https://www.nwo.nl/sites/nwo/files/media-files/Call%20for%20Proposals%20Rubicon%202023-3%20EN.pdf).

Netherlands Organisation for Scientific Research. N.d. "Five questions about grants and award rates." (https://www.nwo.nl/en/five-questions-about-grants-and-award-rates).

Netherlands Organisation for Scientific Research. N.d. "Funding lines." (https://www.nwo.nl/en/funding-lines).

Netherlands Organisation for Scientific Research. N.d. "Governance and organisation." (https://www.nwo.nl/en/governance-and-organisation).

Netherlands Organisation for Scientific Research. N.d. "Open Competition Domain Science – XL." (https://www.nwo.nl/en/calls/open-competition-domain-science-xl).

Netherlands Organisation for Scientific Research. N.d. "NWO-Talent Programme Vici Science domain 2023." (https://www.nwo.nl/en/calls/nwo-talent-programme-vici-science-domain-2023).

Netherlands Organisation for Scientific Research. N.d. "What does the Dutch Research Council do?" (https://www.nwo.nl/en/what-does-the-dutch-research-council-do).

United Kingdom Research and Innovation (UKRI)

United Kingdom Research and Innovation. 2022. "Natural Environment Research Council: GUIDANCE FOR REVIEWERS OF DISCOVERY SCIENCE LARGE GRANTS FULL BIDS APPLICATIONS." (https://www.ukri.org/wp-content/uploads/2022/12/NERC-081222-Publication-GuidanceForReviewersOfDiscoveryScienceLargeGrants.pdf).

United Kingdom Research and Innovation. 2023. "BBSRC: Guidance for Reviewers." (https://www.ukri.org/councils/bbsrc/guidance-for-reviewers/).

United Kingdom Research and Innovation. 2023. "General Guidance: What happens after you have submitted your application." (https://www.ukri.org/councils/innovate-uk/guidance-for-applicants/general-guidance/what-happens-after-you-have-submitted-your-application/).

United Kingdom Research and Innovation. 2023. "How we make decisions." (https://www.ukri.org/apply-for-funding/how-we-make-decisions/).

United Kingdom Research and Innovation. 2023. "Our organization." (https://www.ukri.org/who-we-are/about-uk-research-and-innovation/our-organisation/).

United Kingdom Research and Innovation. 2023. "Role of panel meetings in peer review." (https://www.ukri.org/councils/epsrc/guidance-for-reviewers/peer-review-panels/role-of-panel-meetings-in-peer-review/).

United Kingdom Research and Innovation. N.d. "About UK Research and Innovation." (https://www.ukri.org/who-we-are/about-uk-research-and-innovation/).

United Kingdom Research and Innovation. N.d. "BBSRC Guidance Notes for Reviewers Using the Je-S System." (https://www.ukri.org/wp-content/uploads/2021/03/BBSRC-250723-GuidanceNotesReviewersUsingJointElectronicSubmissionSystem.pdf).

United Kingdom Research and Innovation. N.d. "Biotechnology and Biological Sciences Research Council (BBSRC)." (https://www.ukri.org/councils/bbsrc/).

United Kingdom Research and Innovation. N.d. "Economic and Social Research Council (ESRC)." (https://www.ukri.org/councils/esrc/).

United Kingdom Research and Innovation. N.d. "Engineering and Physical Sciences Research Council (EPSRC)." (https://www.ukri.org/councils/epsrc/).

United Kingdom Research and Innovation. N.d. "EPSRC: Panel Member Guidance." (https://www.ukri.org/wp-content/uploads/2021/10/EPSRC-150823-PanelMemberGuidance.pdf).

United Kingdom Research and Innovation. N.d. "ESRC Peer Reviewer Academic Assessment guidance." (https://www.ukri.org/wp-content/uploads/2022/02/ESRC009022022-PeerReviewAcademicAssessmentGuidance.pdf).

United Kingdom Research and Innovation. N.d. "Natural Environment Research Council (NERC)." (https://www.ukri.org/councils/nerc/).

German Research Foundation (DFG)

German Research Foundation. 2015. "Information on the Proposal, Review and Decision-Making Process." (https://www.dfg.de/en/research_funding/programmes/coordinated_programmes/research_training_groups/proposal_process/index.html).

German Research Foundation. 2021. "Funding at a Glance." (https://www.dfg.de/en/research_funding/programmes/index.html).

German Research Foundation. 2021. "Guidelines for Reviews in the Heisenberg Programme." (https://www.dfg.de/formulare/10_222/10_222_en.pdf).

German Research Foundation. 2021. "Guidelines for the Review of Research Fellowships." (https://www.dfg.de/formulare/10_204/10_204_en.pdf).

German Research Foundation. 2022. "Guidelines for Safeguarding Good Research Practice. Code of Conduct." (https://zenodo.org/records/6472827).

German Research Foundation. 2022. "Guidelines for the Review of Research Grants." (https://www.dfg.de/formulare/10_206/10_206_en.pdf).

German Research Foundation. 2022. "Guidelines Heisenberg Programme" (https://www.dfg.de/formulare/50_03/50_03_en.pdf).

German Research Foundation. 2022. "Individual Research Grants." (https://www.dfg.de/en/research_funding/programmes/individual/research_grants/index.html).

German Research Foundation. 2022. "Walter Benjamin Programme." (https://www.dfg.de/en/research_funding/programmes/individual/walter_benjamin/index.html).

German Research Foundation. 2023. "Arriving at a decision." (https://www.dfg.de/en/research_funding/proposal_funding_process/individual_grants_programmes/arriving_decision/index.html).

German Research Foundation. 2023. "Emmy Noether Programme." (https://www.dfg.de/en/research_funding/programmes/individual/emmy_noether/index.html)

German Research Foundation. 2023. "General Guidelines for Reviews." (https://www.dfg.de/formulare/10_20/10_20_en.pdf).

German Research Foundation. 2023. "General Questions about Proposals and Proposal Submission." (https://www.dfg.de/en/research_funding/faq/faq_submitting_proposal/index.html).

German Research Foundation. 2023. "Guidelines for Reviews in the Emmy Noether Programme." (https://www.dfg.de/formulare/10_210/10_210_en.pdf).

