SENSITIVE

OBSERVER REPORT

CALL	
Call:	**HORIZON-ER-JU-2022-02**
Topic(s):	• HORIZON-ER-JU-2022-FA7-01 • HORIZON-ER-JU-2022-FA7-02 • HORIZON-ER-JU-2022-FA7-03 • HORIZON-ER-JU-2022-ExplR-01 • HORIZON-ER-JU-2022-ExplR-02 • HORIZON-ER-JU-2022-ExplR-03 • HORIZON-ER-JU-2022-ExplR-04 • HORIZON-ER-JU-2022-ExplR-05 • HORIZON-ER-JU-2022-ExplR-06 • HORIZON-ER-JU-2022-ExplR-07
Type(s) of action:	RIA, IA, CSA
Service:	EU- RAIL
Call deadline:	December 14 2022
Submission model:	Single

EVALUATION	
Evaluation model:	Single
Panel(s):	Panel 1 RIA, Panel 1 IA, Panel 2 RIA, Panel 3 RIA, Panel 3 CSA
Observer(s):	Kristin Oxley

TABLE OF CONTENTS

1. BACKGROUND AND SCOPE

Background and scope

This report describes the observer's assessment of the evaluation of the following call:

Call for proposals: HORIZON-ER-JU-2022-02
Deadline: December 14 2022
Budget: EUR 14,7 MILL

This call covers the following topics:

- HORIZON-ER-JU-2022-FA7-01
- HORIZON-ER-JU-2022-FA7-02
- HORIZON-ER-JU-2022-FA7-03
- HORIZON-ER-JU-2022-ExplR-01
- HORIZON-ER-JU-2022-ExplR-02
- HORIZON-ER-JU-2022-ExplR-03
- HORIZON-ER-JU-2022-ExplR-04
- HORIZON-ER-JU-2022-ExplR-05
- HORIZON-ER-JU-2022-ExplR-06
- HORIZON-ER-JU-2022-ExplR-07

This call covers the following types of action: RIA, IA, CSA

The report analyses the efficiency of the procedures, usability of the instruments (including IT tools), conduct and fairness of the evaluation sessions, and compliance with the applicable rules.

The objective is to give independent advice for improving the evaluation processes for EU funding.

2. OBSERVER ASSESSMENT

Methodology

Methodology

The observer assessed the quality of the evaluation process with respect to its fairness, efficiency, transparency, consistency and the application of rules, guidelines and best practices.

Prior to the consensus stage, the observer attended a web-based briefing and followed the progress of the individual assessment phase through SEP, including the development of IERs and draft CRs.

During the consensus stage, which was done remotely, the observer attended the general briefing, consensus meetings and monitored the development of CRs in SEP. Immediately following the consensus stage, interviews with a selection of experts were carried out to collect their views on the assessment process.

The observer analysed written information pertinent to the call such as HE reference documents, proposal submission and evaluation guide, the call text, IERs and CRs, etc. The observer furthermore analysed relevant research literature on grant peer review and carried out comparisons with evaluation procedures at national and international levels.

Assessment

Assessment

Scale of complexity of the evaluation task

This evaluation of proposals received in response to the call HORIZON-ER-JU-2022-02 was a moderately complex evaluation. In total 13 proposals were evaluated. Proposals were evaluated by three panels, consisting of 4-6 experts and one dedicated recorder. Each panel evaluated 3-6 proposals. The number of proposals received was significantly lower than expected. For several topics only one proposal was received and for one topic no proposals were received. Experts commented that the small number of proposals received for some topics appeared to have negative effects on proposal quality. For the future, Europe's Rail Joint Undertaking (EU-RAIL) could thus consider if there is scope for intensifying communication efforts related to the publication of calls, in order to ensure a somewhat larger pool of applications and by consequence potentially higher quality of the proposals ultimately funded.

The EU- RAIL evaluation team was well prepared to deal with the evaluation task. Prior to undertaking the evaluation, experts received comprehensive briefing material – both written information and an online briefing session - which gave guidance on the relevant policy context, the call, and the evaluation task. The information material gave good guidance on how to evaluate and score the award criteria, including good examples of how to draft appropriate IERs and CRs.

Overall, evaluation tasks were handled with impressive professionalism and thoroughness in all stages of the evaluation. The excellent planning and preparations carried out by the evaluation team facilitated a smooth-running process.

