Discovery Stack Pilot: Feasibility and Outcomes of a Scientist-Designed Peer Review Model Separating Quality and Impact

Maureen A. McGargill[1], Beiyun C. Liu[2], Michael S. Kuhns[3], Daniel Mucida[4,5], Isabella Rauch[6], Lauren B. Rodda[6], Meghan A. Koch[7,8], Hugo Gonzalez[9,10,11], Ken Cadwell[12,13], Tanya S. Freedman[14,15,16], Tiffany C. Scharschmidt[17], Richard Sever[18,19], Jose Ordovas-Montanes[20,21,22,23,24], Andrew Oberst[8], Brooke Runnette[1], and Matthew F. Krummel[25]

Affiliations:

1. Solving for Science, 2261 Market St Ste 5966, San Francisco, CA 94114
2. St. Jude Children's Research Hospital, Strategic Communication Education & Outreach, Memphis, TN 38105
3. Department of Immunobiology, The University of Arizona, Tucson, AZ 85724
4. Laboratory of Mucosal Immunology, Rockefeller University, New York, NY 10063
5. Howard Hughes Medical Institute, Rockefeller University, New York, NY 10063
6. Department of Molecular Microbiology and Immunology, Oregon Health and Science University, Portland, OR 97239
7. Basic Sciences Division, Fred Hutchinson Cancer Center, Seattle, WA.
8. Department of Immunology, University of Washington, Seattle, WA, 98109
9. Centro Basal Ciencia & Vida, Fundación Ciencia & Vida, Santiago, Chile
10. Facultad de Ciencias Para el Cuidado de la Salud, Universidad San Sebastian, Santiago, Chile
11. Department of Anatomy, University of California, San Francisco, San Francisco, CA 94143
12. Division of Gastroenterology and Hepatology, Department of Medicine, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA 19104
13. Department of Pathobiology, University of Pennsylvania School of Veterinary Medicine, Philadelphia, PA 19104
14. Department of Pharmacology, University of Minnesota, Minneapolis, MN 55455,
15. Center for Immunology, University of Minnesota, Minneapolis, MN 55455
16. Masonic Cancer Center, University of Minnesota, Minneapolis, MN 55455
17. Department of Dermatology, University of California, San Francisco, CA 94143
18. openRxiv, Davis, CA 95616
19. Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11724
20. Division of Gastroenterology, Hepatology, and Nutrition, Boston Children's Hospital, Boston, MA 02115
21. Broad Institute of MIT and Harvard, Cambridge, MA 02142
22. Harvard Stem Cell Institute, Cambridge, MA 02138
23. Ragon Institute of MGH, MIT and Harvard, Cambridge, MA 02139
24. Program in Immunology, Harvard Medical School, Boston, MA 02115
25. UCSF Bakar ImmunoX Initiative and Department of Pathology, University of California, San Francisco, San Francisco, CA 94143

ABSTRACT

Peer review serves as the cornerstone of scientific quality control. Yet, the current journal-centric system is hindered by long timelines, high publication costs, inconsistent review quality, systemic biases, and editorial gatekeeping. Notably, the system is built around misaligned measures of impact that are tethered to journal branding and conflate scientific rigor (*Quality*) with perceived significance (*Impact*). Here, we report findings from the Discovery Stack Pilot Study, which tested a scientist-designed, journal-independent peer review model. The Discovery Stack model integrates in-line reviewer comments to promote constructive, improvement-focused feedback and generates separate, multimodal assessments of scientific *Quality* and *Impact*. To examine its feasibility and effectiveness, manuscripts enrolled in the pilot were reviewed in parallel with traditional journal review. A total of 162 reviews were completed, and survey data from 86 participants were analyzed to evaluate the experience of both authors and reviewers. The results showed that reviewers effectively evaluated *Quality* and *Impact* as separate dimensions, with *Quality* scores being more consistent across reviewers than *Impact* scores. Importantly, participants strongly supported the core elements of the Discovery Stack model and expressed enthusiasm for its broader adoption to enhance transparency, efficiency, and value in peer review. Future studies will explore integrating this model into a digital platform for reviewing and curating scientific discoveries to improve the production and dissemination of high-quality research.

INTRODUCTION

Manuscript publication is the primary way researchers share new discoveries. Peer review prior to publication remains the central mechanism for evaluating the scientific rigor and validity of these findings (1). Researchers depend on this process to guide future studies, validate results, and uphold their professional reputation. Likewise, funding agencies, regulatory bodies, academic institutions, and industry stakeholders rely on peer-reviewed research, and the prestige of the journals in which it is published, to inform critical decisions on funding, policy, faculty promotion, media communication, product development, and patient care. In short, peer review underpins nearly every aspect of how scientific knowledge is generated, communicated, applied, and valued.

Despite its central role, the current journal-centric peer review and scientific publishing system often fails to meet the needs of scientists and society (2–6). A key limitation is the lack of a clear distinction between scientific *Quality* and *Impact*. *Quality* refers to the rigor and reproducibility of the data supporting a study's conclusions, whereas *Impact* reflects the extent to which the findings advance understanding. When these dimensions are blurred, high-impact or hyped findings can overshadow weak evidence, while rigorous but incremental work is undervalued. Moreover, perceived *Impact* is often inferred from journal prestige rather than the intrinsic merit of the research itself (7, 8). Collectively, these dynamics distort how scientific contributions are valued, reinforcing journal reputation over scientific merit and undervaluing confirmatory studies that are essential for establishing confidence in foundational discoveries.

The traditional publishing process is also slow and inefficient. Manuscripts are considered by only one journal or journal family at a time, and each review cycle may take months. Across disciplines, the average time from submission to acceptance is approximately six months, but frequently extends well beyond a year (9, 10).

The quality, bias, and transparency of peer review are additional concerns. Reviews vary widely in depth and rigor, and critical flaws are sometimes missed (2–5, 10, 11). The lack of formal training and standardized guidelines contributes to this inconsistency (12). Reviewer anonymity, while intended to promote objectivity, can also shield bias and hostility from accountability. Together, these challenges undermine the effectiveness of peer review as a mechanism for quality control and diminish its value to authors.

Compounding these challenges, researchers perform peer review labor without compensation, while also paying to publish and access scientific literature. This model is inequitable, frustrating, and increasingly unsustainable, as publication numbers continue to rise without a proportional increase in the number of available reviewers (9).

The growing prevalence of open-access preprint servers has improved accessibility to new findings (8, 13). However, because these platforms often lack meaningful peer review or metrics of rigor, they have created a new problem: information overload without effective mechanisms to search or filter studies based on *Quality* or *Impact*, leaving readers uncertain about which studies to trust and prioritize.

To address these challenges and accelerate scientific progress, the peer review process must be strengthened and modernized. We posit that an improved system should emphasize a “peer-improvement” mindset, positioning reviewers as collaborators focused on strengthening scientific rigor and benefiting the scientific community, rather than functioning as journal consultants determining binary publication eligibility. Such a system should apply standardized metrics to assess a study’s *Quality* and *Impact* as separate dimensions (14). Moreover, because *Impact* evolves over time through ongoing evaluation and influence on subsequent studies, this metric should remain dynamic and independent of journal branding.

The Discovery Stack Pilot Study tested the feasibility and effectiveness of a new peer review model built on these principles. The pilot evaluated the outcomes of separately assessing manuscript *Quality* and *Impact*, applying standardized metrics to independently measure and report both dimensions, and using an in-line commenting tool to facilitate constructive contextual feedback directly to authors and readers. The pilot also explored mechanisms for improving transparency, accountability, and timeliness, aiming to shorten the time from submission to dissemination. Findings from the Discovery Stack Pilot provide a foundation for further refinement and optimization of approaches to improve scientific publishing and peer review.

RESULTS

Discovery Stack Model

The Discovery Stack Pilot was designed to test a structured, multi-phase peer review process that separates the evaluation of scientific *Quality* from potential *Impact*. The process included three sequential phases: 1) *Quality* Review, 2) Author Response, and 3) *Impact* Review (**Figure 1A**). Each phase was built on the previous one, with *Impact* reviewers able to view compiled *Quality* review feedback and author responses. Detailed procedures for each phase are provided in Materials and Methods.

Manuscripts were eligible for enrollment if they: (1) were posted to a preprint server such as bioRxiv, which enabled testing of in-line comments through the Hypothes.is platform; (2) had been submitted to a traditional journal, allowing for comparison with conventional peer review; and 3) were in the fields of immunology or cancer biology, enabling us to leverage the subject expertise of the scientific advisory board and participating reviewers.

To identify eligible manuscripts, initial invitations were sent to individuals who had previously registered to participate as authors or reviewers in the Discovery Stack Pilot. Outreach was then expanded to authors of recent bioRxiv preprints. Three additional bioRxiv manuscripts were selected by the pilot's scientific advisory board and reviewed without author input.

In total, 18 manuscripts were enrolled (**Figure 1B**, generating 162 reviews: 50 *Quality* and 112 *Impact* reviews (**Figure 1B-C**). Each manuscript received at least two *Quality* and five *Impact* reviews, with an average of 2.8 *Quality* and 6.5 *Impact* reviews per manuscript. Because we hypothesized that *Impact* assessments would be inherently more subjective and variable, we aimed to recruit more *Impact* (six) than *Quality* (three) reviewers per manuscript. Most manuscripts met or exceeded this goal, demonstrating

the feasibility of enrolling manuscripts, recruiting reviewers, and completing both *Quality* and *Impact* assessments.

Reviewers Successfully Distinguished *Quality* from *Impact*

As separate evaluation of a manuscript's *Quality* and *Impact* is a core feature of the Discovery Stack model, we examined whether reviewers evaluated these dimensions independently. First, composite *Quality* and *Impact* scores for each manuscript were plotted in the order of the journal impact factor to which it was submitted, which served as a proxy for the author's perception of *Impact* (**Figure 2A**). This visualization showed several manuscripts rated high in *Quality* but low in *Impact*, suggesting that reviewers were able to uncouple their assessment of *Impact* from *Quality*. To formally test this, we quantified the relationship between these two dimensions. To ensure sufficient evaluations for each manuscript, *Quality* reviewers were also asked to provide a separate assessment of the same manuscript's *Impact*. Given the greater subjectivity of *Impact* assessments, additional reviewers were recruited to evaluate *Impact* only, resulting in more *Impact* than *Quality* scores per manuscript. To avoid detecting differences driven by unequal sample sizes and reviewer variability, we restricted the primary analysis to reviewers who provided both *Quality* and *Impact* scores for the same manuscript. Pearson's correlation showed a significant positive association between *Quality* and *Impact* ($r = 0.70$, $p=0.0017$), and linear regression confirmed that *Quality* significantly predicted *Impact* (**Figure 2B**; $\beta = 0.80$, $R^2 = 0.49$ $p=0.0017$). As expected, poor-quality manuscripts are unlikely to be considered impactful. However, residual analysis indicated substantial divergence. Nine of 17 manuscripts (52.9%) had *Impact* scores outside the 95% confidence interval (**Figure 2B**), with residuals ranging from −0.94 to +0.55 *Impact* points (**Figure 2C**). The standard deviation of residuals (SD = 0.41) illustrates considerable variation around the regression line, indicating that reviewers' *Impact* evaluations frequently diverged from predictions based solely on

Quality. Analyses including all reviewer scores yielded comparable results (r = 0.76, *p*=0.0004; β = 0.77, R^2 = 0.57, *p*=0.0004).

To further assess the independence between *Impact* and *Quality*, manuscripts were grouped into tertiles by composite *Quality* score, and mean *Impact* scores were compared the across groups. Mean *Impact* scores increased with higher *Quality*, but significant differences were observed only between the lowest and highest tertiles, with substantial overlap between adjacent groups (**Figure 2D**). The residual variation (SD = 0.41) was nearly as large as the average difference in *Impact* between tertiles (0.46–0.48), indicating that manuscripts with comparable *Quality* scores frequently differed in their *Impact* ratings.

Next, we examined whether reviewer scores aligned with the journal impact factors to which the manuscripts were submitted. It is important to note that Discovery Stack reviewers were unaware of which journals the authors selected. *Quality* scores were not correlated with journal impact factors (**Figure 2E;** r = -0.47, *p*= 0.059), whereas *Impact* scores were significantly correlated (**Figure 2F;** r = -0.56, *p*=0.0186). Again, analyses including all reviewer scores yielded comparable results (*Quality* r = -0.48, *p*= 0.051; *Impact* r = -0.64, *p*= 0.0059). These findings suggest that reviewers' perceptions of *Impact*, but not *Quality*, modestly aligned with the authors' expectations of significance. Together, these results demonstrate that reviewers treated *Quality* and *Impact* as separate but complementary dimensions of manuscript evaluation.

A limitation of this analysis is that comparisons between *Impact* scores of revised manuscripts to journal impact factors in which manuscripts were ultimately published was not possible, as reviewers only assessed initial submissions and only six of 18 manuscripts had been accepted at the time of analysis. Additionally, it is possible that

reviewers inferred the tier of journal selected by authors based on formatting of the preprint.

Widespread Endorsement for Standardized Metrics and Separate Evaluation

Since separating *Quality* from *Impact* was a central component of this pilot study, we evaluated participants' perceptions of how effectively the platform supported this distinction and whether doing so enhanced the review process. The vast majority of reviewers (93.0%) agreed or strongly agreed that the platform effectively supported separate assessments of *Quality* and *Impact* (**Figure 2G**). Moreover**,** 84.9% of authors and reviewers agreed that separating these dimensions led to more constructive and insightful evaluations than traditional reviews (**Figure 2H)**.

We also examined perceptions of rating standardized attributes as a means to establish quantitative metrics for scientific rigor and *Impact*. Support for standardized metrics was compelling, with 90% of participants agreeing that standardized *Quality* ratings could generate meaningful metrics of rigor (**Figure 2I**), and 82% agreeing that standardized *Impact* ratings could capture perceived significance (**Figure 2J**). Reviewer support exceeded author support for both metrics (92% vs. 72% for *Quality*; 86% vs. 55% for *Impact*), though the author sample was smaller than the reviewer sample (n=11 vs. n=75).

These findings demonstrate broad endorsement of two core elements of the Discovery Stack model: (1) the separation of *Quality* and *Impact,* and (2) the use of standardized metrics to increase transparency, improve review quality, and reduce reliance on journal branding as a measure of scientific value. Because participation in this pilot was voluntary, participants may have been more receptive to alternative models of peer review than the broader scientific community, which should be considered when interpreting these results.

Impact Reviews Exhibit Greater Variability Than *Quality* Reviews

Visual inspection of side-by-side *Quality* and *Impact* scores suggested that, within individual manuscripts, *Impact* scores varied more than *Quality* scores (**Figure 2A**). To compare score variability, reviewer score dispersion was measured using three complementary metrics: standard deviation (SD), range (maximum - minimum score), and interquartile range (IQR; difference between the 75th and 25th percentile). Across all three metrics, *Impact* scores consistently exhibited greater dispersion than *Quality* scores (**Figure 2K-M**). To account for unequal numbers of reviewers, small sample sizes, and non-parametric distributions, we performed bootstrap resampling under two conditions: unmatched, which used all available *Impact* and *Quality* scores, and matched, which resampled an equal number of scores per manuscript. In both conditions, *Impact* scores remained significantly more variable than *Quality* scores across all three metrics, indicating that the observed effect was not due to differences in reviewer counts (**Figure 2N**). These findings underscore the value of evaluating *Quality* and *Impact* as distinct dimensions and support the need for a greater number of *Impact* reviewers to capture the broader range of perspectives on scientific significance.