German Research Foundation. 2023. "Guidelines for Reviews in the Walter Benjamin Programme." (https://www.dfg.de/formulare/10_219/10_219_en.pdf).

German Research Foundation. 2023. "Guidelines for the Review of Reinhart Koselleck Projects." (https://www.dfg.de/formulare/10_203/10_203_en.pdf).

German Research Foundation. 2023. "Heisenberg Programme." (https://www.dfg.de/en/research_funding/programmes/individual/heisenberg/index.html).

German Research Foundation. 2023. "Proposal Information." (https://www.dfg.de/en/research_funding/proposal_funding_process/individual_grants_programmes/proposal_information/index.html).

German Research Foundation. 2023. "Questions about the decision-making process." (https://www.dfg.de/en/research_funding/faq/faq_committee_members/index.html).

German Research Foundation. 2023. "Reinhart Koselleck Projects." (https://www.dfg.de/en/research_funding/programmes/individual/reinhart_koselleck_projects/index.html).

German Research Foundation. 2023. "Research Fellowships." (https://www.dfg.de/en/research_funding/programmes/individual/research_fellowships/index.html).

German Research Foundation. 2023. "What is the DFG?" (https://www.dfg.de/en/dfg_profile/about_the_dfg/what_is_the_dfg/index.html).

Alfred P. Sloan Foundation

Alfred P. Sloan Foundation. 2023. "Grant Application Guidelines." (https://sloan.org/storage/app/media/files/application_documents/Sloan-Grant-Proposal-Guidelines-Research-Projects.pdf).

Alfred P. Sloan Foundation. N.d. "Energy & Environment." (https://sloan.org/programs/research/energy-and-environment).

Alfred P. Sloan Foundation. N.d. "Letters of Inquiry." (https://sloan.org/grants/apply#tab-letters-of-inquiry).

Alfred P. Sloan Foundation. N.d. "Matter-to-Life." (https://sloan.org/programs/research/matter-to-life).

Alfred P. Sloan Foundation. N.d. "The Grant Application Process." (https://sloan.org/grants/apply#tab-the-grant-application-process).

Alfred P. Sloan Foundation. N.d. "Small-Scale Fundamental Physics." (https://sloan.org/programs/research/small-scale-fundamental-physics).

Bill and Melinda Gates Foundation

Bill and Melinda Gates Foundation. 2010. "Our Approach to Shaping, Funding, and Managing Grants." (https://docs.gatesfoundation.org/documents/our-approach-to-grants.pdf).

Bill and Melinda Gates Foundation. N.d. "How we work." (https://www.gatesfoundation.org/about/how-we-work).

Bill and Melinda Gates Foundation. N.d. "Our work." (https://www.gatesfoundation.org/our-work).

Bill and Melinda Gates Foundation. N.d. "Accelerating Catalyzing Solutions for Climate Change's Impact on Health, Agriculture, and Gender Rules and Guidelines." (https://gcgh.grandchallenges.org/sites/default/files/files/climate_change_impact_rules_and_guidelines.pdf).

Bill and Melinda Gates Foundation. N.d. "Strengthening African National Regulatory Authorities Data Systems to Enhance and Track Performance." (https://gcgh.grandchallenges.org/challenge/strengthening-african-national-regulatory-authorities-data-systems-enhance-and-track).

Bill and Melinda Gates Foundation. N.d. "Strengthening African National Regulatory Authorities Data Systems to Enhance and Track Performance Grand Challenges Rules & Guidelines." (https://gcgh.grandchallenges.org/sites/default/files/files/african_national_regulatory_authorities_data_systems_rules_and_guidelines.pdf).

MacArthur Foundation

MacArthur Foundation. 2015. "MacArthur Fellows Program Frequently Asked Questions." (https://www.macfound.org/media/files/macarthur_fellows_program_faq_v4.pdf).

MacArthur Foundation. 2023. "An Opportunity for Transparency: Illustrating Our Grant Process." (https://www.macfound.org/press/perspectives/an-opportunity-for-transparency-illustrating-our-grant-process).

MacArthur Foundation. N.d. "About MacArthur Fellows Program." (https://www.macfound.org/programs/fellows/strategy).

MacArthur Foundation. N.d. "Big Bets Guiding Questions." (https://www.macfound.org/programs/bigbets/guiding-questions).

MacArthur Foundation. N.d. "Climate Solutions Grant Guidelines." (https://www.macfound.org/info-grantseekers/grantmaking-guidelines/climate-solutions).

MacArthur Foundation. N.d. "Climate Solutions Our Goal." (https://www.macfound.org/programs/climate/strategy).

MacArthur Foundation. N.d. "How We Work." (https://www.macfound.org/about/how-we-work/).

MacArthur Foundation. N.d. "MacArthur Fellows FAQs." (https://www.macfound.org/programs/awards/fellows/faq).

MacArthur Foundation. N.d. "MacArthur Foundation Grant Application Process." (https://www.macfound.org/media/article_pdfs/grant-illustration-process.pdf).

MacArthur Foundation. N.d. "Nuclear Challenges Grant Guidelines." (https://www.macfound.org/info-grantseekers/grantmaking-guidelines/nuclear-grant-guide).

MacArthur Foundation. N.d. "Nuclear Challenges Our Goal." (https://www.macfound.org/programs/nuclear/strategy).

Gordon and Betty Moore Foundation

Gordon and Betty Moore Foundation. N.d. "Environmental Conservation." (https://www.moore.org/programs/environmental-conservation).

Gordon and Betty Moore Foundation. N.d. "Grant Development Overview." (https://www.moore.org/docs/default-source/Grantee-Resources/grant-development-review.pdf?sfvrsn=2).

Gordon and Betty Moore Foundation. N.d. "Our approach." (https://www.moore.org/about/our-approach).

Gordon and Betty Moore Foundation. N.d. "Our Grantmaking." (https://www.moore.org/about/our-grantmaking).

Gordon and Betty Moore Foundation. N.d. "Science." (https://www.moore.org/programs/science).