Transparency of the procedures

The procedures were carried out in accordance with very high standards of transparency. They adhered closely to the criteria and processes described in the call and the supporting information. Moderators and recorders ensured that experts addressed all relevant sub-criteria associated with the main criteria, and that they did so in a uniform manner. A separate quality control was carried out by EU-RAIL staff to ensure that the comments matched the scores, and that criteria and sub-criteria were addressed in a uniform manner across different consensus groups.

Consensus reports were of high quality, with judgements that were easily understandable and transparent, and which closely followed the guidelines and information available to applicants through the call.

The use of an independent observer further served to verify that the stated guidelines and procedures were followed in practice. The observer function also provides a way for identifying and airing potential problems and make suggestions/recommendations during and after the evaluation so that corrective actions can be adopted. As the evaluation process in its entirety was carried out in SEP and Teams, this allowed for easy and comprehensive monitoring of the process by the observer.

In sum, the final result was achieved in a transparent way.

Throughput time of the evaluation and the efficiency of the procedures

Excellent throughput time
The throughput time of the evaluation was excellent. In the individual evaluation stage, experts were observed to adhere to the deadlines set, and no significant delays were in evidence. In the consensus stage no major delays were observed. Time set aside for consensus discussions was 1,5 hours, with some additional time (0,5 hours in between proposal discussion and approximately one day at the end) set aside for discussing quality control feedback. This was observed to be sufficient to ensure thorough, yet focused discussions. In order to ensure efficient discussions it was observed to be important to focus primarily on the points of divergence, and particularly on the points that have an influence on scoring. As is to be expected, some consensus discussions took slightly longer to complete than the original scheduling foresaw. Two panels finished half a day later than the original schedule foresaw.

Key to achieving the excellent throughput time was the high-quality preparations, including excellent briefing material with detailed guidance for the development of IERs and CRs. Good IERs enables a more efficient elaboration of CR drafts, and subsequently, a good CR draft was observed to be paramount to ensure a good throughput time for consensus discussions. Quality controllers consistently gave feedback on the first CRs recorder developed prior to consensus discussion, further contributing to their high quality.

Generally, the first proposal that panels discussed was observed to take longer to complete as participants had to get used to using the technology, those that are new to the exercise had to get used to the consensus discussion flow, and all needed to get to know each other and their respective expertise in order to efficiently calibrate their individual evaluations against each other. As one interviewee commented, "we had a learning curve working together, and we progressively got better and better at it".

Also, for proposals with a large divergence of opinions/scores amongst experts in their IERs, agreeing on a consensus report often takes more time than for proposals where experts' opinions/scores are largely aligned. With this in mind, it could be worthwhile to schedule proposals with small divergences of opinion among experts as the first to be discussed. Establishing good working relationships and a sense of team sprit within the panel through

discussion of an "easy" first CR can help ensure more efficient processing of consecutive CRs, including those with large divergences of views.

Highly efficient procedures

Very efficient moderation
Moderation was overall very good, helping ensure efficient consensus discussions. Moderators play a very important role in the consensus process. They ensure evaluation criteria are well understood and correctly applied, and also ensure that discussions move forward at an appropriate pace.

Group interaction can make group members motivate each other and increase the amount of information that is collected and discussed, compared with individual decision making. On the other hand, group interaction can result in poorer decision making because shared responsibility creates a situation in which everyone withdraws and no one really puts in the necessary effort, known as social loafing. The moderator is essential in ensuring that the latter tendency of group interaction is avoided. Overall, moderators in this evaluation helped ensure that the added value of the group discussion was capitalized on. They ensured that experts participated actively in the discussion, they would ask pertinent clarifying questions to encourage experts to go further into relevant details, they would give good guidance on score setting, etc.

Very efficient quality control
The quality control of CRs was ensured by internal staff. This was observed to work very well and seemed to be more efficient than the model previously observed in other EU-level evaluations where this function has been ensured by external experts. This is natural as internal staff will have more expertise in the format required due to their continuous exposure to the evaluation process, which will naturally be difficult for any external expert to achieve. While in previous EU-level evaluations observed, the quality control process has proven at times very time-consuming and straining for experts who have had to go through numerous iterations of adjustments to CR text to comply with quality control requirements, in this evaluation it was very swift and effective. When observing the processing of the quality controllers' comments, it was evident that their comments were overall very good, not going into unnecessary detail, and focusing predominantly on corrections required to establishing a correct alignment between comments and scores.