Identity Disclosure Enhances Perceived Transparency but Affects *Impact* Scores

To promote transparency, reviewers were encouraged to disclose their identity to authors and co-reviewers, although anonymity remained an option to preserve the integrity of the review process. Approximately half of reviewers disclosed their identity: 50% of *Quality* reviewers and 54% of *Impact* reviewers (**Supplementary Figure 1A**). Interestingly, a greater proportion of trainees (73%) than principal investigators (47%) identified themselves (**Supplementary Figure 1B**). The proportion of identified reviewers varied substantially across manuscripts (**Supplementary Figure 1C**; range:

0-100%), suggesting that factors such as authorship or perceived study *Quality* may have influenced identity disclosure decisions. The most common reasons for remaining anonymous were familiarity with the authors and concern about professional repercussions (**Supplementary Figure 1D**).

Most authors (72.7%) agreed that the Discovery Stack review was more transparent than traditional review (**Supplementary Figure 1E**). Although, the number of author responses was small (n=11), authors whose manuscripts had a higher proportion of identified reviewers were more likely to perceive increased transparency (**Supplementary Figure 1F**), suggesting that identity disclosure contributes to perceived transparency.

Quality reviewers' feedback was shared with authors and *Impact*-only reviewers, allowing both groups to evaluate whether reviewer identity influenced scoring. Most *Impact*-only reviewers (89%) and authors (70%) reported observing no clear differences in scores between identified and anonymous reviewers (**Supplementary Figure 1G-H**). However, four out of 10 authors stated that identified reviews were more constructive than anonymous reviews, consistent with comments from *Impact*-only reviewers, noting that anonymous feedback tended to be briefer and less engaged. These mixed responses suggest that reviewer identification may enhance engagement for some reviewers, but not universally.

To directly evaluate whether reviewer identity influenced scores, we compared individual scores from anonymous and identified reviewers across all manuscripts (unpaired) and within manuscripts (paired). The paired analysis assessed whether, for a given manuscript, scores differed between anonymous and identified reviewers. There were no significant differences in *Quality* scores between the two groups (**Supplementary Figure 1I-J**), but *Impact* scores were significantly higher among

identified reviewers (**Supplementary Figure 1K-L**). This difference was consistent in the unpaired analysis across all manuscripts and within manuscripts in the paired analysis. Additionally, the higher *Impact* scores from identified reviewers was not explained by reviewer career stage (**Supplementary Figure 1M**).

In summary, identity disclosure was associated with greater perceived transparency and higher *Impact* scores, suggesting reviewers may be more likely to identify themselves when giving favorable evaluations or that reviewers were more inclined to soften their subjective assessment of *Impact* after agreeing to disclose their identity. These results highlight the need for further study on how identity disclosure influences the review process and its perception.

Discovery Stack Platform Delivers a Better Experience than Traditional Review

A central goal of the Discovery Stack Pilot was to assess whether participants believed that the platform offered a better experience than traditional peer review. Overall, a strong majority (82.6%) rated their experience as "*much better"* or "*slightly better"*, while only 3.5% rated it as worse (**Figure 3A**). Positive ratings were highest among *Impact*-only reviewers (94.0%), followed by authors (81.8%), and *Quality/Impact* reviewers (73.8%). The lower satisfaction among *Quality/Impact* reviewers may reflect the additional effort required to complete both review phases and learn the Hypothes.is platform.

Only three participants reported a worse experience, citing the need to consult setup instructions, perceived platform complexity, or uncertainty about how their reviews affected manuscript outcomes. These challenges are typical of new systems and are expected to diminish with familiarity. Additionally, because the pilot ran in parallel to traditional review, authors were not required to respond to Discovery Stack feedback, which limited reviewers' insight into the impact of their efforts.

A core tenet of the Discovery Stack model is to promote a “peer-improvement” mindset, encouraging reviewers to provide constructive, actionable feedback that enhances scientific rigor. Most participants (75%) agreed that the platform fostered this mindset more effectively than traditional reviews, with stronger agreement among reviewers (81%) than authors (55%) (**Figure 3B**).

To gain qualitative insight, participants were asked open-ended questions about the most beneficial and most challenging aspects of the platform compared to traditional review. Consistent with the positive platform ratings, more participants cited benefits (n=74) than challenges (n=56). The most common benefits were in-line commenting (34.9%), separation of *Quality* and *Impact* reviews (19.8%), and standardized questions and scoring (18.6%) (**Figure 3C**). The most frequent challenges were setting up and learning Hypothes.is (19.8%), Hypothes.is limitations (18.6%), and time demands (10.5%) (**Figure 3D**).

Although challenges with the Hypothes.is tool were most frequently cited, in-line commenting was also the most reported benefit, highlighting both its value as a core feature and the need for technical enhancements. With 82% of participants reporting a better experience than traditional review, these findings provide strong support for Discovery Stack and its potential for broader adoption with continued optimization.

In-line Commenting Improves Review Clarity, Collegiality, and Efficiency

Given that in-line commenting was both the most cited benefit and key area for improvement, we evaluated its effectiveness in improving review clarity, collegiality, and efficiency. The Hypothes.is tool was selected because it enabled contextual annotation of bioRxiv-hosted preprints within private groups, allowing reviewers to leave feedback directly on the manuscript text. Responses from both reviewers and authors strongly supported this feature. Most authors and *Impact*-only reviewers (71%) agreed that in-

line comments were more collegial and constructive than traditional reviews (**Figure 4A**). Notably, all authors found in-line comments easier to respond to and more helpful for identifying needed revisions than traditional review summaries (**Figure 4B**).

Among *Quality/Impact* reviewers, 93% agreed that in-line comments made it easier to highlight errors in logic or clarity, 86% felt in-line commenting improved the constructiveness of feedback, and 71% preferred addressing comments directly to the authors, believing that their in-line comments would be more helpful than a traditional review summary (**Figure 4C**). Although the Hypothes.is tool was new to most reviewers, 73% reported that it was easy to use, and 60% agreed that in-line commenting improved review efficiency. Additionally, 67% agreed that in-line commenting helped them adopt a more collegial tone (**Figure 4C**). Together, these findings indicate that in-line commenting is both feasible and effective, providing clearer, more constructive, and more collegial feedback than traditional review.

Authors Find Discovery Stack *Quality* Reviews Rigorous and Helpful

We next evaluated whether the Discovery Stack model effectively directed reviewers to focus on scientific rigor by asking authors about their perception of the *Quality* reviewer feedback. Most authors (81.8%) agreed that *Quality* reviews focused on scientific rigor, and 72.7% found the standardized *Quality* and *Impact* scores helpful for understanding how their manuscript was evaluated (**Figure 5A**)**.** Likewise, 81.8% agreed the *Quality* Assessment Form effectively summarized the key strengths and weaknesses, and more than 60% reported revising their manuscript based on reviewers' feedback. In-line commenting was also well received, with 87.5% of authors agreeing that the opportunity to interact with reviewers through Hypothes.is provided a good method to clarify feedback and expedite the review process (**Figure 5A**).

When asked to compare the overall Discovery Stack feedback to traditional review, 45.5% of authors rated it better, and 45.5% rated it similar, and only one rated it worse. (**Figure 5B**). These findings indicate that Discovery Stack delivered rigorous, constructive, and actionable *Quality* reviews. Although limited by a small author sample (n=11), the results support the models' potential to improve consistency and utility of peer review.

Impact **Assessment Viewed as Useful and Strengthened by Prior** ***Quality*** **Review**

Next, we examined participants' impressions of the *Impact* review and its value in assessing the broader significance and potential influence of a manuscript's findings. Most reviewers (77%) reported that evaluating *Impact* independently made it easier to assess significance (**Figure 5C**), consistent with the model's premise that separating *Quality* and *Impact* enables a more nuanced evaluation.

Reviewers also found the *Impact* review to be well-designed and intuitive; 93% agreed the steps were clear and manageable, 88% found the *Impact* Assessment Form helpful for estimating a manuscript's "must-read" value, and 89.4% agreed the form included essential criteria for evaluating significance (**Figure 5C**). Nearly all respondents (91.8%) agreed that identifying the criteria applicable to each manuscript was essential for an accurate evaluation (**Figure 5C**). Simultaneously, reviewers acknowledged challenges of reliably assessing *Impact*, with 58.1% agreeing that assessing *Impact* requires more than three reviewers, and 89.2% agreeing that a study's true *Impact* often becomes clear only over time, highlighting the importance of dynamic, evolving *Impact* metrics.

To facilitate the separate evaluation, the pilot was designed so that *Quality* reviews were completed before the *Impact* assessment. This allowed *Impact* reviewers to consider *Quality* review feedback while evaluating significance. More than 90% of *Impact*-only reviewers agreed that access to the *Quality* reviews improved their ability to

evaluate *Impact* (**Figure 5D**). These findings indicate that reviewers viewed the *Impact* assessment as well-structured, valuable, and strengthened by prior *Quality* review.

Reviewer Engagement Relies on Outreach, Familiarity, and Trainee Involvement

Recruiting reviewers is a persistent challenge in peer review and a critical barrier to timely evaluations (9, 15). This challenge was amplified for the Discovery Stack Pilot, as most researchers were unfamiliar with the platform and each manuscript required six reviewers. Acceptance rates varied dramatically depending on reviewers' familiarity with the study and whether they received personal outreach. Among individuals with no prior connection to the pilot, the acceptance rate was only 8.6% (**Supplementary Figure 2A**). In contrast, reviewers already enrolled in the pilot accepted invitations at a much higher rate of 58.8% ($p < 0.0001$, Fisher's Exact test).

To improve reviewer participation among new reviewers, members of the Scientific Advisory Board (SAB) sent follow-up emails to individuals in their respective networks, which substantially raised the acceptance rate among new reviewers to 53.3% (**Supplementary Figure 2A**). Personal outreach increased acceptance odds by more than 12-fold (Odds Ratio (OR)=12.2, $p<0.0001$), while prior enrollment increased odds by 15-fold (OR=15.2, $p<0.0001$). With these strategies in place, for the 17 bioRxiv manuscripts reviewed in the pilot, 283 review invitations were sent (160 for *Quality*/*Impact* and 123 for *Impact*-only reviews), yielding an overall acceptance rate of 32.5% (**Supplementary Figure 2B-C**), which is similar to the 32.3% acceptance rate reported in 2022 by Clarivate's ScholarOne platform, which supports over 8,000 journals (15). Overall, 90% of accepting reviewers were either familiar with the study or personally contacted by someone in their network. Survey data reinforced this trend, with 73% of reviewers and 67% of authors reporting they heard about the pilot study prior to participating or alternatively received outreach from a colleague after the initial

reviewer request (**Supplementary Figure 2D**). These findings underscore that outreach and professional networks substantially improve reviewer engagement.

Although trainees who reviewed independently of their advisors represented a small proportion of reviewers (27.8%; **Supplementary Figure 2E**), their acceptance rate was markedly higher than that of principal investigators. After adjusting for familiarity and outreach, trainees remained 15 times more likely to accept review invitations (OR=15.3, $p<0.0001$; **Supplementary Figure 2F**). These findings highlight that actively training and recruiting trainees may be an effective strategy for increasing reviewer engagement.

Expedited Review Is Achievable with Reviewer Accountability and Backup Plans

Another persistent challenge in peer review is the prolonged duration between manuscript submission and publication, which slows the dissemination of scientific findings. This was evident for manuscripts included in the pilot. At the time of analysis, only six of 18 manuscripts had been published. The mean from submission to publication for the published manuscripts was 231 days. The remaining manuscripts were still under review or undergoing revisions, with an average of 395 days since submission (**Figure 6A**).

Across the pilot, the average time from reviewer recruitment to completion of *both* the *Quality* and *Impact* reviews was 86 days (range: 42–125) (**Figure 6B-C**). The *Quality* phase averaged 51 days and the *Impact* phase 33 days. Reviewer recruitment averaged 15-16 days per phase. The most time-consuming component was completion of *Quality* reviews (35 days) (**Figure 6D**).

Since the pilot included two review phases, total duration was not directly comparable to traditional review. Therefore, we compared *Quality* review duration to the initial stage of journal review. Discovery Stack *Quality* review averaged 52 days, which was

comparable to the author-reported journal average of 58 days (range: 30–152) (**Figure 6E**). Nevertheless, over half of authors (55%) reported receiving Discovery Stack *Quality* reviews faster than journal reviews (**Figure 6F**).

To understand why the overall review process was not faster, given that reviewers agreed to complete reviews within 14 days, we examined factors contributing to delays in review completion. Individual turnaround times were close to the expected timelines, with *Quality* reviews averaging 18 days and *Impact* reviews 14 days (**Figure 6G**). Moreover, 61% (roughly two out of three) of *Quality* reviewers and 78% of *Impact* reviewers submitted their reviews within three days of the deadline. Visualization of individual reviewer times per manuscript demonstrated that delays were typically due to a single late reviewer rather than widespread delays across all reviewers (**Figure 6H**).

To investigate whether the time required to complete each review contributed to delays in review completion, reviewers were asked to estimate their time investment. *Quality/Impact* reviewers spent an average of 3.50 ± 0.97 hours per review, and *Impact*-only reviewers spent 1.66 ± 0.84 hours (**Figure 6I**). When asked to evaluate the efficiency of the review process, considering both the time invested and the value provided to authors, 90.5% of *Quality/Impact* reviewers rated it as equal to or more efficient than traditional review (**Figure 6J**).

Together, these findings indicate that although the Discovery Stack model was new to reviewers, the time commitment was reasonable, reviewers largely adhered to deadlines, and overall delays were driven by isolated late reviews. Therefore, expedited peer review is achievable if delays associated with a single late reviewer are mitigated through clear accountability and contingency plans.

Scientists Show Strong Interest in New Publishing Models but Hesitate to Submit

To gauge interest in alternative publishing approaches, we first asked participants to share their perspectives on needed reforms to scientific publishing. Two open-ended questions invited participants to describe what they believe needs to change in the current system, and if designing a new model from scratch, to identify its three most essential principles. The most prominent issues for change focused on inefficiencies in the peer review process, unconstructive or biased reviews, and excessive reviewer demands (**Supplementary Figure 3A-C)**. These priorities align closely with the goals of Discovery Stack model.

Participants were then asked whether they would use a new platform that applies scientist-developed metrics to evaluate and curate peer-reviewed research. Across all participants, 92% indicated they would “definitely” or “probably” use such a platform, with only 5% undecided and 4% not interested (**Figure 7A**).

When asked whether they would submit an original research paper to such a platform, 71% indicated they would “definitely” or “probably” submit a manuscript, 24% were undecided, and only 6% were not interested (**Figure 7A**). The lower willingness to submit manuscripts likely reflects the continued influence of traditional journals as the key metric of researcher productivity. Together, these results reveal both a strong interest for publishing models that emphasize transparency and rigor, and the challenge of overcoming the hesitancy to depart from traditional journals. Addressing this adoption barrier will likely require strategies to increase the platform’s visibility, its credibility, and institutional recognition.