Appendix B:
Quality Assurance

- To ensure the overall quality of the project, we conducted the following quality assurance procedures: We consulted with a librarian to develop and calibrate our search strategy. The librarian conducted the database searches and returned the results. We replicated the searches to confirm the number of results.
- A primary reviewer screened records identified through the database and Google searches, and a senior reviewer verified all screening decisions.
- Each publication that met the screening criteria was reviewed first by a primary reviewer and a second time by a senior reviewer. These reviews served to verify the study meets the inclusion criteria, is classified as the correct study type, and that the information entered in the review template is accurate and complete.
- One team member completed all tabulations, and a task leader verified them.
- An independent reviewer and the deputy project director reviewed the written report, focusing on relevance, method appropriateness, accurate interpretation, objective conclusions, transparency, writing clarity, and presentation.
- Our editors edited the written report for clarity, succinctness, and consistency.
- Our production staff made this report visually appealing and 508 compliant.

Appendix C:
Acronym Keys

Exhibit C.1. Acronyms used in this report.

Acronym	Definition
AD	average deviation
AI	artificial intelligence
AIBS	American Institute of Biological Sciences
ARC	Australian Research Council
BBSRC	Biotechnology and Biological Sciences Research Council
CHIPS	Creating Helpful Incentives to Produce Semiconductors (for America Fund Act)
COST	European Cooperation in Science and Technology
CV	curriculum vitae
DFG	German Research Foundation
DOD	United States Department of Defense
DOE	United States Department of Energy
DORA	Declaration on Research Assessment
ED	United States Department of Education
EPSRC	Engineering and Physical Sciences Research Council (United Kingdom Research and Innovation)
ERC	European Research Council
ESRC	Economic and Social Research Council (United Kingdom Research and Innovation)
EU	European Union
FET-Open	Future and Emerging Technologies Program
FWF	Austrian Science Fund
FY	fiscal year
HRC	Health Research Council of New Zealand
ICC	intraclass correlation
IHRM	Interactive Heuristic Review Mechanism
JIF	journal impact factor
KIC	(Netherlands Organisation for Scientific Research) Knowledge and Innovation Covenant
MSCA	Marie Skłodowska-Curie Actions
NASA	National Aeronautics and Space Administration
NEA	National Endowment for the Arts
NEH	National Endowment for the Humanities
NERC	Natural Environment Research Council (United Kingdom Research and Innovation)
NGF	(Netherlands Organisation for Scientific Research) National Growth Fund
NIH	National Institutes of Health
NSB	National Science Board
NSERC	Natural Sciences and Engineering Research Council of Canada
NSF	U.S. National Science Foundation
NWA	(Netherlands Organisation for Scientific Research) Dutch Research Agenda
NWO	Netherlands Organization for Scientific Research
PI	principal investigator
REA	Research Executive Agency

Acronym	Definition
RNC	Research Council of Norway
RRI	Responsible Research and Innovation (European Union Framework Programs)
SIAMPI	Social Impact Assessment Methods for research and funding instruments through the study of Productive Interactions
STEM	science, technology, engineering, and mathematics
TF-IDF	term frequency-inverse document frequency
TWG	technical working group
UKRI	United Kingdom Research and Innovation
USDA	United States Department of Agriculture
VA	United States Department of Veterans Affairs

References

Arnott, James, Christine Kirchhoff, Ryan Meyer, Alison M. Meadow, and Angela T. Bednarek. 2020. "Sponsoring Actionable Science: What Public Science Funders Can Do to Advance Sustainability and the Social Contract for Science." *Current Opinion in Environmental Sustainability* 42:38–44. doi: 10.1016/j.cosust.2020.01.006.

Avin, Shahar. 2015. "Funding Science by Lottery." Pp. 111–26 in *European Studies in Philosophy of Science*, edited by Dennis Dieks, Maria Carla Galavotti, and Wenceslao J. Gonzalez. Springer.

Ayoubi, Charles, Michele Pezzoni, and Fabiana Visentin. 2021. "Does It Pay to Do Novel Science? The Selectivity Patterns in Science Funding." *Science and Public Policy* 48(5):635–48. doi: 10.1093/scipol/scab031.

Baimpos, Theodoros, Nils Dittel, and Roumen Borissov. 2020. "Unravelling the Panel Contribution Upon Peer Review Evaluation of Numerous, Unstructured and Highly Interdisciplinary Research Proposals." *Research Evaluation* 29(3):316–26. doi: 10.1093/reseval/rvz013.

Bedessem, Baptiste. 2020. "Should We Fund Research Randomly? An Epistemological Criticism of the Lottery Model as an Alternative to Peer Review for the Funding of Science." *Research Evaluation* 29(2):150–57. doi: 10.1093/reseval/rvz034.

Benneworth, Paul Stephen, and Julia Olmos-Peñuela. 2022. "An Openness Framework for Ex Ante Evaluation of Societal Impact of Research." *Research Evaluation*. doi: 10.1093/reseval/rvac023.

Biegelbauer, Peter, Thomas Palfinger, and Sabine Mayer. 2020. "How to Select the Best: Selection Procedures of Innovation Agencies." *Research Evaluation* 29(3):289–99. doi: 10.1093/reseval/rvaa011.

Bol, Thijs, Mathijs de Vaan, and Arnout van de Rijt. 2018. "The Matthew Effect in Science Funding." *Proceedings of the National Academy of Sciences of the United States of America* 115(19):4887–90. doi: 10.1073/pnas.1719557115.

Bollen, Johan. 2018. "Who Would You Share Your Funding With?" *Nature* 560(7717):143. doi: 10.1038/d41586-018-05887-3.

Bollen, Johan, Stephen R. Carpenter, Jane Lubchenco, and Marten Scheffer. 2019. "Rethinking Resource Allocation in Science." *Ecology and Society* 24(3):29. doi: 10.5751/ES-11005-240329.

Bollen, Johan, David J. Crandall, Damion Junk, Ying Ding, and Katy Börner. 2014. "From Funding Agencies to Scientific Agency." *EMBO Reports* 15(2):131–33. doi: 10.1002/embr.201338068.