In addition to ensuring high-quality feedback to applicants and efficient calibration of scoring across proposals, the quality control function also served to improve efficiency of discussions. When uncertain about how a criterion should be scored, whether something should be termed a shortcoming or a minor shortcoming, etc., panellists and moderators were observed to cut lengthy discussions by referring to the quality control function of ensuring coherent approaches across reports in accordance with the HE guidelines.

Recorder task handled efficiently

The recorders developed a draft consensus report based on the IERs prior to consensus discussion. During the consensus phase they supported the moderator in seeking consensus and drafted the final consensus report. They were observed to handle their task in a very efficient, high-quality manner, displaying good drafting skills, good command of the English language and good ability to summarize the discussion.

Efficiency, reliability and usability of the procedures, including the IT-tools

This evaluation was carried out remotely using virtual consensus meetings, in addition to a remote individual evaluation step. This overall worked impressively well.

Studies that compare physical panel meetings with panel meetings that use teleconferencing or digital meetings, conclude that the choice of communication channel has little effect on the overall outcome of the evaluation process. Nevertheless, several of these studies find that the changes made to assessments as a result of the panel discussions were smaller for panel meetings conducted via teleconference compared to physical panel meetings, and that discussion time was also shorter than in physical panel meetings.

Other studies that focus to a greater extent on panel members' experience of online panel meetings highlight various disadvantages. Some find that a larger proportion of those who participate in physical panel meetings feel that the meetings are useful for clarifying divergent opinions compared to those who participated in digital panel meetings. Other studies find that experts experience digital panel meetings as more demanding because they are more taxing in terms of ensuring consistent focus, and also put higher demands on participants' technical and communication skills. Online panel meetings also have the disadvantage that it is harder to interpret body language and non-verbal communication compared to physical panel meetings, and panel members report that they miss the social and network-building aspects of gathering for a physical panel. Furthermore, there are indications that people with different expertise who have to collaborate may have difficulties communicating if they do not know each other beforehand and that this problem can be exacerbated if communication takes place digitally.

At the same time, digital panel meetings also have several advantages. The literature highlights the large cost reductions involved and the reduced environmental impacts. Digital panel meetings can also open participation to a more varied group of panel members, including for example researchers who have children and therefore cannot travel as much. Digital meetings also mean that travelling time is eliminated, potentially making it easier to recruit panel members due to the reduced time commitment involved. Previous studies have found that the most common reason for declining requests to participate in peer review was lack of time.

In sum, according to the literature, the positives associated with online evaluation appear to outweigh the negatives. This conclusion is shared by all experts interviewed in connection with this evaluation, and observations also align with this conclusion.

Experts highlighted that they missed the informal communication which onsite evaluation allows for, as this would enable smoother consensus discussions. Others highlighted that the limited ability to read body language in an online context made consensus discussions somewhat more cumbersome. However, all recognised that online discussions also have strong positive sides, including a reduced carbon footprint and a reduced time commitment on their side. Some acknowledged that they would not have been able to participate in the evaluation if they had been obliged to travel to Brussels.

Through a very good combination of online and offline work, this evaluation managed to capitalise on the positives associated with online evaluation while minimising the potential negative effects. Discussion sessions were overall kept quite short, normally around two hours. The maximum observed was three hours. In such cases, where discussion run for more than two hours, taking a pause is advisable. This is so because long hours in front of a screen are more tiring than on-site discussions. According to a 2021 survey of 3,288 grant reviewers for the US National Institutes of Health, reviewers report shorter attention spans and lower engagement during video grant-review meetings than in those held face-to-face. Compared with in-person grant-review meetings, 46% of respondents said that they paid less attention during the video meetings, and 51% said that their engagement was worse[1]. For the future, a standardized approach to taking breaks could be considered; short five-minute breaks per hour for example, in order to ensure experts stay focused.