Key Features That Will Drive Adoption of a New Platform for Peer Review

The Discovery Stack Pilot tested multiple features that could inform the design of a new publishing platform. Participants rated these elements to reveal which features would most affect their willingness to use a novel platform for evaluating and curating peer-reviewed research. The most frequently prioritized elements included a peer-improvement" mindset (88.2%), a "peer-approved" status for validated manuscripts (87.1%), faster publication timelines (83.5%), a scientist-owned platform free of journal revenue and branding (81.2%), reviewer feedback metrics (76.5%), scientist-developed *Quality* metrics (76.5%), and evolving *Impact* metrics based on *ongoing* community input (71.8%) (**Figure 7B**).

These preferences mirror priorities expressed in the open-ended responses describing needed changes in the current system and core principles that should guide an improved system (**Supplementary Figure 3A-C**), reinforcing support for a "peer-improvement" mindset, a transparent "peer-approved" status, and faster progression from preprint to publication.

Participants were also invited to suggest additional features they would like to see in a future platform. The most frequently mentioned feature was the formal adoption of the platform by funding agencies, hiring committees, and promotion committees, reflecting the importance of institutional recognition for platform credibility. Other suggestions included easier access to raw data, code, and protocols; moderation of comments and vetting of users and reviewers; and an initial quality screen before review.

Collectively, these results demonstrate that several of the core features tested in the Discovery Stack Pilot align with researchers' priorities and highlight the need for institutional recognition of the platform to overcome hesitancy to depart from traditional journals.

DISCUSSION

The Discovery Stack Pilot evaluated the feasibility and value of a scientist-designed peer review model that separates scientific rigor (*Quality*) from perceived significance (*Impact*), uses standardized metrics, incorporates in-line commenting, and promotes a peer-improvement mindset. Conducting reviews in parallel with traditional journal reviews enabled a direct comparison of feasibility and effectiveness. Several important findings emerged. Notably, reviewers successfully evaluated *Quality* distinctly from *Impact*, and *Impact* scores showed greater variability, highlighting the need for larger reviewer sample sizes. Reviewer engagement strongly depended on targeted outreach and leveraging personal networks. Delays in review completion were typically due to a single late reviewer, underscoring the importance of accountability and backup plans. Identity disclosure improved perceived transparency, but was also associated with higher *Impact* scores, warranting further study. Finally, participants strongly endorsed core elements of the model, demonstrating its feasibility, perceived value, and readiness for broader adoption. Because participation was voluntary, it is important to acknowledge the potential for selection bias. Individuals who chose to enroll may have been more motivated to support alternative models than the broader scientific community, which should be considered when interpreting the strength of the observed support.

A major outcome of this pilot was empirical support for separating *Quality* from *Impact*, as we had previously proposed separating these two dimensions to address a long-standing problem in which perceptions of significance can mask concerns about rigor, and vice versa (14). Although *Quality* and *Impact* were positively associated, as expected, when poor rigor limited perceived importance, residual and tertile analyses revealed frequent divergence. Additionally, *Impact* scores better aligned with journal impact factors than *Quality* scores. Together, these data demonstrate that reviewers

assessed rigor independently of significance. Importantly, most reviewers (93.0%) agreed that the platform effectively supported separate assessments, and 84.9% found that separating these dimensions led to more constructive and insightful reviews. Separating *Quality* and *Impact* would substantially benefit science by assigning value to high-*Quality* work, independent of novelty, creating space for careful replication studies, incremental clarifications, and contrarian findings that the current novelty-oriented model often overlooks.

Data from the pilot also revealed a greater dispersion in *Impact* scores relative to *Quality*, reflecting inherent subjectivity in evaluating significance, which is sensitive to field coverage, methodological preferences, translational focus, and individual research agendas. Operationally, this finding supports the use of a larger number of *Impact* reviewers per manuscript to capture a representative distribution of viewpoints. It also argues for dynamic *Impact* metrics that can evolve as replication, uptake, and downstream influence become visible, rather than fixing perceptions at the time of initial review as is the case with the current publication model.

The pilot also aimed to test the feasibility and utility of standardized metrics. Participants strongly endorsed metrics for both *Quality* and *Impact*, 90% for *Quality* and 82% for *Impact*. In addition to providing data to generate metrics, the Likert-scale questions aligned reviewer attention to the core elements of rigor (methodology, controls, statistics, concordance between data and conclusions) or significance (extent of advance in scientific understanding, filling of a knowledge gap, “must-read” value), supporting more consistent evaluation across manuscripts. In future implementations of the platform, a threshold for high-*Quality* papers may emerge, above which manuscripts would be designated as rigorous and trustworthy. While overreliance on metrics alone risks oversimplification (16), these hazards can be mitigated when metrics are based on clearly defined attributes, coupled with narrative feedback and in-line annotations, and

routinely evaluated. Implemented in this way, rich metrics can function as important decision aids that enable readers to filter and search by features they find most meaningful when deciding what science to read and value.

Timeliness remains one of the most significant challenges in peer review. Despite emphasizing deadlines and streamlining reviews by using in-line comments instead of lengthy narratives, the Discovery Stack *Quality* review phase averaged about eight weeks, matching the initial journal review rather than shortening it as intended. Individual reviewers generally adhered to the expected timelines, and most delays were caused by a single reviewer. Replacing a late reviewer rarely shortens the overall review time because confirming a replacement takes time, and the new reviewer requires time to complete the review. While many journals address this by relying on two reviewers instead of three, this approach risks weakening the overall quality of the evaluation. The Discovery Stack Pilot demonstrated that reducing the review duration requires more than workflow simplification. Rather, clear accountability and pre-planned backup coverage are required. Practical steps include confirming four reviewers at the outset, recruiting “alternate” reviewers, and rewarding timely submissions with monetary compensation and recognition. Notably, a recent study combining pre-recruitment of expert reviewers, compensation, and strict deadlines achieved review turnaround times of less than one week, illustrating how innovative recruitment strategies and incentives can expedite the review process (17).

Reviewer recruitment remains another bottleneck in peer review and adds to the total review time. In our pilot, confirming three *Quality* reviewers averaged two weeks per manuscript, a nontrivial delay that accounted for one quarter of the eight-week *Quality* review phase (9, 15). Familiarity with the project and personal outreach markedly improved reviewer acceptance rates. Additionally, trainees were especially responsive to accepting review assignments. These insights indicate that visible recognition,

modest compensation, and formal training for early-career scientists can expand the reviewer pool, while also providing a path for trainees to become constructive critics and establish their own scientific reputations.

Survey responses revealed broad dissatisfaction with current publishing norms and strong enthusiasm for reform. Participants consistently cited poor review caliber, slow review process, excessive reviewer demands, for-profit publishing models, lack of reviewer compensation, and high cost to publish and access scientific papers as top priorities for change. The broad consensus on these priorities underscores the need for bold changes. Encouragingly, 92% of participants indicated their willingness to use a scientist-designed platform to evaluate and curate peer-reviewed research. Features most likely to drive adoption align with the Discovery Stack model, including a peer-improvement mindset, a “peer-approved” status to indicate scientific validation, faster timelines to publication, scientist ownership, and reviewer feedback metrics. Although learning the in-line annotation tool was the most frequently mentioned challenge of the Discovery Stack model, it was also cited as one of the top benefits, streamlining the *Quality* review time, encouraging collegiality, and improving clarity. In future implementations, in-line commenting could shorten the review time by allowing authors to address feedback in real time. With these top-ranked features, the Discovery Stack model provides a strong foundation for refinement and scaling.

Although willingness to submit manuscripts was strong (71%), it lagged behind readiness to use a platform for evaluation and curation, reflecting dependence on journal branding and impact factors for career advancement and funding decisions, which remains the primary barrier to adopting a new model. Although flawed, journal impact factors remain the primary metric used by funding and academic institutions to assess research values. While reform movements are emerging (7, 16, 18), widespread

change will depend on adoption by major stakeholders and a firm understanding of how alternative metrics function, which was a central goal of this pilot.

This study has limitations, including a modest sample size concentrated in immunology and cancer biology, incomplete publication outcomes for some manuscripts, and reliance on an external annotation tool. Additionally, most evaluations were of initial submissions, which limited analysis of score dynamics across revisions. Because participation was voluntary, the study population may not fully represent the broader scientific community, introducing potential selection bias. These constraints are typical of feasibility studies and can be addressed in future iterations.

In summary, the Discovery Stack Pilot demonstrated that reviewers reliably evaluated *Quality* and *Impact* as separate dimensions, and that *Impact*, by its nature, varies more than Quality. The model's core elements of standardized metrics, in-line annotation, peer-improvement mindset, and distinct *Quality* and *Impact* evaluations enhanced clarity and constructiveness, earning considerable support. Together, these results provide a practical foundation for a peer review system for manuscripts and discoveries that addresses key limitations of the current system and is ready for expansion and adoption.

MATERIALS AND METHODS

Study Design

The Discovery Stack Pilot consisted of three sequential phases designed to test a structured, peer review process that separately evaluated *Quality* and *Impact*. These phases were: 1) *Quality* Review, 2) Author Response, and 3) *Impact* Review.

Quality Review Phase

The *Quality* review phase provided a structured, standardized evaluation of each manuscript's scientific rigor. Participants were informed that *Quality* referred to the extent to which a study's experimental design, methodology, controls, statistical analyses, and sample sizes were appropriate and sufficient, and whether the data supported the stated conclusions. To support consistency and emphasize that improving scientific *Quality* was the primary objective of this phase, reviewers received a *Quality Assessment Checklist* (**Supplementary Figure 4**) directing them to evaluate four key areas: 1) whether the conclusions were adequately supported by the data and free of contradictions, 2) the soundness of the experimental design, controls, methodology, and statistical analyses, 3) the integrity and reproducibility of the data, including sample sizes and number of replicates, and 4) the clarity and presentation of the manuscript.

Quality reviewers provided feedback through two complementary mechanisms. First, in-line comments were added directly to the manuscript preprint using the Hypothes.is tool, mirroring how scientists typically provide feedback to colleagues during manuscript preparation. Comments were initially posted in private groups visible only to the editor, then compiled and shared in groups visible to all reviewers and the authors. Comments from reviewers who wished to remain anonymous were de-identified before sharing.

Second, reviewers completed a standardized *Quality* Assessment Form consisting of six short-answer questions and multiple Likert-scale questions (**Supplementary Figure 5A-B**). This form was designed to capture summary-level assessments and test the feasibility and utility of using structured, quantitative metrics to evaluate scientific rigor. All assessment forms were designed, distributed, and collected using the Formaloo platform.

Author Response Phase

After *Quality* reviews were submitted, the editor compiled and shared feedback with authors and other reviewers. In-line comments were shared by inviting authors and reviewers to a shared Hypothes.is group containing all reviewer annotations. *Quality* Assessment Form feedback was compiled in a PDF that included short-answer responses and graphical representations of the Likert-style questions (**Supplementary Figure 5C**). Both in-line comments and assessment form feedback were labeled by reviewer. Reviewers who disclosed their identity were named, while anonymous reviewers were designated “Reviewer #”.

Authors were encouraged, though not required, to reply directly to reviewers in Hypothes.is. Both authors and reviewers were encouraged to use Hypothes.is for open, constructive, and interactive discussions aimed at improving the manuscript’s overall *Quality*. Authors were not expected to conduct additional experiments for the pilot.

Impact Review Phase

Impact was defined as the extent to which a study’s findings advance scientific understanding, fill critical knowledge gaps, influence multiple fields, or have therapeutic relevance. Reviewers were informed that a manuscript may be of high *Quality* yet have

limited *Impact* if presents a modest or highly specialized advance relevant to only a small audience.

To ensure sufficient evaluations for each manuscript, *Quality* reviewers were also asked to provide a separate assessment of the manuscript's potential *Impact*. After completing the *Quality* Assessment Form, reviewers were directed to a separate *Impact* Assessment Form that included a short-answer question asking whether specific strengths or weaknesses identified in the *Quality* Review influenced their perception of *Impact*, a multiple-choice question, and five Likert-scale questions assessing key dimensions of *Impact* such as transformative potential, generalizability, technological advancement, addressing a critical knowledge gap, or providing novel mechanistic insights (**Supplementary Figure 6A-D**). This structured framework enabled *Impact* to be evaluated independently of *Quality* while capturing the diversity of perspectives on scientific significance.

Impact-only Review Phase

Since assessment of *Impact* is inherently subjective, additional reviewers were recruited to focus solely on the *Impact* evaluation. The goal was to obtain a minimum of three *Quality* and six *Impact* reviews per manuscript. *Impact*-only reviewers were recruited after the *Quality* phase was completed and they received the manuscript preprint, *Quality* reviewers in-line comments, compiled *Quality* Assessment Form feedback, and the authors' responses. *Impact*-only reviewers completed the same *Impact* Assessment Form described above (**Supplementary Figure 6A-D**), focusing exclusively on assessing significance rather than scientific rigor.

Reviewer Experience, Training, Expectations, and Compensation

Reviewers were selected based on their subject matter expertise, identified through author recommendations, nominations by members of the scientific advisory board,

previously enrolled participants, or PubMed searches by the editor. Most reviewers participating in the study were principal investigators. Trainees, which included staff scientists, postdoctoral fellows, and graduate students, also completed reviews, but only with prior endorsement from their PI confirming their readiness to participate. Additionally, several reviews were completed collaboratively between principal investigators and trainees as a training exercise. Most reviewers brought substantial experience to the process: 83% previously reviewed 10 or more manuscripts, 49% reported over a decade of experience submitting and reviewing papers, and 80% had been engaged in peer review for at least five years (**Supplementary Figure 2G-H**).

Before receiving review materials, reviewers were invited to attend a brief Zoom onboarding session or receive detailed instructions via email. Most *Quality* reviewers (88%) attended a Zoom onboarding meeting, whereas most *Impact*-only reviewers (90%) received detailed instructions via email. Onboarding covered Hypothes.is setup, pilot goals, and emphasized the value of a peer-improvement mindset, defined as offering clear, constructive, and actionable feedback to improve scientific rigor and suggest additional experiments only when essential to support the conclusions, feasible for the research group, and within the scope of the study. *Quality* reviewers agreed to complete reviews within 14 days and *Impact*-only reviewers within 10 days. Timeliness was communicated during reviewer recruitment, reinforced during onboarding sessions, emphasized in the review instructions, and reiterated in follow-up reminder emails. To acknowledge their contributions and reflect our commitment to a platform that compensates reviewers, *Quality* reviewers were offered $30, and *Impact*-only reviewers were offered $20 per review.

Manuscript Enrollment

Five manuscripts were initially enrolled following email invitations to previously enrolled participants. Subsequently, invitations were sent to authors of recent bioRxiv

manuscripts, resulting in the enrollment of 10 additional manuscripts. Three additional bioRxiv manuscripts were selected by the scientific advisory board and reviewed without author input. In total, 18 manuscripts were reviewed in the Discovery Stack Pilot. One manuscript (DSPM118) was posted on SSRN rather than bioRxiv. Since Hypothes.is did not function well on SSRN, additional reviewers were not recruited for this manuscript, and it was excluded from several analyses.