Bornmann, Lutz. 2013. "What is Societal Impact of Research and How Can It be Assessed? A Literature Survey." *Journal of the American Society for Information Science and Technology* 64(2):217-233. doi: 10.1002/asi.22803

Bornmann, Lutz, Ruediger Mutz, and Hans-Dieter Daniel. 2007. "Gender Differences in Grant Peer Review: A Meta-Analysis." *Journal of Informetrics* 1(3):226–38. doi: 10.1016/j.joi.2007.03.001.

Brouns, M. L. M. 2000. "The Gendered Nature of Assessment Procedures in Scientific Research Funding: The Dutch Case." *Higher Education in Europe* 25(2):193–99. doi: 10.1080/713669261.

Brunet, Lucas, and Ruth Müller. 2022. "Making the Cut: How Panel Reviewers Use Evaluation Devices to Select Applications at the European Research Council." *Research Evaluation* 31(4):486–97. doi: 10.1093/reseval/rvac040.

Bulathsinhala, Nadika A. 2015. "Ex-ante Evaluation of Publicly Funded R&D Projects: Searching for exploration." *Science and Public Policy* 42(2): 162-175. doi: 10.1093/scipol/scu035.

Cañibano, Carolina, F. Javier Otamendi, and I. Andújar. 2009. "An Assessment of Selection Processes among Candidates for Public Research Grants: The Case of the Ramón y Cajal Programme in Spain." *Research Evaluation* 18(2):153–61. doi: 10.3152/095820209x444968.

Chen, Christine Yifeng, Sara S. Kahanamoku, Aradhna Tripati, Rosanna A. Alegado, Vernon R. Morris, Karen Andrade, and Justin Hosbey. 2022. "Systemic Racial Disparities in Funding Rates at the National Science Foundation." *eLife* 11:e83071. doi: 10.7554/eLife.83071.

Cousens, Roger. 2019. "Why Can't We Make Research Grant Allocation Systems More Consistent? A Personal Opinion." *Ecology and Evolution* 9(4):1536–44. doi: 10.1002/ece3.4855.

Creating Helpful Incentives to Produce Semiconductors (CHIPS) for America Fund Act of 2022, Public Law 117–167, 136 Statutes at Large 1366. (2022).

Cunha, Jorge, Paula Varandas Ferreira, Maria Madalena Teixeira De Araújo, and Enrique Ares. 2012. "Social Return of R&D Investments in Manufacturing Sector: Some Insights from an Exploratory Case Study." *AIP Conference Proceedings* 1431:43–53. doi: 10.1063/1.4707549.

Davis, Michael, and Kelly Laas. 2013. "'Broader Impacts' or 'Responsible Research and Innovation'? A Comparison of Two Criteria for Funding Research in Science and Engineering." *Science and Engineering Ethics* 20(4):963–83. doi: 10.1007/s11948-013-9480-1.

De Los Reyes, Andres, and Mo Wang. 2012. "Applying Psychometric Theory and Research to Developing a Continuously Distributed Approach to Making Research Funding Decisions." *Review of General Psychology* 16(3): 298-304. doi: 10.1037/a0027250.

Devyatkin, Dmitry, Roman Suvorov, Ilya Tikhomirov, and Oleg Grigoriev. 2016. "Feature Selection Techniques for Scientific Projects Funding Criteria Analysis." In 2016 IEEE 8th International Conference on Intelligent Systems (IS), pp. 167-172.

Dinov, Ivo D. 2019. "Flipping the Grant Application Review Process." *Studies in Higher Education* 45(8):1737–45. doi: 10.1080/03075079.2019.1628201.

Dresler, Martin. 2022. "FENS-Kavli Network of Excellence: Postponed, Non-competitive Peer Review for Research Funding." *European Journal of Neuroscience*. doi: 10.1111/ejn.15818.

Dwitayanti, Yevi, and M. Miftakul Amin. 2023. "The Group Decision Support System Model of Research Proposal Assessment Using Researcher Track Record and Research Output." In *AIP Conference Proceedings* 2683(1). AIP Publishing.

Elhorst, J. Paul, and Dries Faems. 2021. "Evaluating Proposals in Innovation Contests: Exploring Negative Scoring Spillovers in the Absence of a Strict Evaluation Sequence." *Research Policy* 50(4):104198. doi: 10.1016/j.respol.2021.104198.

External Funds Service. 2023. "Future and Emerging Technologies [FET]." Retrieved December 6, 2023. (https://sfe.lnl.infn.it/future-and-emerging-technologies-fet/#:~:text=The%20FET%20programme%20has%20three%20complementary%20lines%20of,of%20knowledge%20and%20excellence%20around%20them.%20More%20items).

Falk-Krzesinski, Holly J., and Stacey C. Tobin. 2015. "How Do I Review Thee? Let Me Count the Ways: A Comparison of Research Grant Proposal Review Criteria Across US Federal Funding Agencies." *The Journal of Research Administration* 46(2):79–94.

Fang, Ferric C., and Arturo Casadevall. 2016a. "Grant Funding: Playing the Odds." *Science* 352(6282):158. doi: 10.1126/science.352.6282.158-a.

Fang, Ferric C., and Arturo Casadevall. 2016b. "Research Funding: The Case for a Modified Lottery." *mBio* 7(2). doi: 10.1128/mbio.00422-16.

Feliciani, Thomas, Michael Morreau, Junwen Luo, Pablo Lucas, and K. Shankar. 2022. "Designing Grant-Review Panels for Better Funding Decisions: Lessons from an Empirically Calibrated Simulation Model." *Research Policy* 51(4):104467. doi: 10.1016/j.respol.2021.104467.

Franzoni, Chiara, and Paula Stephan. 2022. "Uncertainty and Risk-Taking in Science: Meaning, Measurement and Management in Peer Review of Research Proposals." *Research Policy* 52(3):104706. doi: 10.1016/j.respol.2022.104706

Gallo, Stephen A., Joanne Sullivan, and DaJoie R. Croslan. 2022. "Scientists from Minority-Serving Institutions and Their Participation in Grant Peer Review." *BioScience* 72(3):289–99. doi: 10.1093/biosci/biab130.