The technical infrastructure overall worked very well, with no technical difficulties observed. While in previous observations observed, problems with bandwidth have often required participants to keep their cameras off, here there were no such issues, mitigating problems associated with inability to observe body language. Furthermore, participants made active use of the online tools available to facilitate discussions, for example using the chat function to suggest text for inclusion in the consensus report. Overall, this is the best functioning evaluation the observer has observed to date in terms of its technical functionality.

SEP was also observed to work well with no issues. However, panels used word as the main working tool to develop their CRs and receive feedback on these from QC. This worked very well and is superior to the use of SEP in this respect, as SEP functionality for cooperative working on documents is poor. However, this practice is more vulnerable with respect to confidentiality, see section below.

Impartiality, fairness and confidentiality of the evaluation

Research suggests that peer review may suffer from a number of biases, including cronyism, gender bias and cognitive particularism (tendency of experts to favour their own field or way of thinking). This evaluation had a number of procedures in place to avoid bias and ensure impartial and fair reviews. The observer is convinced that the evaluation was impartial, fair and confidential, constituting an international best-practice example in this respect.

The evaluation was impartial

Cronyism is a concern for all research funders, and the majority have detailed conflict of interest processes in place to counter the presence or perception of such biases. This evaluation clearly informed experts on the rules regarding conflicts of interests in numerous instances – through the briefings, through the contracts and through the information material supplied. There was no evidence of unmanaged conflicts of interest.

The evaluation was fair

All proposals were evaluated with parity, and criteria and scoring scales were interpreted in a uniform fashion and applied coherently.

Research offers mixed judgements on the fairness of grant peer review, identifying variable levels of decision-making consistency between review panels, suggesting the outcome of evaluations are not only down to the quality of proposals, but also to a large degree down to the "luck of the reviewer draw" [2] A recent study comparing the outcome of two independent panels evaluating identical proposals found an agreement rate of 83%.[3] A previous study was less favourable, showing agreement levels of 65–69%.[4] However, consistency of judgment between panels in this evaluation appeared very high, likely exceeding the agreement levels found in cases evidenced in the literature due to excellent procedure aimed at securing consistency of scoring across panels.

As mentioned above, experts are given extensive information and guidance on how to correctly assess proposals. Recent studies suggest that this type of training for peer reviewers is effective in raising the reliability of the evaluation exercise.[5]

Furthermore, panels consistently checked back with proposals previously evaluated to ensure that they were applying a uniform evaluation approach, and dedicated quality controllers helped bolster the fairness of the

[1] Kaplan D, Lacetera N, Kaplan C (2008) Sample Size and Precision in NIH Peer Review. PLoS ONE 3(7): e2761. doi: 10.1371/journal.pone.0002761

[2] Bornmann L, Mutz R, Daniel HD: Latent Markov modeling applied to grant peer review. J Informetr. 2008; 2(3): 217–228; Cole S, Cole JR, Simon GA: Chance and consensus in peer review. Science. 1981; 214(4523): 881–886;

[3] Clarke P, Herbert D, Graves N, et al.: A randomized trial of fellowships for early career researchers finds a high reliability in funding decisions. J Clin Epidemiol. 2016; 69: 147–151.

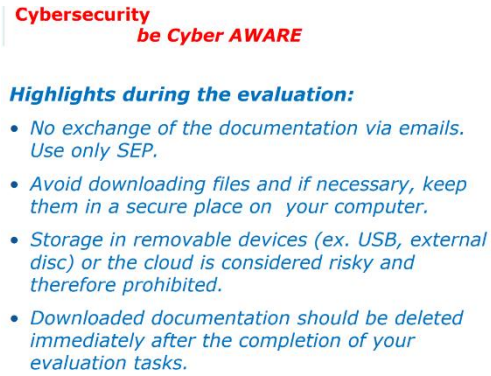
[4] Fogelholm M, Leppinen S, Auvinen A, et al.: Panel discussion does not improve reliability of peer review for medical research grant proposals. J Clin Epidemiol. 2012; 65(1): 47–52.

[5] Sattler DN, McKnight PE, Naney L, et al.: Grant Peer Review: Improving Inter-Rater Reliability with Training. PLoS One. 2015; 10(6): e0130450.

evaluation exercise further by assessing the consistent operationalisation of evaluation criteria and consistent application of the scoring scale across proposals and across panels.

The evaluation was confidential
The importance of keeping the evaluation confidential was underlined to evaluators on numerous occasions - in the written information material supplied, in the general briefings, etc.