At the conclusion of the pilot, 11 of 18 authors (61.1%) completed the author feedback survey, which captured information about the submission status of their manuscripts. Of these 11 manuscripts, eight were first-time submissions to a traditional journal, while three were revised drafts following one round of journal revision. As of October 8, 2025, only 6/18 manuscripts reviewed in the pilot had been accepted for publication in a traditional journal.

Development and Validation of Composite *Quality* and *Impact* Metrics

Quality and *Impact* metrics were developed using standardized assessment forms containing Likert-scale questions rating defined attributes of each dimension on a 1-5 scale (1= strongly agree, best; and 5=strongly disagree, worst (**Supplementary Figures 5-6**).

Quality scores were derived from 13 Likert-scale questions addressing experimental design, controls, statistical analysis, reproducibility, and data support for conclusions. Each question was weighted according to its relative importance, and weighted responses were averaged to yield a single composite *Quality* score per review (**Supplementary Figure 7A**). All manuscripts received at least two composite *Quality* scores, most received three. To evaluate the effect of weighting, weighted and unweighted averages were compared to the average score of a key item ("Overall, the data support the conclusions presented in the paper"), which served as a benchmark of

overall rigor. Weighted scores aligned with unweighted averages but trended toward the benchmark (**Supplementary Figure 7C**).

Impact scores were generated from one multiple-choice question and five Likert-scale items. In the multiple-choice question, reviewers selected features contributing to a manuscript's *Impact* (**Supplementary Figure 6 and 7B**). Each feature was weighted by importance, and the weighted sums were normalized to a 1-5 scale and combined with the weighted average of the five Likert-scale responses to generate a single composite *Impact* Score for each review. All manuscripts received at least five composite *Impact* scores, and most (15 of 17) received six or more. Unweighted and weighted averages, with and without the multiple-choice question, were compared and showed similar results (**Supplementary Figure 7D**).

Most manuscripts reviewed in the pilot were initial submission. However, three manuscripts were revised versions after one round of review. Composite *Quality* and *Impact* scores of initial and revised submissions were compared to determine if revised manuscripts should be analyzed separately. No significant differences, or even trends toward higher scores were observed in the revised manuscripts, so all manuscripts were grouped together for further analyses (**Supplementary Figure 7E-F**).

Composite *Quality* and *Impact* Score Analysis

Associations were tested using Pearson's correlation and linear regression models using GraphPad Prism. Data were analyzed either using all available *Quality* and *Impact* scores or using only scores from reviewers who evaluated both *Quality* and *Impact* to ensure any differences observed were not due to different reviewers. To compare variability between *Quality* and *Impact* scores, SD and range were determined in Excel, IQR percentiles were calculated using NumPy's default linear interpolation method, and were compared using the Wilcoxon matched-pairs signed-rank test in

Prism. Manuscript DSPM136 was excluded as an outlier (z-score = -3.14, >3 SD from mean difference). Bootstrap resampling was performed using custom Python scripts to assess the SD, range, and IQR of *Quality* and *Impact* scores. For each of 2,000 bootstrap iterations, reviewer scores were resampled with replacement for each manuscript under two conditions: Unmatched resampled all available scores, while matched resamples an equal number of scores per manuscript, equal to the smaller of the two counts for that manuscript. For each resampled dataset, we calculated SD, range, and IQR for *Impact* and *Quality* scores, computed the difference (*Impact* − *Quality*) for each metric, and averaged these differences across manuscripts. This yielded a bootstrap distribution of differences for each variability metric and sampling condition. Mean differences, 95% percentile confidence intervals, and two-sided p-values were estimated from these distributions.

Survey Design and Analysis

At the completion of the pilot study, surveys were distributed to 93 individuals who participated as authors, *Quality/Impact* reviewers, or *Impact*-only reviewers. Surveys were generated and distributed using the Formaloo platform. Three distinct surveys were developed, each tailored to the participant's role. Individuals who served in multiple roles received a separate survey for each role. In total, 101 survey invitations were sent, and 86 completed responses were received, yielding an 85% completion rate.

Questions used a mix of Likert-scale, multiple-choice, and open-ended formats. Quantitative responses were summarized as percentages of total respondents per question. Open-ended responses were coded thematically, and frequencies were tallied to identify common themes. For ranking questions, responses marked "very important" or "absolutely essential" were combined to calculate the proportion of participants prioritizing each feature.

Statistical analyses were performed in GraphPad Prism using standard functions for correlation, range, and dispersion calculations. Figures were generated in Prism and assembled in PowerPoint.

FIGURE LEGENDS

Figure 1. The Discovery Stack Model and Manuscript Enrollment. (**A**) The Discovery Stack model includes three sequential phases: 1) *Quality* Review, 2) Author Response, and 3) *Impact* Review. (**B**) Eighteen manuscripts were reviewed in the pilot. Manuscripts are grouped according to method of enrollment with the number of completed *Quality* and *Impact* reviews for each manuscript depicted. (**C**) Number of *Quality* and *Impact* reviews completed per manuscript. Manuscript DSPM118 was posted on SSRN rather than bioRxiv and excluded from this analysis due to compatibility issues between SSRN and Hypothes.is.

Figure 2. Effective Separation of *Quality* and *Impact* with Strong Support for Standardized Metrics. (**A**) Composite *Quality* and *Impact* scores arranged by impact factor of the submission journal. The submission journal for one manuscript (DSPL165) was unknown, so this manuscript was excluded from analyses in A-F. (**B**) The relationship between composite *Quality* and *Impact* scores for each manuscript was tested using Pearson's correlation and linear regression, restricting analysis to reviewers who completed both assessments. The shaded region shows the 95% confidence interval (CI) of the regression line. (**C**) Residuals from the linear regression model were calculated to assess variation in *Impact* not explained by *Quality*. The shaded region represents ±1 standard deviation (SD = 0.41). (**D**) Manuscripts were grouped into tertiles based on *Quality* scores and *Impact* scores were compared across tertiles using Kruskal–Wallis tests (H = 9.46, p=0.0032), followed by Dunn's multiple comparisons. (**E-F**) Correlations between *Quality* (**E**) or *Impact* (**F**) scores and journal impact factor were evaluated using Pearson's correlation. (**G-J**) Survey data from *Quality/Impact* reviewers (n=42), *Impact*-only reviewers (n=33), and authors (n=11) were tallied, and the percentage of each group selecting each response is shown. (**K-M**) Variability between composite *Quality* and Impact scores, within each manuscript, was

quantified using (**K**) standard deviation (SD), (**L**) range, and (**M**) interquartile range (IQR), and compared using Wilcoxon matched-pairs signed-rank test. $p<0.05$*; $p<0.001$***. (**N**) Bootstrap resampling was performed to compare variability between *Impact* and *Quality* scores using SD, range, and IQR under matched and unmatched sampling conditions. SD distributions are shown. Mean differences were 0.23 (SD), 0.74 (range), and 0.37 (IQR) for unmatched sampling (all $p < 0.01$), and 0.17 (SD), 0.29 (range), and 0.15 (IQR) for matched sampling (all $p < 0.05$). Manuscript DSPM136 was excluded as an outlier from analyses in K-M (z-score = -3.14, >3 SD from mean difference) and DSPM118 was excluded as additional *Impact* reviewers were not recruited due to Hypothes.is compatibility issues.

Figure 3. High Satisfaction with the Discovery Stack Platform. (**A**) Participants rated their overall experience with the Discovery Stack platform compared to traditional review. The percent of authors (n=11), *Quality/Impact* reviewers (n=42), and *Impac*t-only reviewers (n=33) selecting each rating is shown. (**B**) Authors (n=11) and *Quality/Impact* reviewers (n=42) were asked whether the Discovery Stack platform fostered a greater "Peer-Improvement" mindset than traditional peer review. (**C**) Participants were asked in an open-ended question what aspects of the Discovery Stack platform they found most beneficial compared to traditional review. A total of 74 responses were collected, categorized into key thematic areas, and the percent of participants (n=86) who mentioned each theme is shown. All responses are shown. (**D**) Participants identified aspects of the platform they found most challenging. A total of 56 open-ended responses were collected, categorized into thematic categories, and the percentage of participants (n=86) mentioning each challenge is shown. All responses are shown.

Figure 4. In-line Commenting Improves Review Clarity, Collegiality, and Efficiency. (**A**) Authors (n=8-9) and *Impact*-only reviewers (n=32) were asked whether

they were able to read the *Quality* reviewers' in-line Hypothes.is comments without difficulty, and whether they found the comments more constructive and helpful than those typically received through traditional review. (**B**) Authors (n=8-9) were asked whether responding to reviewers' in-line comments would be easier than addressing traditional review comments, and whether the in-line format helped clarify which aspects of the paper needed improvement. (**C**) *Quality/Impact* reviewers (n=40-42) were asked a series of questions focused on providing in-line comments using the Hypothes.is platform.

Figure 5. Discovery Stack Enables Rigorous *Quality* Review and Supports Meaningful *Impact* Assessment. (**A**) Authors (n=8-11) were asked to rate their level of agreement with statements related to the feedback they received during the *Quality* review. (**B**) Authors (n=11) were asked to rate the overall usefulness of feedback they received from the Discovery Stack review compared to traditional peer review. (**C**) *Quality/Impact* (n=42) and *Impact*-only (n=32) reviewers responded to questions focused on their experience with the *Impact* assessment. (**D**) *Impact*-only reviewers (n=32) were asked whether the *Quality* reviewers feedback from the assessment form and in-line comments aided their assessment of manuscript *Impact*.

Figure 6. Expedited Review Is Achievable with Reviewer Accountability and Backup Plans. (**A**) For manuscripts still under review at the time of analysis (10/8/25), the time from author-reported submission to analysis is plotted. For published manuscripts, time from submission to publication reported by the journal is shown. The submission dates for manuscripts DSPL178 and DSPL165 are unknown, so they are excluded from this analysis. (**B**) Days required to recruit three *Quality* reviewers, complete *Quality* reviews, recruit three additional *Impact*-only reviewers, and complete *Impact* reviews is shown for each manuscript. DSPM118 was excluded from analyses in B-D as additional *Impact* reviewers were not recruited due to Hypothes.is compatibility

issues. (**C**) Total duration of the *Quality* phase (including recruitment and review completion), the *Impact* phase, and combined review time for each manuscript. (**D**) Average recruitment and review times per manuscript were analyzed with Kruskal-Wallis and Dunn's multiple comparison post-test. $p<0.05$*; $p<0.01$**. (**E**) Authors (n=11) reported the duration of initial journal review, which was compared to the *Quality* review duration using Wilcoxon matched-pairs signed-rank test. (**F**) Authors (n=11) reported whether they received Discovery Stack *Quality* reviews faster than the initial traditional review. (**G**) Individual *Quality* or *Impact* review times were compared using the Mann-Whitney test ($p=0.0002$***). (**H**) Individual *Quality* Review times were plotted by manuscript. DSPM118 was excluded. (**I**) *Quality/Impact* reviewers (n=40) estimated the time to complete the review, including reading the pre-print, commenting in Hypothes.is, completing both the *Quality* and *Impact* Assessment Forms. *Impact*-only reviewers (n=33) estimated the time to read the pre-print along with *Quality* reviewers' feedback and complete the *Impact* Assessment Form. (**J**) *Quality/Impact* reviewers (n=42) rated the efficiency of Discovery Stack review, considering both the time spent and value it provided to authors, compared to traditional peer review.

Figure 7. Strong Interest in a New Scientist-Driven Publishing Platform. (**A**) Participants (n=85) were asked whether they would be interested in using a new platform that utilizes scientist-developed metrics to evaluate and curate peer-reviewed research, and whether they would submit an original research paper to a new platform that utilizes scientist-developed metrics to assess *Quality* and *Impact*. (**B**) Participants (n=85) were asked to rank the importance of features that would influence their likelihood of using a novel platform for evaluating and curating peer-reviewed and peer-approved research. Features were rated from "absolutely essential" to "not important". The percent of respondents selecting each rank is shown.

Supplementary Figure 1. Reviewer Identity Disclosure Enhances Perceived Transparency but Influences *Impact* Scores. (**A**) The percent of *Quality* (n=50) or *Impact*-only (n=114) reviewers who disclosed their identity or remained anonymous. (**B**) The percent of PIs (n=129) and trainees (n=33) who disclosed their identity or remained anonymous. (**C**) The percent of identified reviewers per manuscript. (**D**) Reviewers who remained anonymous (n=35) were asked a follow-up multiple-choice question regarding the reasons that influenced their choice. (**E**) Authors (n=11) were asked if they perceived the Discovery Stack review process to be more transparent than traditional review. (**F**) Transparency ratings for authors that completed surveys (n=11) were compared to percent of identified authors for each manuscript using Pearson's correlation. (**G**) Authors (n=10) and *Impact*-only reviewers (n=18) were asked whether they noticed a difference in scoring between identified and anonymous reviewers (**H**) Authors (n=10) were asked follow-up questions regarding their perception of feedback from identified and anonymous reviewers. (**I, K**) Composite (**I**) *Quality* and (**K**) *Impact* scores were compared between identified and anonymous reviewers across all manuscripts using the Mann-Whitney test (p=0.0082** for *Impact*). (**J, L**) Within each manuscript, the average (**J**) *Quality* and (**L**) *Impact* scores from identified and anonymous reviewers were compared using the Wilcoxon matched-pairs signed rank test (p=0.0258* for *Impact*). Manuscripts with no identified reviewers (*Quality* n=2, *Impact* n=1) or no anonymous reviewers (*Quality* n=3; *Impact* n=3) were excluded. (**M**) Composite *Quality* and *Impact* scores were compared between PIs (*Quality* n=37; *Impact* n=92) and trainees (*Quality* n=12; *Impact* n=21). Differences were tested using the Mann-Whitney test.

Supplementary Figure 2. Reviewer Engagement Depends on Outreach, Networks, and Trainee Involvement. (**A**) The positive response rate was compared between individuals who were already familiar with the study (DSP-enrolled), new to the study

(New to DSP), or new but contacted by a known colleague on the Scientific Advisory Board (New to DSP + email) using two-sided Fisher's exact tests (p<0.0001). Odds ratios (ORs) and 95% CIs were calculated from the contingency table. P-values were adjusted for multiple pairwise comparisons using the Bonferroni correction. Error bars represent 95% CIs. DSP-enrolled vs. New to DSP (OR=15.2, 95% CI: 7.1–32.7, p<0.0001, adjusted), New to DSP + email vs. New to DSP vs. (OR=12.2, 95% CI: 5.8–25.7, p<0.0001, adjusted), DSP-enrolled vs. New to DSP + email significant (OR=1.25, 95% CI: 0.64–2.42, p=0.61, adjusted). (**B**) Number of reviewer invitations per manuscript to recruit *Quality/Impact* and *Impact*-only reviewers. (**C**) The percent of individuals that responded "yes", "no", or did not respond to reviewer invitations per manuscript. The overall acceptance was 32.5%. (**D**) Survey respondents indicated whether they were unfamiliar with the study, previously enrolled, or recruited by a colleague. (**E**) Percent of *Quality* (n=48) or *Impact* (n=110) reviewers at each career stage. Among *Quality* reviewers, 73% were Principal Investigators (PIs) while 80% of *Impact* reviewers were PIs. Only independent trainee reviewers (i.e., not co-reviewing with a PI) were included in these percentages. Other includes industry positions. (**F**) A multivariable logistic regression model estimated the odds of reviewer acceptance as a function of position (PI vs. trainee) and participant type (DSP-enrolled, New to DSP, New to DSP + email) as predictors. Only independent trainee reviewers, not co-reviewers, were included in the analysis. OR and 95% CI were obtained by exponentiating the logistic regression coefficients. Statistical significance was assessed using Wald tests (OR=15.3, 95% CI: 4.2–56.2, p<0.0001). (**G**) Survey participants reported their review experience. (**H**) The number of years of peer review experience reported by participants.