Gallo, Stephen A., Lisa Thompson, Karen B. Schmaling, and Scott R. Glisson. 2018. "Risk Evaluation in Peer Review of Grant Applications." *Environment Systems and Decisions* 38(2):216–29. doi: 10.1007/s10669-018-9677-6.

Gallo, Stephen A., Lisa Thompson, Karen B. Schmaling, and Scott R. Glisson. 2019. "The Participation and Motivations of Grant Peer Reviewers: A Comprehensive Survey." *Science and Engineering Ethics* 26(2):761–82. doi: 10.1007/s11948-019-00123-1.

Gallo, Stephen A., Michael P. LeMaster, and Scott R. Glisson. 2016. "Frequency and Type of Conflicts of Interest in the Peer Review of Basic Biomedical Research Funding Applications: Self-Reporting Versus Manual Detection." *Science and Engineering Ethics* 22(1):189–97. doi: 10.1007/s11948-015-9631-7.

Gross, Kevin, and Carl T. Bergstrom. 2019. "Contest Models Highlight Inherent Inefficiencies of Scientific Funding Competitions." *PLoS Biology* 17(1):e3000065. doi: 10.1371/journal.pbio.3000065.

Gunn, Andrew, and Michael Mintrom. 2017. "Evaluating the Non-Academic Impact of Academic Research: Design Considerations." *Journal of Higher Education Policy and Management* 39(1):20–30. doi: 10.1080/1360080x.2016.1254429.

Gurwitz, David, Elena Milanesi, and Thomas Koenig. 2014. "Grant Application Review: The Case of Transparency." *PLoS Biology* 12(2): e1002010.

Győrffy, Balázs, Boglorka Weltz, and István Szabó. 2023. "Supporting Grant Reviewers Through the Scientometric Ranking of Applicants." *PLOS ONE* 18(1):e0280480. Doi: 10.1371/journal.pone.0280480.

Győrffy, Balázs, Peter C. Herman, and István Szabó. 2020. "Research Funding: Past Performance Is a Stronger Predictor of Future Scientific Output than Reviewer Scores." *Journal of Informetrics* 14(3):101050. doi: 10.1016/j.joi.2020.101050.

Harnagel, Audrey. 2019. "A Mid-Level Approach to Modeling Scientific Communities." *Studies in History and Philosophy of Science* 76:49–59. doi: 10.1016/j.shpsa.2018.12.010.

Heinze, Thomas. 2008. "How to Sponsor Ground-Breaking Research: A Comparison of Funding Schemes." *Science and Public Policy* 35(5):302–18. doi: 10.3152/030234208x317151.

Hesselberg, Jan-Ole, Knut Inge Fostervold, Pål Ulleberg, and Ida Svege. 2021. "Individual versus General Structured Feedback to Improve Agreement in Grant Peer Review: A Randomized Controlled Trial." *Research Integrity and Peer Review* 6(1). doi: 10.1186/s41073-021-00115-5.

Heyard, Rachel, Manuela Ott, Georgia Salanti, and Matthias Egger. 2022. "Rethinking the Funding Line at the Swiss National Science Foundation: Bayesian Ranking and Lottery." *Statistics and Public Policy* 9(1):110–21. doi: 10.1080/2330443x.2022.2086190.

Hirt, Julian, Thomas Nordhausen, Christian Appenzeller-Herzog, and Hannah Ewald. 2022. "Citation Tracking for Systematic Literature Searching: A Scoping Review." medRxiv (Cold Spring Harbor Laboratory). doi: 10.1101/2022.09.29.22280494.

Holbrook, J. Britt, and Robert Frodeman. 2011. "Peer Review and the Ex Ante Assessment of Societal Impacts." *Research Evaluation* 20(3):239–46. doi: 10.3152/095820211x12941371876788.

Holbrook, J. Britt, and Steven Hrotic. 2013. "Blue Skies, Impacts, and Peer Review." *RT. A Journal on Research Policy and Evaluation* 1(1). doi: 10.13130/2282-5398/2914.

Holbrook, J. Britt. 2010. "The Use of Societal Impacts Considerations in Grant Proposal Peer Review: A Comparison of Five Models." *Technology & Innovation* 12(3):213–24. doi: 10.3727/194982410X12895770314078.

Horbach, Serge P. J. M., Joeri K. Tijdink, and Lex M. Bouter. 2022. "Partial Lottery Can Make Grant Allocation More Fair, More Efficient, and More Diverse." *Science and Public Policy* 49(4):580–82. doi: 10.1093/scipol/scac009.

Horsley, Tanya, Orvie Dingwall, and Margaret Sampson. 2011. "Checking Reference Lists to Find Additional Studies for Systematic Reviews." *The Cochrane Library* 2011(8). doi: 10.1002/14651858.mr000026.pub2.

Hug, Sven E., and Mirjam Aeschbach. 2020. "Criteria for Assessing Grant Applications: A Systematic Review." *Palgrave Communications* 6(1). doi: 10.1057/s41599-020-0412-9.

Jerrim, John, and Robert Vries. 2020. "Are Peer Reviews of Grant Proposals Reliable? An Analysis of Economic and Social Research Council (ESRC) Funding Applications." *Social Science Journal* 60(1):91–109. doi: 10.1080/03623319.2020.1728506.

Kerr, Norbert L., and R. Scott Tindale. 2004. "Group Performance and Decision Making." *Annual Review of Psychology* 55(1):623–55. doi: 10.1146/annurev.psych.55.090902.142009.

Kuhlisch, Wiltrud, Magnus Roos, Jörg Rothe, Joachim Rudolph, Björn Scheuermann, and Dietrich Stoyan. 2015. "A Statistical Approach to Calibrating the Scores of Biased Reviewers of Scientific Papers." *Metrika* 79(1):37–57. doi: 10.1007/s00184-015-0542-z.

Kuhn, Thomas. 1962. *The Structure of Scientific Revolutions.* University of Chicago Press: Chicago.

Kuppler, Matthias. 2022. "Predicting the Future Impact of Computer Science Researchers: Is There a Gender Bias?" *Scientometrics* 127(11):6695–732. doi: 10.1007/s11192-022-04337-2.