However, for the future, guidance to experts could also usefully include some general information on the importance of IT security in order to secure confidentiality. Given that the evaluation process in its entirety is now carried out online, increased attention to IT security is warranted and other EU evaluations observed have started to include this kind of information in its briefings. Example:



UPDATE YOUR DEVICES, SOFTWARE & ANTI-VIRUS Regular updates of your personal devices, software and anti-virus will install patches to fix known vulnerabilities and increase your cybersecurity level.	*PASSWORD* Do not reuse passwords. To make them complex use a pass-phrase instead, and combine it with two-factor authentication if available.
SECURING 'SOCIAL BEHAVIOUR Be cautious with what you share on social networks. What is public about your private life might get into the wrong hands. Manage privacy settings wisely.	*THE SECRET BEHIND THE 'S'* When navigating the web, remember to check your sources: *httpS://* in a link, means the URL you are visiting uses secure measures.
DON'T GET PHISHED Phishing may come from known e-mail addresses, like your bank or familiar sites. Always double check the source and if it looks suspicious delete it or block it.	*WHAT ABOUT MOBILE AND COMPUTER APPS?* Remember to read applications' requirements and to check the source. Do not give them access to all your mobile's features, and delete them if not used.
BE AWARE, CONNECT WITH CARE Avoid using free Wi-Fi, rather opt for a trusted VPN service. Disconnect Wi-Fi, location service and Bluetooth when you don't need them.	*BACK UP WHAT YOU CANNOT GET BACK* Get into the habit of regularly doing a back up of all your files on an external hard drive. This will prevent you from losing them in case of an attack.
BE BOTH ADMIN AND USER OF YOUR COMPUTER To enhance security on your private computer create two profiles: Admin and User. The User one will request your Admin password to allow any kind of download or action.	*COVER IT UP* Most devices have a camera incorporated. Use a camera cover to avoid incidents in case the device is compromised.

For the future, the use of email to exchange consensus reports back and forth between moderators, recorders and QCs should be reconsidered due to the confidentiality risks involved. In other evaluations observed, a shared drive has been used for this communication, and could be explored as a more secure option for the future.

Conformity of the evaluation with the applicable rules (including guidance documents)

The evaluation was conducted in full conformity with applicable rules and guidelines published in the context of the call. These were communicated to experts in the various instructions and briefings given and reinforced during the evaluation through the active moderation of discussions, and further checked and verified through the QC process. Based on previous experience, which revealed that experts found it particularly challenging to understand the guidelines regarding lump sum and how these should be followed up in their evaluation, dedicated measures aimed at clarifying and aiding experts in this regard were foreseen. This included a dedicated briefing on the subject, a lump sum and Q&A session foreseen for each proposal in order to clarify issues that proved unclear in the course of discussion, and the inclusion of a financial expert in each panel. This was excellent.

Quality of the evaluation process in comparison with similar national/international evaluation procedures

In the context of carrying out a PhD on the subject of panel peer review of research grant applications, the observer has over the last two years spent more than 130 days observing more than 120 panels at national, Nordic and EU-level. The observer also knows the literature on grant panel peer review well, including the large empirical literature detailing peer review procedures in use internationally. Based on this knowledge base, EU-RAIL's procedures are comparably of excellent quality.

The process was very reliable; experts are given extensive information and guidance in how to correctly assess proposals. Experts reported that they were very pleased with both the information they received prior to carrying out their task and the guidance they received both in the consensus phase.

Furthermore, the evaluation constitutes a best-practice example in terms of its fairness. In line with the literature on peer review, several procedures are in place to help guard against bias and ensure impartial and fair reviews, including strict rules on conflicts of interest, uniform use of evaluation criteria and scores, and efficient quality control mechanisms.

In contrast to many national level evaluations observed, the process was organised in such a manner that it ensured that the value added associated with grant panel discussions could be reaped. The value added of panel discussions compared to just averaging the scores of a number of individual assessments is that a panel can draw on a larger pool of knowledge as each member has unique knowledge they can contribute, enabling a more holistic assessment of proposals. In a panel context, members are also able to process the information they possess more thoroughly through the discussion, asking questions, considering alternative options, weighing arguments against one another, etc. Through such information processing, the group is better positioned than individuals to identify errors. While many national evaluations observed fail to reap these benefits due to insufficient resources and time invested in both the individual and consensus part of the evaluation process, in this evaluation staff assisted experts in developing high quality IERs and CRs and ensured sufficient discussion time and effective moderation which enabled experts to draw effectively on each other's' expertise.