Supplementary Figure 3. Top Priorities for Improved Scientific Publishing (**A**) Participants (n=86) were asked what they believe needs to change about the current

scientific publishing and peer review systems. A total of 65 open-ended responses were collected, categorized into key thematic areas, and tallied. The number of respondents mentioning each concern is shown. (**B**) Participants were also asked to imagine that scientific publishing didn't exist and to identify three core principles they would prioritize if building a system from scratch. A total of 53 open-ended responses were collected, categorized by theme, and tallied. (**C**) Responses across both questions were averaged and ranked to highlight the most broadly recognized priorities for reform.

Supplementary Figure 4. *Quality* Assessment Checklist. The *Quality* Assessment checklist was given to *Quality/Impact* reviewers to highlight key elements to consider when evaluating the *Quality* of a manuscript. It offers a comprehensive framework for reviewing conclusions, assessing experimental design, ensuring data integrity, and evaluating clarity and scholarly analysis. This guide contains resources adapted from (19) https://doi.org/10.5281/zenodo.5484087.

Supplementary Figure 5. *Quality* Review Assessment Form. The *Quality* Review Assessment Form included six short-answer questions, followed by a set of Likert-scale items designed to evaluate specific attributes of manuscript *Quality*. These questions focused on the rigor and reproducibility of the data.

Supplementary Figure 6. *Impact* Review Assessment Form. The *Impact* Review Assessment Form included: 1) a short-answer question asking whether specific strengths or weaknesses identified in the *Quality* Review influenced their perception of the manuscript's *Impact*, 2) a multiple-choice question allowing reviewers to select features they believed contributed to the manuscript's *Impact*, and 3) five Likert-scale questions assessing key dimensions of the manuscript's *Impact*.

Supplementary Figure 7. Calculation and Comparison of *Quality* and *Impact* Composite Scores. Composite *Quality* and *Impact* review scores were calculated to

condense the full set of ratings from each review into a single metric. (**A**) Composite *Quality* scores were derived from 13 Likert-scale questions that were weighted based on relative importance. Composite *Quality* score = ∑ (Question Score x Importance weight)/ ∑ (Importance weight). (**B**) Composite *Impact* scores were based on responses to one multiple-choice question and five Likert-scale items. To integrate responses across two question types, each feature in the multiple-choice question was assigned a weight reflecting its relative importance to *Impact*, the weighted sum of selected features was calculated and then normalized to a 1-5 scale using a linear transformation. Transformed Q2 = 6 - (1 + ((∑Q2-1)*4/(36-1))). In this scale, selecting no features corresponds to a score of 5 (lowest *Impact*) and selecting features totaling the maximum score achieved (36) corresponds to a score of 1 (highest *Impact*). This transformed Q2 value was then combined with the weighted average of the five Likert-scale responses to generate a single composite *Impact* score for each review. (**C**) The weighted *Quality* composite score, unweighted average of Likert-scale responses, and the average score for the first statement (Overall, the data support the conclusions presented in the paper) (**D**) The weighted composite *Impact* score (including Q2), the unweighted average of Q2-Q7, and the unweighted average of Q3-Q7 were compared across manuscripts. (**E-F**) Weighted composite scores for (**E**) *Quality* or (**F**) *Impact* were compared between manuscripts enrolled as initial submissions versus revised versions. The version of three manuscripts (DSPL165, DSPL173, DSPL178) was unknown, so are excluded. No significant differences were observed by Mann-Whitney test.

AUTHOR CONTRIBUTIONS

MAM: Conception of methods, coordination of study, data analysis and presentation, writing and editing of manuscript.

BCL: Conception of study, ongoing coordination during the study, initial setup of study.

KC, CR, MK, MK, BR, SS, AG, LBR, SR, TS, IR, JG, JO-M, SS, DM, AO, IB, HGV, JS, TF, NJ: Conception of study, ongoing coordination during the study.

MFK: Conception of study, ongoing coordination during the study, writing and editing of manuscript.

ACKNOWLEDGEMENTS

We thank Solving For Science for funding and support associated with undertaking this study. We thank the NIH for respective funding for all the labs involved in this study. We thank Igor Brodsky, Julien Gaillard, Ananda Goldrath, Nikhil Joshi, Savan Ram, Carla Rothlin, Sunny Shin, Sara Suliman, and Joe Sun for helpful advice during the planning, implementation, and analysis of the pilot.

REFERENCES

1. A. Ghasemi, P. Mirmiran, K. Kashfi, Z. Bahadoran, Scientific Publishing in Biomedicine: A Brief History of Scientific Journals. *Int. J. Endocrinol. Metab.* **21**, e131812 (2023).

2. T. Jefferson, M. Rudin, S. Brodney Folse, F. Davidoff, Editorial peer review for improving the quality of reports of biomedical studies. *Cochrane Database Syst. Rev.* **2007**, MR000016 (2007).

3. R. Smith, Peer review: a flawed process at the heart of science and journals. *J. R. Soc. Med.* **99**, 178–182 (2006).

4. C. Superchi, *et al.*, Development of ARCADIA: a tool for assessing the quality of peer-review reports in biomedical research. *BMJ Open* **10**, e035604 (2020).

5. C. Superchi, *et al.*, Tools used to assess the quality of peer review reports: a methodological systematic review. *BMC Med. Res. Methodol.* **19**, 48 (2019).

6. J. P. Tennant, T. Ross-Hellauer, The limitations to our understanding of peer review. *Res. Integr. Peer Rev.* **5**, 6 (2020).

7. L. Jancovich, C. Pitches, D. Stevenson, Failures in impact evaluation. *Res. Eval.* **34**, rvaf033 (2025).

8. M. Pagliaro, Publishing scientific articles in the digital era. *Open Sci. J.* **5** (2020).

9. M. A. Hanson, P. G. Barreiro, P. Crosetto, D. Brockington, The strain on scientific publishing. *Quant. Sci. Stud.* **5**, 823–843 (2024).

10. J. Huisman, J. Smits, Duration and quality of the peer review process: the author's perspective. *Scientometrics* **113**, 633–650 (2017).

11. L. Nelson, *et al.*, Robustness of evidence reported in preprints during peer review. *Lancet Glob. Health* **10**, e1684–e1687 (2022).

12. A. M. Lyons-Warren, W. W. Aamodt, K. M. Pieper, R. E. Strowd, A structured, journal-led peer-review mentoring program enhances peer review training. *Res. Integr. Peer Rev.* **9**, 3 (2024).

13. R. Sever, *et al.*, bioRxiv: the preprint server for biology. [Preprint] (2019). Available at: https://www.biorxiv.org/content/10.1101/833400v1 [Accessed 11 September 2025].

14. M. Krummel, *et al.*, Universal Principled Review: A Community-Driven Method to Improve Peer Review. *Cell* **179**, 1441–1445 (2019).

15. A. Dance, Stop the peer-review treadmill. I want to get off. *Nature* **614**, 581–583 (2023).

16. S. Helmer, D. B. Blumenthal, K. Paschen, What is meaningful research and how should we measure it? *Scientometrics* **125**, 153–169 (2020).

17. D. A. Gorelick, A. Clark, Fast & Fair peer review: a pilot study demonstrating feasibility of rapid, high-quality peer review in a biology journal. [Preprint] (2025). Available at: https://www.biorxiv.org/content/10.1101/2025.03.18.644032v1 [Accessed 20 October 2025].

18. A. Rushforth, Beyond impact factors? Lessons from the Dutch attempt to transform academic research assessment. *Res. Eval.* **34**, rvaf035 (2025).

19. A. Foster, S. Hindle, K. M. Murphy, D. Saderi, Open Reviewers Reviewer Guide. (2021).

Figure 1. The Discovery Stack Model and Manuscript Enrollment

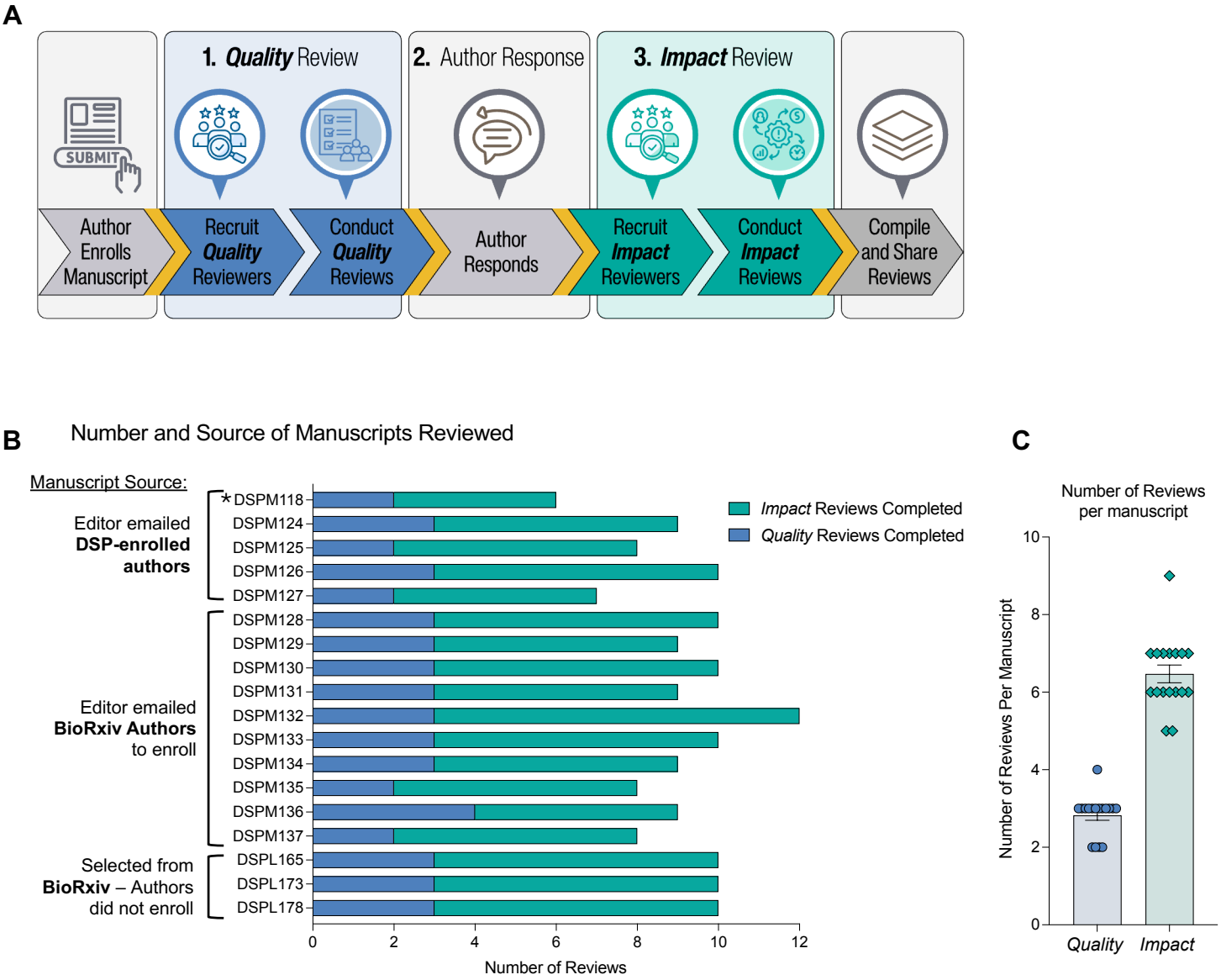


Figure 2. Effective Separation of *Quality* and *Impact* with Strong Support for Standardized Metrics

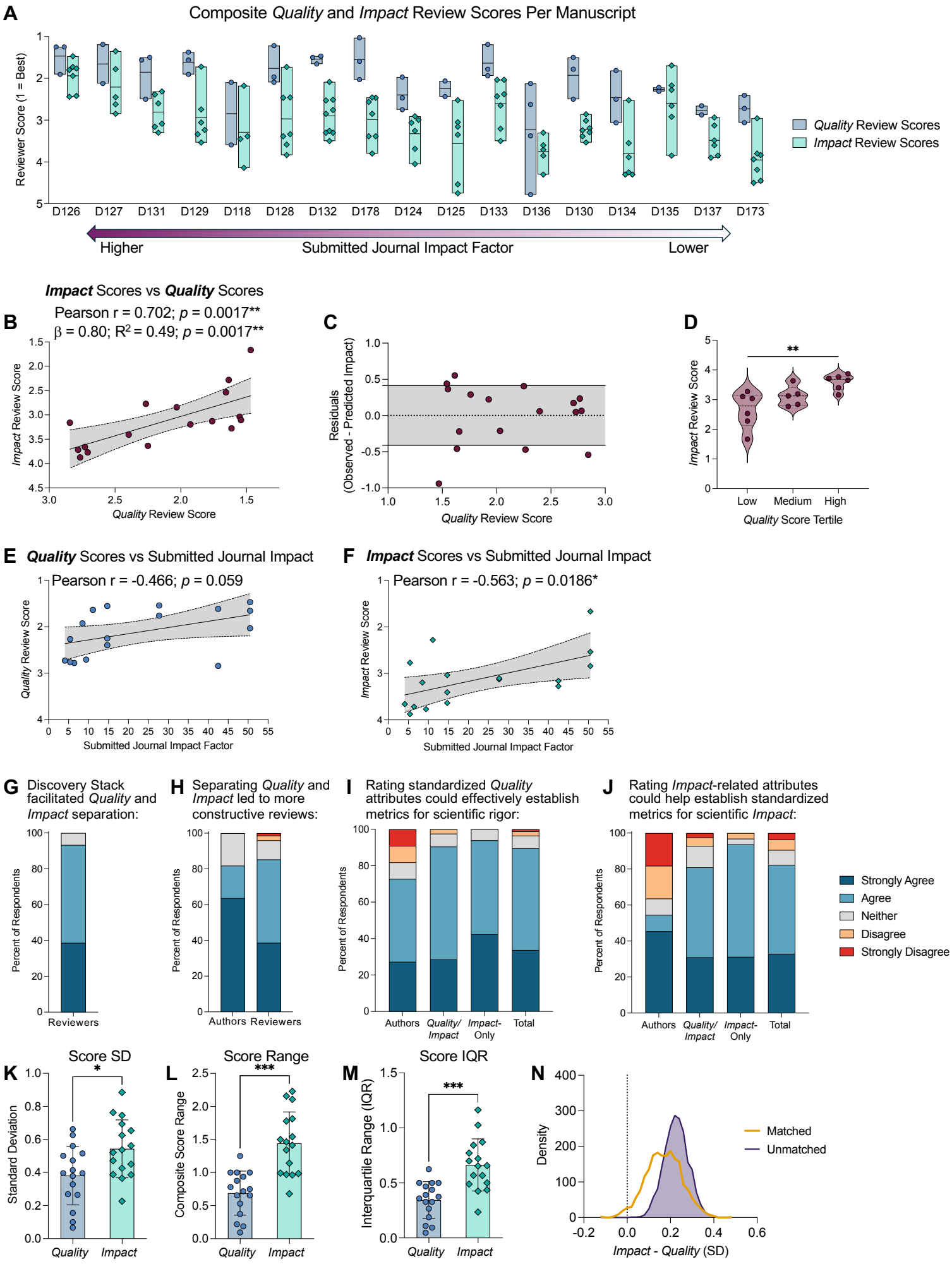


Figure 3. High Satisfaction with the Discovery Stack Platform

Figure 4. In-line Commenting Improves Review Clarity, Collegiality, and Efficiency