Langfeldt, Liv, Ingvild Reymert, and Dag W. Aksnes. 2021. "The Role of Metrics in Peer Assessments." *Research Evaluation* 30(1):112–26. doi: 10.1093/reseval/rvaa032.

Langfeldt, Liv. 2001. "The Decision-Making Constraints and Processes of Grant Peer Review, and Their Effects on the Review Outcome." *Social Studies of Science* 31(6):820–41. doi: 10.1177/030631201031006002.

Lee, Carole J. 2015. "Commensuration Bias in Peer Review." *Philosophy of Science* 82(5):1272–83. doi: 10.1086/683652.

Lee, Carole J., Cassidy R. Sugimoto, Guo Zhang, and Blaise Cronin. 2012. "Bias in Peer Review." *Journal of the Association for Information Science and Technology* 64(1):2–17. doi: 10.1002/asi.22784.

Liljequist, David, Britt Elfving, and Kirsti Skavberg Roaldsen. 2019. "Intraclass Correlation–A Discussion and Demonstration of Basic Features." *PLOS ONE* 14(7):e0219854.

Luo, Junwen, Lai Ma, and K. Shankar. 2021. "Does the Inclusion of Non-Academic Reviewers Make Any Difference for Grant Impact Panels?" *Science and Public Policy* 48(6):763–75. doi: 10.1093/scipol/scab046.

Ma, Long, Junwen Luo, Thomas Feliciani, and K. Shankar. 2020. "How to Evaluate Ex Ante Impact of Funding Proposals? An Analysis of Reviewers' Comments on Impact Statements." *Research Evaluation* 29(4):431–40. doi: 10.1093/reseval/rvaa022.

Marsh, Herbert W., Lutz Bornmann, Rüdiger Mutz, Hans-Dieter Daniel, and Alison J. O'Mara. 2009. "Gender Effects in the Peer Reviews of Grant Proposals: A Comprehensive Meta-Analysis Comparing Traditional and Multilevel Approaches." *Review of Educational Research* 79(3):1290–1326. doi: 10.3102/0034654309334143.

Marsh, Herbert W., Upali W. Jayasinghe, and Nigel W. Bond. 2008. "Improving the Peer-Review Process for Grant Applications: Reliability, Validity, Bias, and Generalizability." *American Psychologist* 63(3):160–68. doi: 10.1037/0003-066x.63.3.160.

Materia, Valentina Cristiana, Stefano Pascucci, and Christos Kolympiris. 2015. "Understanding the Selection Processes of Public Research Projects in Agriculture: The Role of Scientific Merit." *Food Policy* 56:87–99. doi: 10.1016/j.foodpol.2015.08.003.

Mohan, Premila, and Ramasamy Brakaspathy. 2018. "SERB Merit Review Process: Adapting to Emerging Challenges." *Current Science* 114(9):1835–39.

Mongeon, Philippe, Christine Brodeur, Catherine Beaudry, and Vincent Larivière. 2016. "Concentration of Research Funding Leads to Decreasing Marginal Returns." *Research Evaluation* 25(4):396–404. doi: 10.1093/reseval/rvw007.

Mow, Karen E. 2011. "Peers Inside the Black Box: Deciding Excellence." *The International Journal of Interdisciplinary Social Sciences* 5(10):175–84. doi: 10.18848/1833-1882/cgp/v05i10/51914.

Murray, Dennis L., Douglas W. Morris, Claude Lavoie, Peter R. Leavitt, Hugh J. MacIsaac, Michael E. J. Masson, and Marc-André Villard. 2016. "Bias in Research Grant Evaluation Has Dire Consequences for Small Universities." *PLOS ONE* 11(6):e0155876. doi: 10.1371/journal.pone.0155876.

Mutz, Rüdiger, Lutz Bornmann, and Hans-Dieter Daniel. 2012. "Heterogeneity of Inter-Rater Reliabilities of Grant Peer Reviews and Its Determinants: A General Estimating Equations Approach." *PLOS ONE* 7(10):e48509. doi: 10.1371/journal.pone.0048509.

Mutz, Rüdiger, Lutz Bornmann, and Hans-Dieter Daniel. 2016. "Funding Decision-Making Systems: An Empirical Comparison of Continuous and Dichotomous Approaches Based on Psychometric Theory." *Research Evaluation* rvw002. doi: 10.1093/reseval/rvw002.

The National Science Act of 1950, Public Law 81-507. (1950).

Neufeld, Jörg, Nathalie Huber, and Antje Wegner. 2013. "Peer Review-Based Selection Decisions in Individual Research Funding, Applicants' Publication Strategies and Performance: The Case of the ERC Starting Grants." *Research Evaluation* 22(4):237–47. doi: 10.1093/reseval/rvt014.

Olbrecht, Meike, and Lutz Bornmann. 2010. "Panel Peer Review of Grant Applications: What Do We Know from Research in Social Psychology on Judgment and Decision-Making in Groups?" *Research Evaluation* 19(4):293–304. doi: 10.3152/095820210x12809191250762.

Öztayşi, Başar, Sezi Çevik Onar, Kerim Göztepe, and Cengiz Kahraman. 2017. "Evaluation of Research Proposals for Grant Funding Using Interval-Valued Intuitionistic Fuzzy Sets." *Soft Computing* 21(5):1203–18. doi: 10.1007/s00500-015-1853-8.

Page, Matthew J., Joanne E. McKenzie, Patrick M. Bossuyt, Isabelle Boutron, Tammy Hoffmann, Cynthia D. Mulrow, Larissa Shamseer, Jennifer Tetzlaff, Elie A. Akl, Sue Brennan, Roger Chou, Julie Glanville, Jeremy Grimshaw, Asbjørn Hróbjartsson, Manoj M. Lalu, Tianjing Li, Elizabeth Loder, Evan Mayo-Wilson, Steve McDonald, Luke A. McGuinness, Lesley A. Stewart, James Thomas, Andrea C. Tricco, Vivian Welch, Penny Whiting, and David Moher. 2021. "The PRISMA 2020 Statement: An Updated Guideline for Reporting Systematic Reviews." *The BMJ* n71. doi: 10.1136/bmj.n71.