Overall quality of the evaluation

The evaluation constitutes an example of international best practice. It was conducted overall to the highest professional and quality standards. Both EU-RAIL staff and external experts demonstrated high competence and professionalism, and extensive quality assurance and calibration procedures were implemented. In sum, this meant proposals were evaluated and ranked in accordance with their merit in an efficient, fair, reliable and transparent manner, with high-quality feedback to applicants as the final result.

Other remarks

Quality of the documentation provided to experts beforehand
The quality of the documentation received by evaluators prior to starting evaluations was excellent. Comprehensive written briefing material was made available, including detailed guidelines on the development of appropriate IERs/CRs. Comprehensive guidelines were also available on how to use SEP to carry out the evaluation, as well as a useful video tutorial on how to submit expense claims.

Quality of the briefing sessions
Briefings were very informative and well-structured. Briefings carried out prior to the individual evaluation were tailored to the different needs of the experts and the recorders, and a shorter briefing for all involved was carried out immediately preceding the consensus discussions. In this manner experts were provided with a good overview of the evaluation procedure, the different roles involved and received guidance on how to carry out the evaluation in a timely and tailored manner.

In many national level evaluations observed, briefings are only carried out prior to the consensus stage. The fact that comprehensive briefings are also organised prior to the individual phase is very positive, as this helps ensure a uniform understanding and application of criteria and scoring scales at an early stage in the evaluation, with less effort having to be invested in this at the consensus stage.

The understanding by experts of the call (context, topics), of the evaluation process and their role
Evaluators appeared to have a clear understanding of the call, the evaluation process, and their role, and they reported that they were very pleased with the information and guidance they received both during the individual and the consensus phase of the evaluation. This is in contrast to international studies of peer reviewers, where for example a survey of experts used in application assessment at nine different research funding agencies internationally showed that only 16% considered that the research councils provided clear guidelines for the task.

As previous experience had shown experts to struggle with the guidelines regarding lump sum and how these should be followed up in their evaluation, dedicated measures aimed at clarifying and aiding experts in this regard were foreseen, which functioned very well.

Criteria and scoring scheme: appropriateness, completeness, relevance, clarity, consistency in application
Research on peer review shows that experts often have quite different opinions on how assessment criteria should be interpreted and emphasized in the overall assessment, and different practices regarding how assessments are translated into scores.

This evaluation had a number of effective measures in place to ensure that the evaluation criteria and scores were understood and used by experts in a uniform manner, ensuring a fair evaluation. The scoring scheme and the three evaluation criteria were thoroughly explained in the briefings. Moderators actively ensured that experts used the criteria and scores appropriately, ensuring the right issues were addressed under the right criteria, and that double penalisation was avoided.

However, one of the panels estimated that the “methodology” in excellence and the “pathway to achieve the expected outcomes” in impact led to some confusions, and recommended more clarity in this regard during briefings.

Furthermore, the formulation of the scoring scheme, whereby the number of shortcomings identified in the proposal largely determined the score, lead to a practice where panels simply counted shortcomings in order to determine the score for each criterion. In general, very limited attention was paid to the strong points of the proposal when determining the scores. As an example, one panellist was observed saying: "My suggestion is to avoid qualifying the strengths saying this is good and very good etc, because in any case they do not impact the final score. Let's now just count the shortcomings". Another panellist echoed this interpretation in an interview, saying that the consensus discussions was a learning exercise, and among the things she learnt was that "strengths don’t count". When questioned about their views on this practice, experts testified that in their individual assessment they tended to put more equal emphasis on both strengths and shortcomings, and they felt the consensus process should allow for this to a greater extent too.