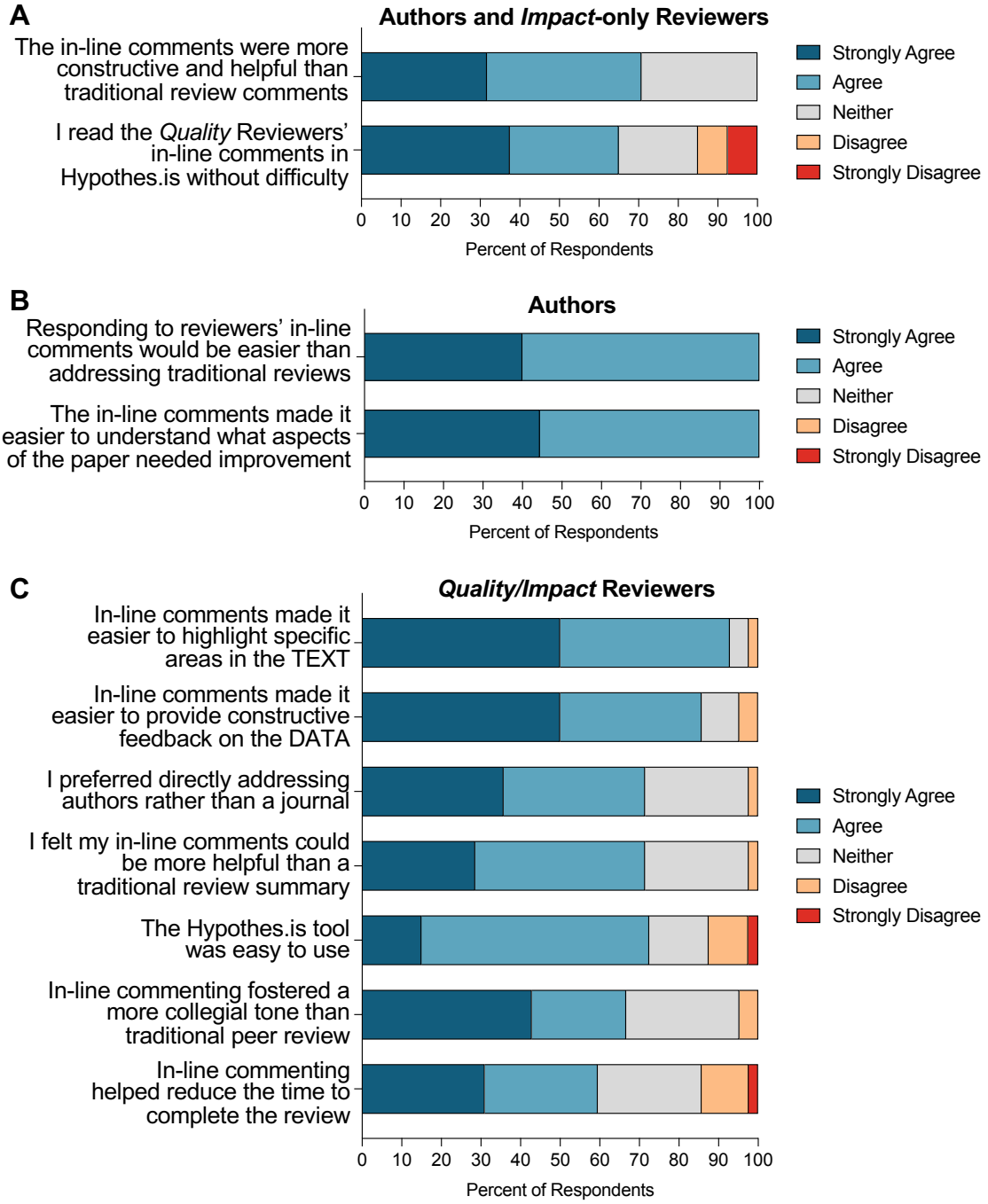


Figure 5. Discovery Stack Enables Rigorous *Quality* Review and Supports Meaningful *Impact* Assessment

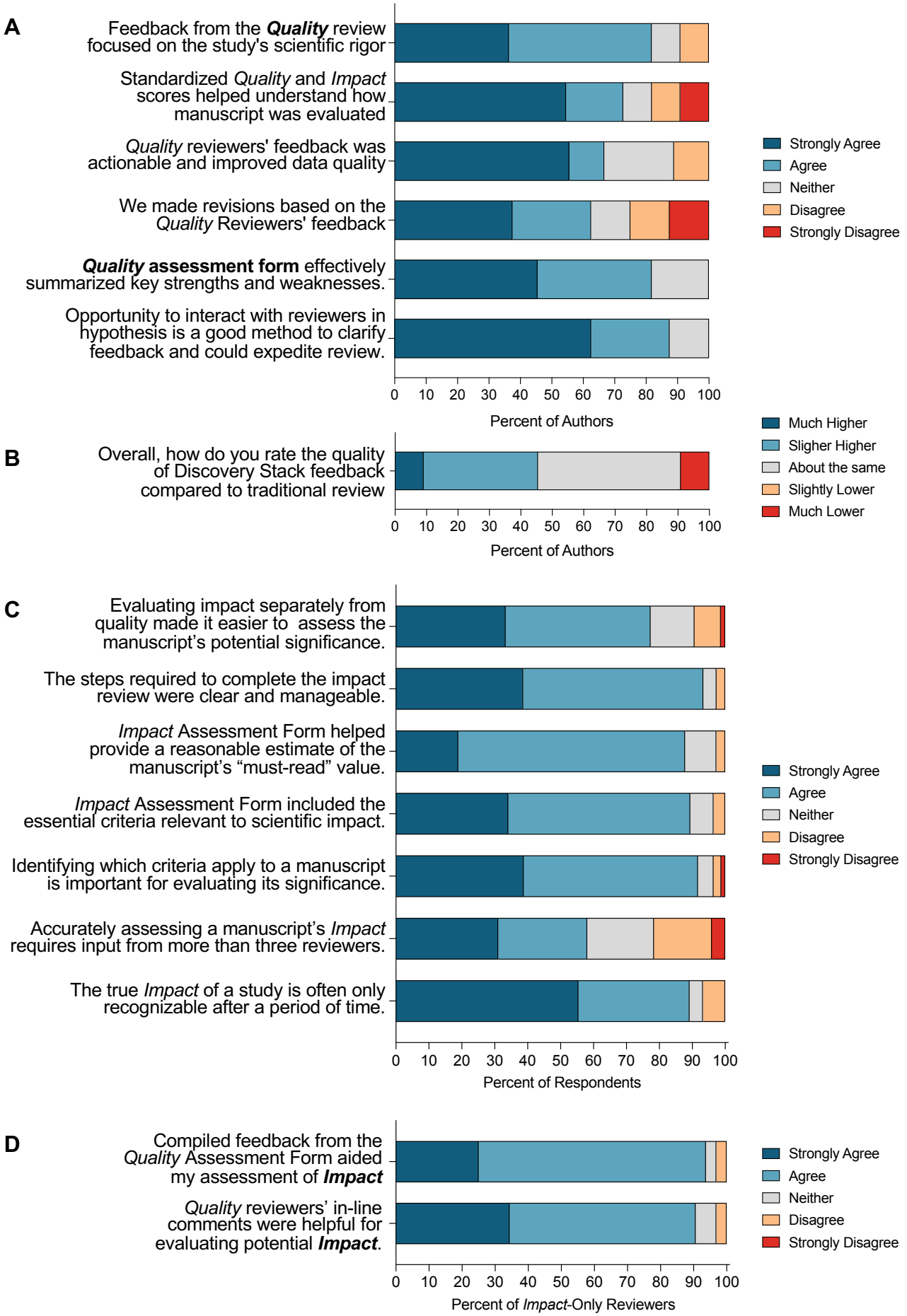


Figure 6. Expedited Review Is Achievable with Reviewer Accountability and Backup Plans

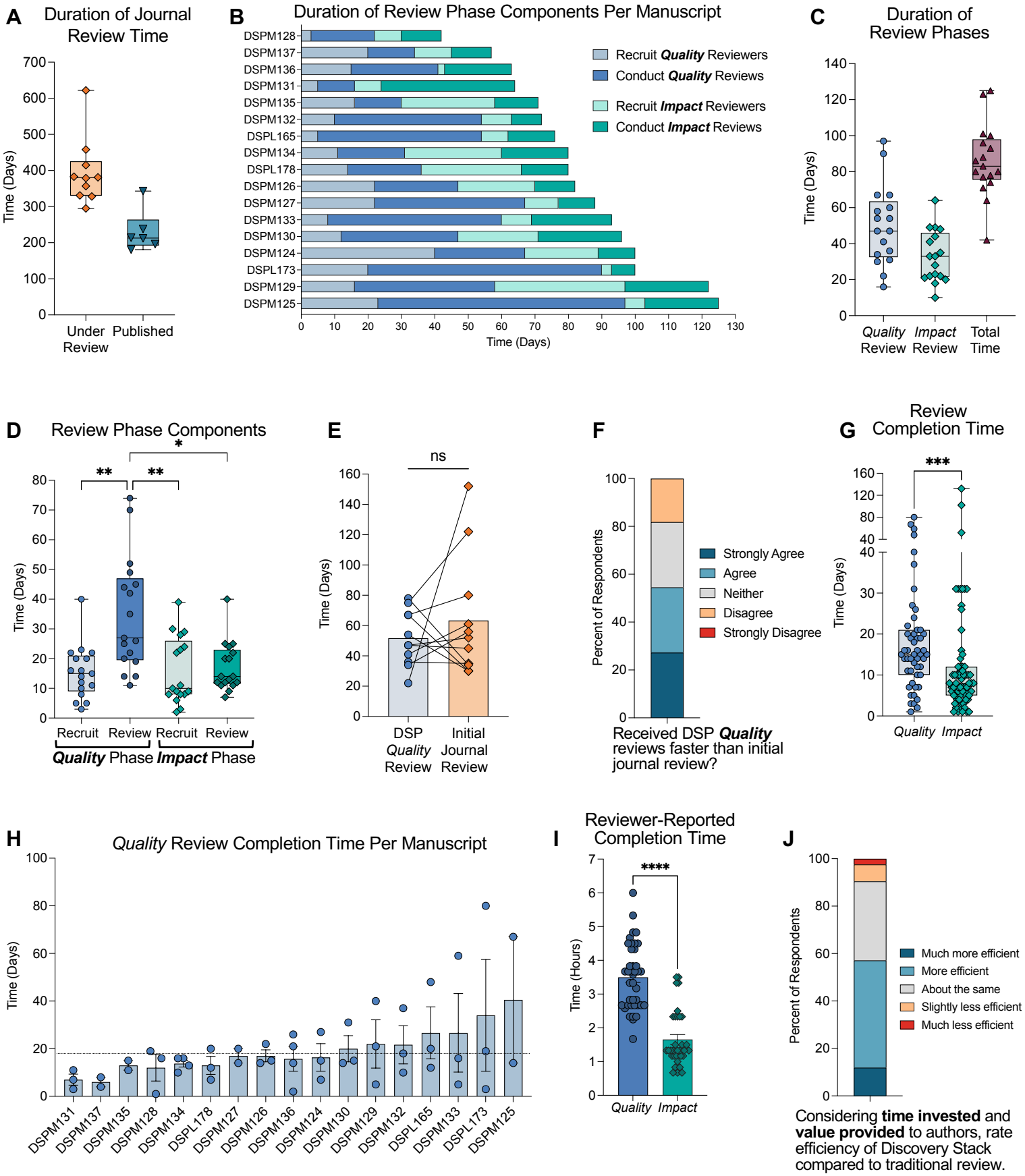
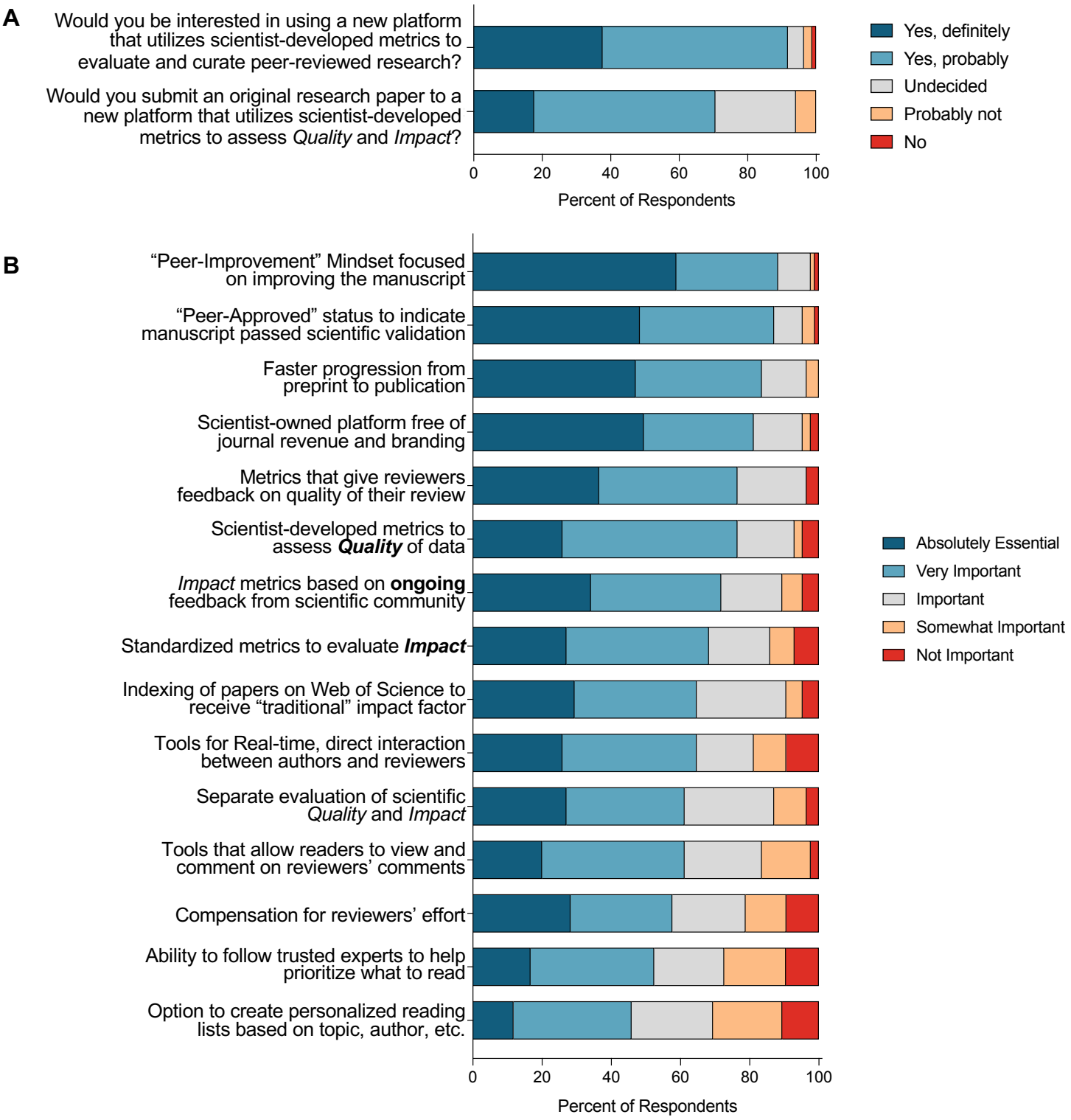
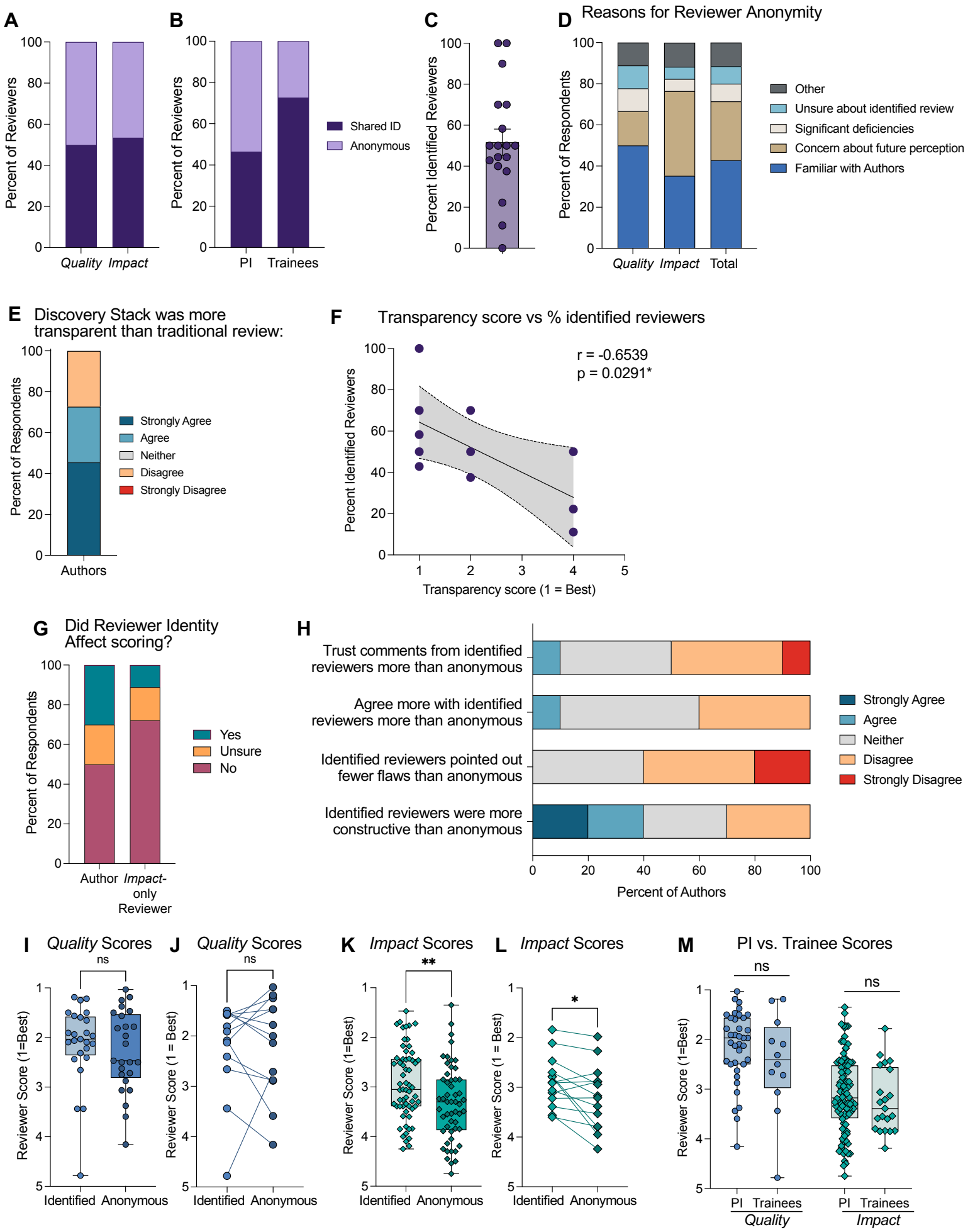