Parreiras, Roberta, Illya Kokshenev, Maria Audália Marques De Carvalho, A. C. M. Willer, C. F. Dellezzopolles, D. B. Nacif, and José Aldo Santana. 2019. "A Flexible Multicriteria Decision-Making Methodology to Support the Strategic Management of Science, Technology and Innovation Research Funding Programs." *European Journal of Operational Research* 272(2):725–39. doi: 10.1016/j.ejor.2018.06.050.

Pedersen, David Budtz, Jonas Følsgaard Grønvad, and Rolf Hvidtfeldt. 2020. "Methods for Mapping the Impact of Social Sciences and Humanities—A Literature Review." *Research Evaluation* 29(1):4–21. doi: 10.1093/reseval/rvz033.

Peterson, Helen, and Liisa Husu. 2022. "Online Panel Work Through a Gender Lens: Implications of Digital Peer Review Meetings." *Science and Public Policy* 50(3):371–81. doi: 10.1093/scipol/scac075.

Philipps, Axel. 2021. "Science Rules! A Qualitative Study of Scientists' Approaches to Grant Lottery." *Research Evaluation* 30(1):102–11. doi: 10.1093/reseval/rvaa027.

Pina, David G., Ivan Buljan, Darko Hren, and Ana Marušić. 2021. "A Retrospective Analysis of the Peer Review of More than 75,000 Marie Curie Proposals Between 2007 and 2018." *eLife* 10:e59338. doi: 10.7554/elife.59338.

Piro, Fredrik Niclas, Pål Børing, Lisa Scordato, and Dag W. Aksnes. 2020. "University Characteristics and Probabilities for Funding of Proposals in the European Framework Programs." *Science and Public Policy* 47(4):581–93. doi: 10.1093/scipol/scaa037.

Porter, Robert. 2005. "What Do Grant Reviewers Really Want, Anyway?" *The Journal of Research Administration* 36(2):47.

Portney, Leslie G., and Mary P. Watkins. 2000. *Foundations of Clinical Research: Applications to Practice*. New Jersey: Prentice Hall.

Ramos-Vielba, Irene, Duncan A. Thomas, and Kaare Aagaard. 2022. "Societal Targeting in Researcher Funding: An Exploratory Approach." *Research Evaluation* 31(2):202–13. doi: 10.1093/reseval/rvab044.

Ranga, Marina, Namrata Gupta, and Henry Etzkowitz. 2012. *Gender Effects in Research Funding*. Bonn: Deutsche Forschungsgemeinschaft. Retrieved January 18, 2024 (https://www.dfg.de/resource/blob/170570/e48fab44b49274b83e2b5aeb382145d0/studie-gender-effects-data.pdf)

Razmgir, Maryam, Sirous Panahi, Leila Ghalichi, Seyed Ali Javad Mousavi, and Shahram Sedghi. 2021. "Exploring Research Impact Models: A Systematic Scoping Review." *Research Evaluation* 30(4): 443–57. doi: 10.1093/reseval/rvab009.

Reale, Emanuela, and Antonio Zinilli. 2017. "Evaluation for the Allocation of University Research Project Funding: Can Rules Improve the Peer Review?" *Research Evaluation* 26(3):190–98. doi: 10.1093/reseval/rvx019.

Recio-Saucedo, Alejandra, Ksenia Crane, Katie Meadmore, Kathryn Fackrell, Hazel Church, Simon Fraser, and Amanda Blatch-Jones. 2022. "What Works for Peer Review and Decision-Making in Research Funding: A Realist Synthesis." *Research Integrity and Peer Review* 7(1). doi: 10.1186/s41073-022-00120-2.

Rip, Arie. 2000. "Higher Forms of Nonsense." *European Review* 8(4):467–85. doi: 10.1017/s1062798700005032

Roebber, Paul J., and David M. Schultz. 2011. "Peer Review, Program Officers and Science Funding." *PLOS ONE* 6(4):e18680. doi: 10.1371/journal.pone.0018680.

Roumbanis, Lambros. 2019. "Peer Review or Lottery? A Critical Analysis of Two Different Forms of Decision-Making Mechanisms for Allocation of Research Grants." *Science, Technology, & Human Values* 44(6):994–1019. doi: 10.1177/0162243918822744.

Roumbanis, Lambros. 2022. "Disagreement and Agonistic Chance in Peer Review." *Science, Technology, & Human Values* 47(6):1302–33. doi: 10.1177/01622439211026016.

Santana, Carlos. 2022. "Why Citizen Review Might Beat Peer Review at Identifying Pursuitworthy Scientific Research." *Studies in History and Philosophy of Science* 92:20–26. doi: 10.1016/j.shpsa.2022.01.012.

Santos, João M. 2022. "Quis Judicabit Ipsos Judices? A Case Study on the Dynamics of Competitive Funding Panel Evaluations." *Research Evaluation* 32(1):70–85. doi: 10.1093/reseval/rvac021.

Sato, Sayaka, Pascal Gygax, Julian F. Randall, and Marianne Schmid Mast. 2020. "The Leaky Pipeline in Research Grant Peer Review and Funding Decisions: Challenges and Future Directions." *Higher Education* 82(1):145–62. doi: 10.1007/s10734-020-00626-y.

Seeber, Marco, Jef Vlegels, and Mattia Cattaneo. 2022. "Conditions That Do or Do Not Disadvantage Interdisciplinary Research Proposals in Project Evaluation." *Journal of the Association for Information Science and Technology* 73(8):1106–26. doi: 10.1002/asi.24617.

Seeber, Marco, Jef Vlegels, Elwin Reimink, Ana Marušić, and David G. Pina. 2021. "Does Reviewing Experience Reduce Disagreement in Proposals Evaluation? Insights from Marie Skłodowska-Curie and COST Actions." *Research Evaluation* 30(3):349–60. doi: 10.1093/reseval/rvab011.

Shaw, Jamie. 2023. "Peer Review in Funding-by-Lottery: A Systematic Overview and Expansion." *Research Evaluation* 32(1):86–100. doi: 10.1093/reseval/rvac022.

Smaldino, Paul E., Matthew A. Turner, and Pablo Contreras Kallens. 2019. "Open Science and Modified Funding Lotteries Can Impede the Natural Selection of Bad Science." *Royal Society Open Science* 6(7):190194. doi: 10.1098/rsos.190194.