While the focus on negatives is not a problem exclusive to this evaluation, but rather a general problem for peer review[6], the strong focus on the number of shortcomings as a means for establishing the score appears to further exacerbate the negativity focus that peer reviewers are prone to. This in turn might lead to the discrimination of very innovative proposals, as taking a big step forward is fraught with more difficulty and thus the potential for more shortcomings to be identified, compared to taking smaller, more incremental steps forward[7]. When the strong points of a proposal are not sufficiently balanced against the weak points, the end result might be a competitive advantage for incremental over highly innovative proposals. While this conservative nature of peer review is a general problem not confined to EU-level evaluation processes, the 2022 study of the proposal evaluation system for the EU R&I framework programme[8] estimates that there is evidence that EU proposal evaluation processes are anti-innovation.

For the future then, means of better balancing the strong points of proposals against shortcomings should be sought. For example, equal attention could be paid to correctly qualify the strong points of a proposal as is currently paid to qualifying the shortcomings. While all panels in this evaluation were observed to focus extensively on qualifying the shortcomings and weaknesses of a proposal in accordance with the scoring scheme, labelling them minor shortcomings, shortcomings or significant weaknesses, very limited attention was paid to the qualification of strong points. Efforts could be made to also label the strong points in accordance with the scoring scheme – assessing if they are excellent, very good, good, etc., and base the scoring decision in equal measure on both the strong and weak points of applications.

The process of the individual evaluations and the actors involved

The process of the individual evaluation proceeded according to schedule and no major issues were observed or reported. During the individual evaluation phase, each expert reads the assigned proposals and writes an IER for each proposal. Following the completion of the IERs, designated recorders develop draft Consensus Reports based on the relevant IERs. The quality of the IERs and draft CRs was overall very good. Recorders consistently received feedback on the first CR developed, ensuring CRs constituted a very good basis for the efficient discussion of proposals during the consensus phase.

The process of the consensus meetings and the actors involved

Consensus discussions were observed to be very well structured and thorough. Some panels started off discussions with a tour de table where panellists presented their overall assessment of the proposal. While this procedure was not applied in all groups, it could be considered to make this a standard element in the evaluation process for the future. It had the added value that panellists got a good overall discussion of the strengths and weaknesses of the project before diving into the details, as well as an impression of the main points of divergence. This facilitated the more detailed discussion of strengths and weaknesses.

Experts are obliged to provide comments and assessments for all the relevant sub-criteria included under each of the three evaluation criteria and they were observed to do so systematically and thoroughly. This is very positive as research shows that structured and detailed discussions help to ensure a uniform and fair treatment of proposals since the same elements are evaluated in all proposals and no other elements than those included in the criteria influence the evaluation.

Experts were instructed to agree on comments before discussing scores. This is positive as it helps ensure that consensus discussions are not reduced to simple averaging exercises. Focusing on comments rather than scores ensures a more thorough and in-depth discussion and can also make it easier for evaluators to focus less on their original and preferred score, making for more flexible discussions where consensus can be reached with greater ease. This is confirmed in other EU-level evaluations. In 2019 and 2020, Marie Skłodowska-Curie Actions trialled a system whereby experts gave no scores in their IERs, scores were only set once an agreement had been reached on comment. Feedback received on the trial stated that the consensus meetings were smoother and that there were fewer inconsistencies between scores and comments.[9]

Panels differed in whether they discussed scores immediately after agreeing comments for a criterion, or whether they agreed all scores once all consensus text had been agreed. While both approaches overall worked well, the approach whereby panels agreed the score for a criterion immediately after finishing the text appeared more efficient as the discussions was then fresh in participants' minds. When left until the end, score discussions could

6 Van den Besselaar, P., Sandström, U., & Schiffbaenker, H. (2018). Studying grant decision-making: a linguistic analysis of review reports. Scientometrics, 117(1), 313-329.

7 Lane, J. N., Teplitskiy, M., Gray, G., Ranu, H., Menietti, M., Guinan, E. C., & Lakhani, K. R. (2022). Conservatism gets funded? a field experiment on the role of negative information in novel project evaluation. Management Science, 68(6), 4478-4495.

8 Rodriguez-Rincon, D., Feijao, C., Stevenson, C., Evans, H., Sinclair, A., Thomson, S., & Guthrie, S. (2022). Study on the proposal evaluation system for the EU R&I framework programme. Final Report.

9 Rodriguez-Rincon, D., Feijao, C., Stevenson, C., Evans, H., Sinclair, A., Thomson, S., & Guthrie, S. (2022). Study on the proposal evaluation system for the EU R&I framework programme. Final Report.

sometimes result in a reopening of the consensus discussion of the criterion. Agreeing scores successively as discussions progress could thus be considered as the standard process for the future.