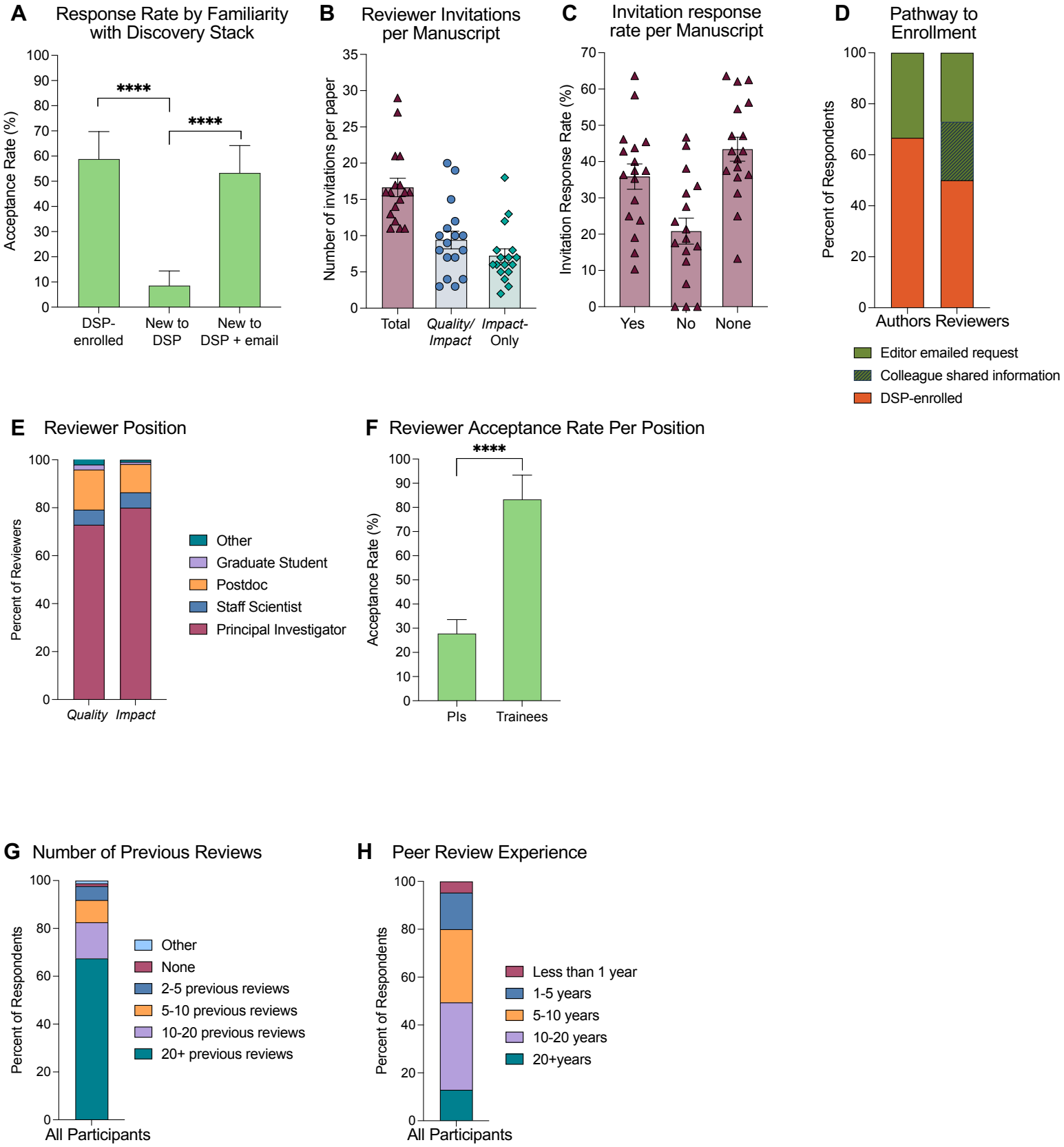
Figure 7. Strong Interest in a New Scientist-Driven Publishing Platform



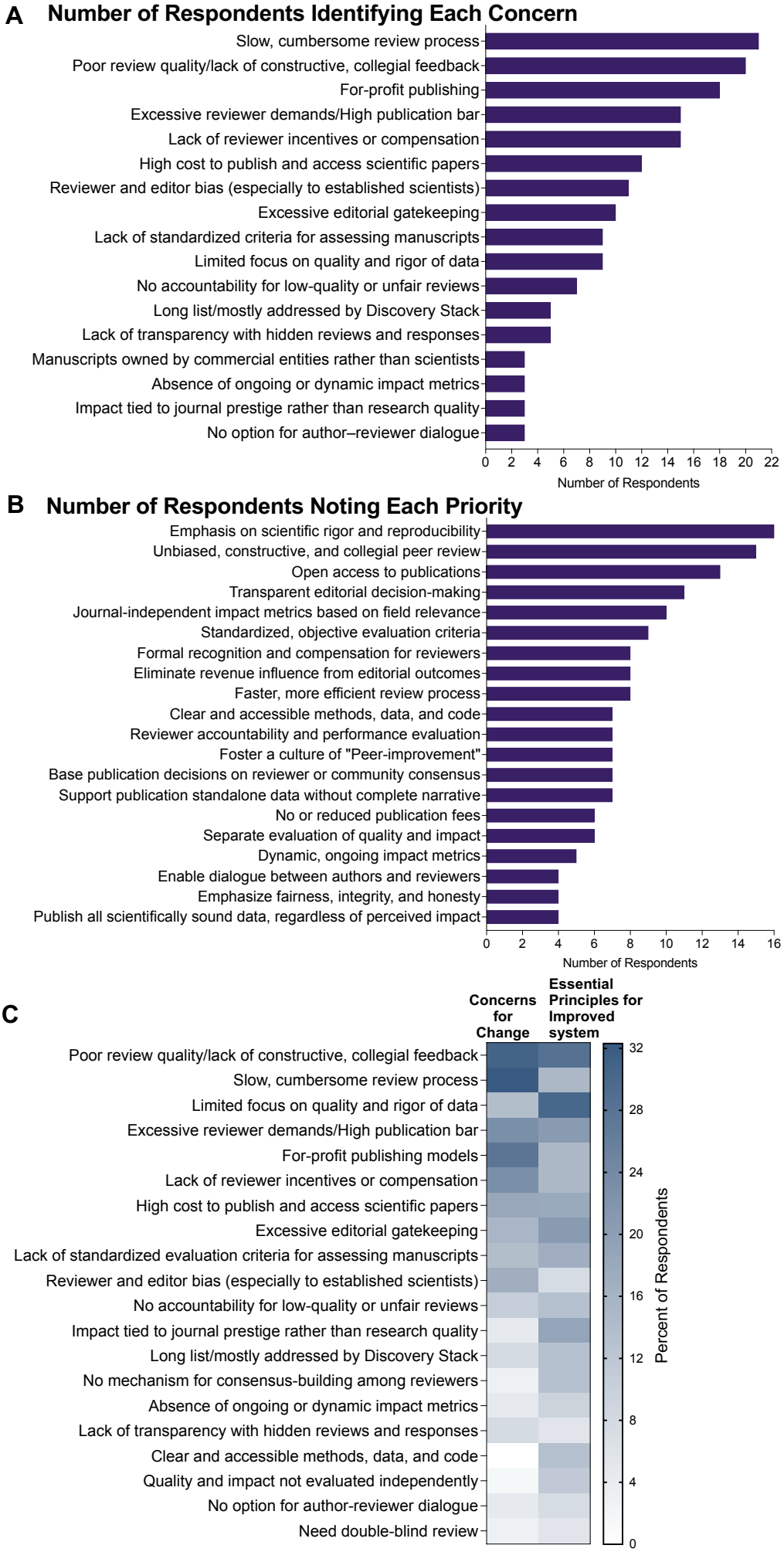
Supplementary Figure 1. Reviewer Identity Disclosure Enhances Perceived Transparency but Affects *Impact* Scores



Supplementary Figure 2. Reviewer Engagement Depends on Outreach, Networks, and Trainee Involvement



Supplementary Figure 3. Top Priorities for Improved Scientific Publishing



Supplementary Figure 4. *Quality* Assessment Checklist

Quality Assessment Checklist

Use hypothes.is to denote the following ON the manuscript PDF:

h. Evaluate Conclusions

- Identify conclusions that are not adequately supported by the data.
- Highlight any contradictory conclusions presented within the manuscript.
- Point out if the authors overlooked significant confounding variables that may impact the conclusions.
- Suggest how conclusions could be clarified so they are supported by the data, or indicate experiments required to support the conclusions.

h. Assess Experimental Design

- For each figure, indicate experiments with improper controls.
- Note if controls are inconsistent with established findings in the field.
- Annotate use of inappropriate methodology.
- Mark incorrect statistical analyses.

h. Review Data Integrity

- Identify experiments not repeated a standard of three times.
- Indicate if the sample size or number of subjects per group is insufficient.
- Note if experimental methods are inadequately outlined to permit reproduction.
- Denote poor data quality.

h. Consider Clarity and Scholarly Analysis

- Indicate if the authors sufficiently discuss how the findings are consistent with or extend upon the field's knowledge.
- Point out if the authors cited and discussed alternative models that could fit with the data.
- Highlight ineffective data presentation.
- Note experiments that require technical clarification.
- Indicate missing or inappropriate references.
- Identify typos, spelling, grammar, readability concerns.

Typically, concerns with **conclusions**, **experimental design**, and **data integrity** are considered **major** issues significantly impacting the quality of a manuscript. While issues with clarity are still important but typically will not affect the overall conclusions.

This guide contains resources adapted from: Foster, A., Hindle, S., Murphy, K. M., & Saderi, D. (2021). Open Reviewers Reviewer Guide. Zenodo. https://doi.org/10.5281/zenodo.5484087

Supplementary Figure 5. *Quality* Review Assessment

A Short-Answer Questions

Q1 – Provide a brief summary of **YOUR** conclusions based on the data presented.

Q2 – What are notable strengths of the manuscript?

Q3 – What are major weaknesses that impact the Quality of the manuscript?
(*Examples of major weaknesses include: Conclusions not supported by the data, contradictory conclusions, inappropriate methodology, or unsuitable statistical analysis, etc.*)

Q4 – What revisions would be required for the data to support the authors' conclusions?

Q5 – the manuscript contains conclusions not supported by the data, how could the authors alter the STATED CONCLUSIONS to better fit the data?