Steiner Davis, Miriam L. E., Tiffani Conner, Kate Miller-Bains, and Leslie Shapard. 2020. "What Makes an Effective Grants Peer Reviewer? An Exploratory Study of the Necessary Skills." *PLOS ONE* 15(5):e0232327. doi: 10.1371/journal.pone.0232327.

Tennant, J. P., and Ross-Hellauer, T. 2020. "The Limitations to Our Understanding of Peer Review." *Research Integrity and Peer Review* 5:6. doi: 10.1186/s41073-020-00092-1.

U.S. National Science Foundation. 2022. *FY 2022: Agency Financial Report.* Alexandria, VA: U.S. National Science Foundation.

U.S. National Science Foundation. 2023. *Proposal and Award Policies and Procedures Guide.* Retrieved October 4, 2023 (https://nsf-gov-resources.nsf.gov/2022-10/nsf23_1.pdf?VersionId=7yfheI.bNrekBK7F5cKu9riXFbi1YjRX).

Vaesen, Krist, and Joel Katzav. 2017. "How Much Would Each Researcher Receive If Competitive Government Research Funding Were Distributed Equally among Researchers?" PLOS ONE 12(9):e0183967. doi: 10.1371/journal.pone.0183967.

van Arensbergen, Pleun, Inge van der Weijden, and Peter van den Besselaar. 2013. "Academic Talent Selection in Grant Review Panels." Pp. 25-54 in *(Re)searching Scientific Careers*, edited by K. Prpic, K.; I. van der Weijden, and N. Aseulova. St. Petersburg: IHST/RAS & SSTNET/ES

van Arensbergen, Pleun, Inge van der Weijden, and Peter van den Besselaar. 2014a. "The Selection of Talent as a Group Process. A Literature Review on the Social Dynamics of Decision Making in Grant Panels." *Research Evaluation* 23(4):298–311. doi: 10.1093/reseval/rvu017.

van Arensbergen, Pleun, Inge van der Weijden, and Peter van den Besselaar. 2014b. "Different Views on Scholarly Talent: What Are the Talents We Are Looking for in Science?" *Research Evaluation* 23(4):273–84. doi: 10.1093/reseval/rvu015.

van den Besselaar, Peter, and Pleun van Arensbergen. 2013. "Talent Selection and the Funding of Research." *Higher Education Policy* 26(3):421–27. doi: 10.1057/hep.2013.16.

van den Besselaar, Peter, Ulf Sandström, and Hélène Schiffbaenker. 2018. "Studying Grant Decision-Making: A Linguistic Analysis of Review Reports." *Scientometrics* 117(1):313–29. doi: 10.1007/s11192-018-2848-x.

van der Lee, Romy, and Naomi Ellemers. 2015. "Gender Contributes to Personal Research Funding Success in The Netherlands." *Proceedings of the National Academy of Sciences of the United States of America* 112(40):12349–53. doi: 10.1073/pnas.1510159112.

van Leeuwen, Thed N., and Henk F. Moed. 2012. "Funding Decisions, Peer Review, and Scientific Excellence in Physical Sciences, Chemistry, and Geosciences." Research Evaluation 21(3):189–98. doi: 10.1093/reseval/rvs009.

Veletanlić, Emina, and Creso M. Sá. 2020. "Implementing the Innovation Agenda: A Study of Change at a Research Funding Agency." *Minerva* 58(2):261–83. doi: 10.1007/s11024-020-09396-4.

Wallenius, Jyrki, James S. Dyer, Peter C. Fishburn, Ralph E. Steuer, Stanley Zionts, and Kalyanmoy Deb. 2008. "Multiple Criteria Decision Making, Multiattribute Utility Theory: Recent Accomplishments and What Lies Ahead." *Management Science* 54(7):1336–49. doi: 10.1287/mnsc.1070.0838.

Wang, Y., Li, X., and Zheng, Y. 2011. "The Interactive Heuristic Review Mechanism: A New Method of Assessing Pioneering Research Projects of the National Natural Science Foundation of China." *Research Evaluation*, 20(4), 267-274.

Xu, Haiyun, Jos Winnink, Huawei Wu, Hongshen Pang, and Chao Wang. 2021. "Using the Catastrophe Theory to Discover Transformative Research Topics." *Research Evaluation* 31(1):61–79. doi: 10.1093/reseval/rvab027.

Zoller, Frank A., Eric Zimmerling, and Roman Boutellier. 2014. "Assessing the Impact of the Funding Environment on Researchers' Risk Aversion: The Use of Citation Statistics." *Higher Education* 68(3):333–45. doi: 10.1007/s10734-014-9714-4.

Acknowledgements, Disclosures, and Citation

Acknowledgements

We would like to thank Christopher Monk, Taylor Rhodes, Erika Rissi, and other NSF staff for their helpful guidance and feedback. Barry Bozeman, Cydney Dupree, J. Britt Holbrook, Kristen Intemann, Julie Risien, and Sean Watts reviewed an earlier version of this draft and provided helpful feedback. We also appreciate our Mathematica colleagues who contributed to the work, including Jennifer Brown, Molly Cameron, Jackie Drummond, Lindsay Fox, Nima Rahimi, and Margaret Sullivan.

Disclosures

This report was prepared for the U.S. National Science Foundation's (NSF) Evaluation and Assessment Capability (EAC) Section within the Office of Integrative Activities (OIA) under contract number 49100421D0011. The views expressed are those of the authors and should not be attributed to NSF, nor does mention of trade names, commercial products, or organizations imply endorsement of same by the U.S. Government.

Citation

Chandler, Jesse, Emily Rosen, Kimberley Raue, Katlyn Lee Milless and Danielle Rockman. 2024. *A Review of Funder Instructions and Grant Reviewer Practices for Assessing the Intellectual Merit and Other Impacts of Research*. Alexandria, VA: U.S. National Science Foundation.

Mathematica Inc.

Our employee-owners work nationwide and around the world.
Find us at **mathematica.org** and **edi-global.com**.

Mathematica
Progress Together