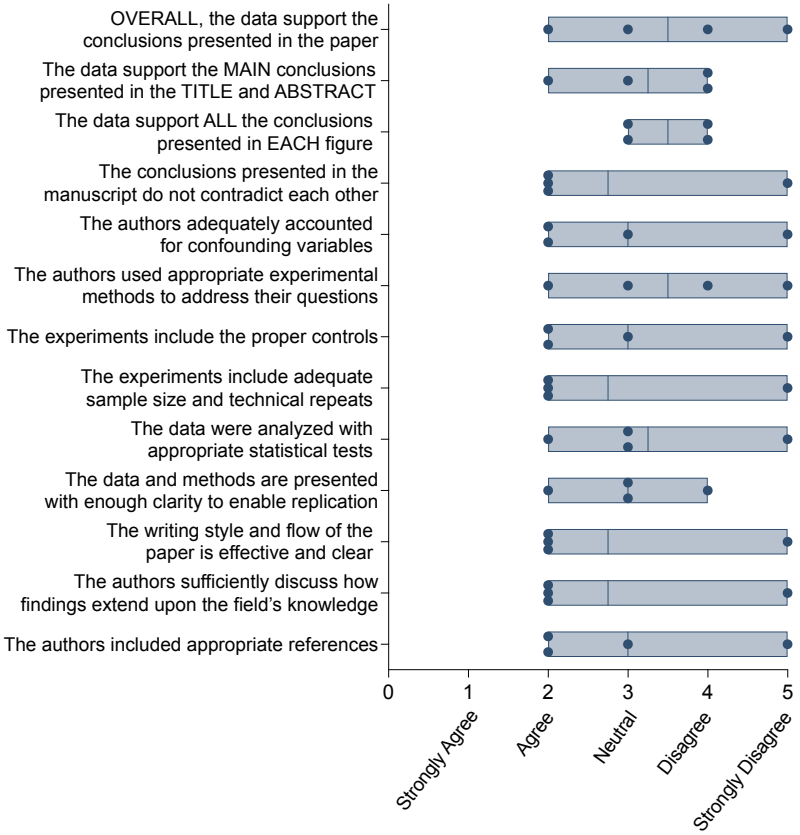
Q6 – If the manuscript contains conclusions not supported by the data, what KEY EXPERIMENTS or data are required to support the authors' conclusions?

B Likert-Scale Questions

Q7 – **Rank your level of agreement with the following statements:**
1. *Strongly Agree*; **2.** *Agree*; **3.** *Neither*; **4.** *Disagree*; **5.** *Strongly Disagree*

- OVERALL, the data support the conclusions presented in the paper.
- The data support the MAIN conclusions presented in the TITLE and ABSTRACT.
- The data support ALL the conclusions presented in EACH figure.
- The conclusions presented in the manuscript do not contradict each other.
- The authors adequately accounted for confounding variables.
- The authors used appropriate experimental methods to address their questions.
- The experiments include the proper controls.
- The experiments include adequate sample size and technical repeats.
- The data were analyzed with appropriate statistical tests.
- The data and methods are presented with enough clarity to enable replication.
- The writing style and flow of the paper is effective and clear.
- Authors sufficiently discuss how findings are consistent with or extend field's knowledge.
- The authors included appropriate references.

C Responses to Likert-Scale Questions for a manuscript



Supplementary Figure 6. *Impact* Review Assessment

A Short-Answer Questions

Q1 – Are there particular strengths and weaknesses mentioned in the Quality Review that contribute to the perceived Impact in the current form?

Q2 – What features contribute to the Impact of this manuscript? Select **ALL** that apply

- **Transformative impact:** This work fundamentally alters scientific understanding, with wide-reaching implications.
- **Significant technological advance:** The work introduces a major technological innovation that may benefit one or more fields.
- **Valuable data resource:** The manuscript provides an important data resource that may benefit current or future researchers.
- **Critical knowledge gap filled**: The work fills a crucial gap in knowledge.
- **Clinical and therapeutic relevance**: The manuscript offers concrete clinical or therapeutic insights for one or more diseases.
- **Novel mechanistic insights:** The findings include new mechanistic data, offering clarity in cellular or molecular mechanisms.
- **Addresses an urgent need**: The work tackles an unmet and urgent need.
- **None of the above**

B Likert-Scale Questions

Q3 – How likely are you to recommend this manuscript to your colleagues?

1. **Extremely enthusiastic:** A must-read paper with groundbreaking findings.
2. **Very enthusiastic:** A thought-provoking paper that offers valuable insights.
3. **Somewhat enthusiastic:** Good, interesting finding relevant to specialized audience.
4. **Not enthusiastic:** The significance of the findings is unclear.

Q4 – Who do you think may be interested in the reported findings?

1. **A very large audience:** General public and researchers across several diverse fields.
2. **A large audience:** Wide range of researchers in multiple, related fields.
3. **A moderate audience:** Researchers in a particular field.
4. **A small audience:** Niche group of researchers, representing a subset within a field.
5. **A very small audience:** Highly specialized researchers in a very specific area.

Q5 – Indicate the level of technological advance presented in the manuscript.

1. **Revolutionary advancement:** Groundbreaking innovation benefiting multiple fields.
2. **Highly advanced:** A significant advance enhancing capabilities within the field.
3. **Moderately advanced:** Meaningful enhancements valuable within a specific area.
4. **Slightly advanced:** Incremental improvements with limited impact outside a niche area.
5. **No advancement:** Similar to existing solutions.

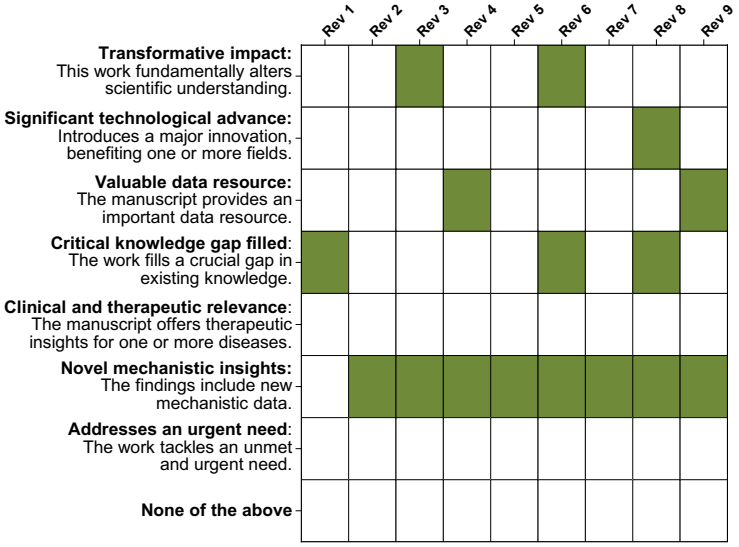
Q6 – Indicate the extent that the findings add new knowledge and advance the field?

1. **Extremely significant:** Major breakthrough, transforming the field, changing paradigms.
2. **Very significant:** Substantial new knowledge, significantly advancing the field.
3. **Moderately significant:** Valuable insights, contributing to existing knowledge.
4. **Slightly significant:** Some new knowledge, providing an incremental advance.
5. **Not significant:** Minimal new knowledge and do not meaningfully advance the field.

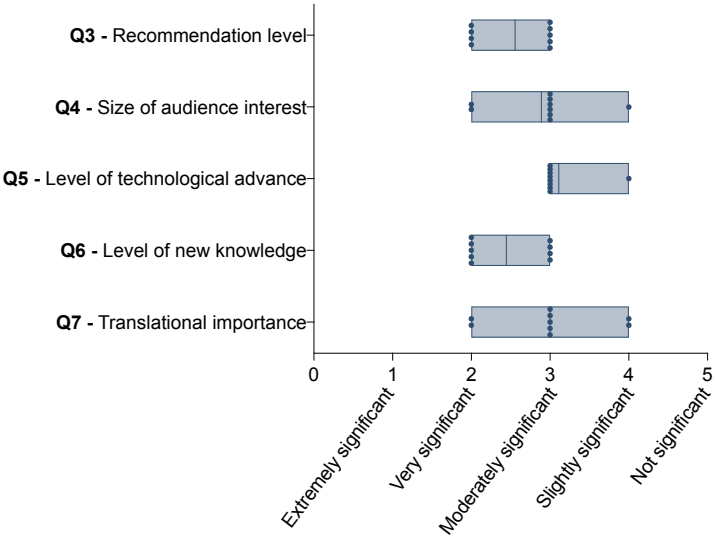
Q7 – What is the level of translational importance of the findings?

1. **Extremely important:** Wide-reaching implications and address an urgent problem.
2. **Very important:** Significant insights and could influence a broad audience or field.
3. **Moderately important:** Data are valuable but primarily relevant to a specific group.
4. **Slightly important:** Somewhat relevant, with potential impact on a small group.
5. **Not important:** The data have minimal clinical relevance.

C Responses to Multiple-Choice Question 2 for a manuscript



D Responses to Likert-Scale Questions for a manuscript



Supplementary Figure 7. Calculation and Comparison of *Quality* and *Impact* Composite Scores

A

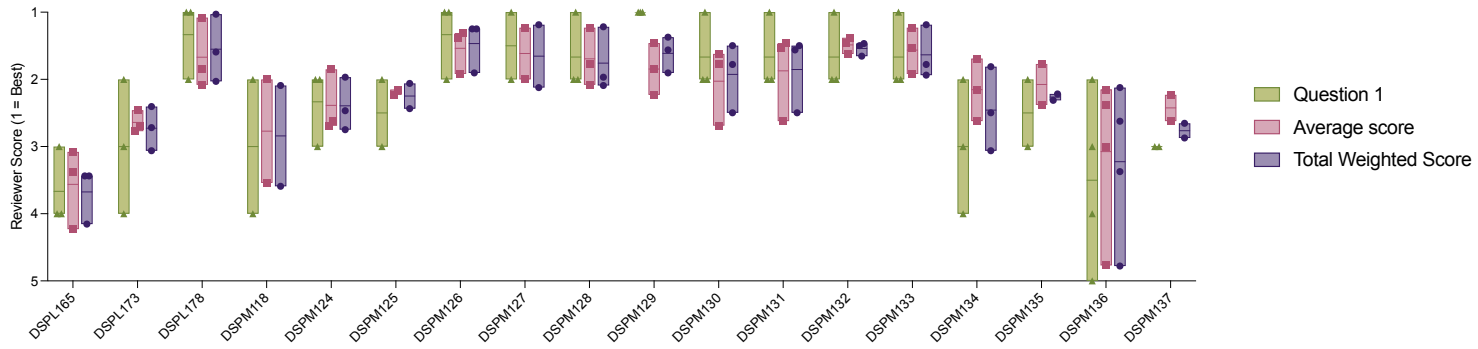
Quality Assessment: Question Weight	**Weight**
Q1_OVERALL, data support conclusions.	10
Q2_Data support MAIN conclusions in TITLE and ABSTRACT.	3
Q3_Data support ALL conclusions in EACH figure.	3
Q4_Conclusions do not contradict each other.	2
Q5_Authors accounted for confounding variables.	2
Q6_Authors used appropriate experimental methods.	2
Q7_Experiments include proper controls.	2
Q8_Experiments include adequate sample size and technical repeats.	2
Q9_Data were analyzed with appropriate statistical tests.	2
Q10_Data and methods include enough clarity to enable replication.	1
Q11_Writing style and flow of the paper is effective and clear.	1
Q12_Authors discuss how findings are consistent with or extend field knowledge.	1
Q13_The authors included appropriate references.	1
∑ (Importance weight)	32

B

Impact Assessment: Q2 Weight of Features Contributing to Impact	**Weight**
Transformative impact: Fundamentally alters scientific understanding	10
Significant technological advance: Major technological innovation	5
Valuable data resource: Important data resource.	5
Critical knowledge gap filled: Fills a crucial gap in existing knowledge.	8
Clinical and therapeutic relevance: Concrete therapeutic insights.	5
Novel mechanistic insights: The findings include new mechanistic data	8
Addresses an urgent need: The work tackles an unmet and urgent need.	5
None of the above	1
∑ (Importance weight)	46

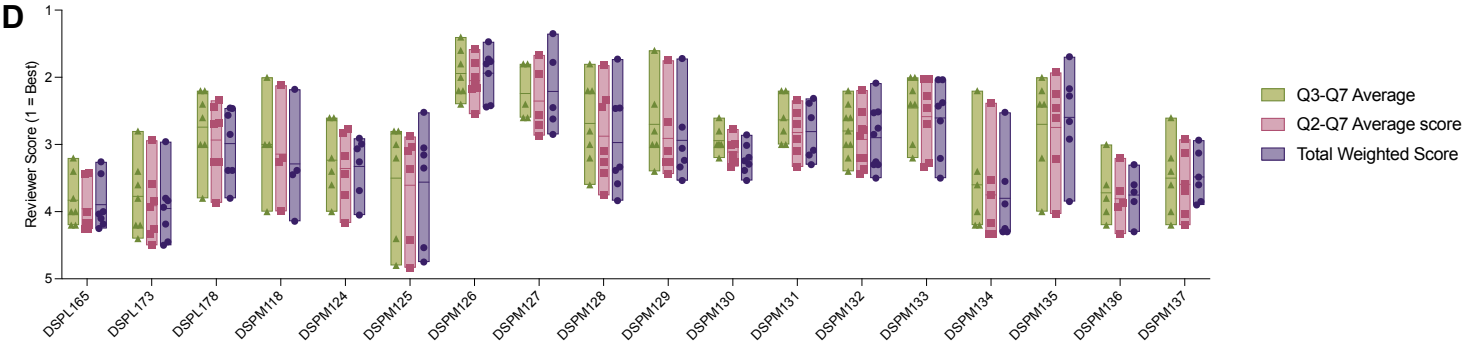
Impact Assessment: Question Weight	**Weight**
Q2_Features that contribute to manuscript's impact:	10
Q3_How likely to recommend manuscript to colleagues.	10
Q4_Size of audience interested in reported findings.	8
Q5_Level of technological advance in the manuscript	2
Q6_Extent findings add new knowledge and advance field.	8
Q7_Level of translational importance.	2
∑ (Importance weight)	40

C Comparison of average ***Quality*** score, average weighted score, and average of "Question 1"

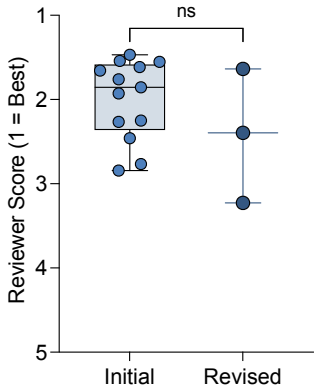
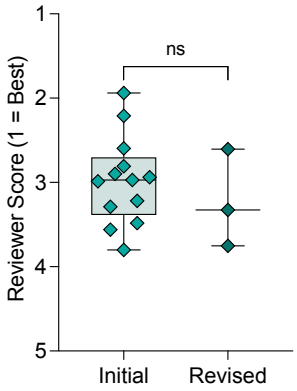


Comparison of average ***Impact*** Q3-Q7, average Q2-Q7, and weighted average

D



E *Quality* Review Scores



F *Impact* Review Scores