The standard of experts was high, with a good mix of expertise, gender and sector. This is important, as research on decision-making shows that groups with heterogeneous members with complementary skills make better group decisions than homogenous groups.[10] Experts were observed to actively draw on each other's expertise and were overall very active in discussions. This was confirmed by experts in interviews. They consistently commented positively about their fellow experts, and felt that the panel groups were very well composed, with the different experts each holding slightly different but complimentary expertise, resulting in high quality assessment. All experts felt that the consensus discussions constituted a considerable added value over what could have been achieved by simply averaging the individual assessments, highlighting that through discussions they were able to discover mistakes, see new perspectives of the proposals, and that they quite often changed their opinion on important aspects of the proposals due to the input of their fellow panellists.

Panels consisted of 4-6 experts. Based on a review of over 30 European research funding organisations, the European Science Foundation concluded that "In general, the aim should be to provide at least three expert assessments before a final decision is made".[11] The number of experts assessing each proposal in this evaluation thus exceeds international best practice. It is positive that four experts were used rather than the EU evaluation minimum number of three, as this makes the evaluation exercise less vulnerable to unforeseen events. For example, in one of the meetings observed one of the experts did not show up for discussions as agreed, only turning up two hours later. While in the instance observed, the group decided to halt discussions and await the missing expert, the fact that they were six experts in the panel meant that discussions could theoretically have continued as the number of experts was well above the minimum required. As this kind of event is to be expected when evaluations are carried out online, it is very positive to use a number of experts that is robust in the case of such unforeseen events.

Quality of evaluation summary reports
The quality of ESRs was excellent, with thorough and helpful feedback that matched the scores, enabling applicants to understand the basis for the decision to fund or reject the proposal. Experts were mindful of ensuring that all proposals were judged according to the same standards, in a uniform manner. The efforts deployed in this respect, e.g. through the moderation and the quality checks is highly commendable.

Working conditions for evaluators
Compared to national and Nordic evaluations observed, renumeration is generous and time foreseen for discussions allowed for thorough and detailed evaluation of proposals. Overall, experts interviewed reported to be content with the time available for evaluation and with the renumeration offered, although some commented that the time used for the individual evaluation phase exceeded the renumeration offered. Some also pointed out that EU-level renumeration has been fixed over numerous years, while private sector renumeration for similar work has increased progressively to take account of inflation, etc, making EU-level evaluation renumeration increasingly uncompetitive compared to similar work on offer. Some also remarked critically on working hours during the consensus phase, which on occasion lasted substantially longer than foreseen in the scheduling. The fact that the number of proposals received was substantially lower than estimated also proved problematic for some experts that had gone to great lengths to clear their calendar to enable individual evaluation of a large number of proposals and consensus discussions over numerous days, only to end up with less than half of the work estimated.

While experts had the option to contact the observer to give their views on the evaluation process, it might prove an added value if experts' feedback is also collected in a more systematic manner. In other EU-evaluations observed, experts routinely receive an online questionnaire soliciting their views on the evaluation process after completion of the evaluation, and this might also be trialled by EU-RAIL in the future in order to enable even more systematic work on continuous improvements to the evaluation exercise.

Overall conduct of staff
EU-RAIL staff was highly professional and very helpful to both experts and the observer. All moderators had the necessary skills and contributed to creating a positive and productive atmosphere among experts and all experts interviewed only had positive remarks to offer regarding the staff.

Recommendations

Recommendations

- **Consider if communication efforts related to the publication of calls can be strengthened**
  The number of proposals received in response to the call was significantly lower than expected. Intensifying communication efforts related to the publication of calls could thus be considered, in order to ensure a somewhat larger pool of applications and by consequence potentially higher quality of the proposals ultimately funded.

[10] Levi, D. (2007): Group dynamics for teams
[11] ESF (2011): European Peer Review Guide: Integrating Policies and Practices into Coherent Procedures.

- **Consider scheduling the sequence of proposal discussions strategically**
  Proposals with small divergences of opinion among experts could be discussed first. Establishing good working relationships and a sense of team sprit within the panel through discussion of an "easy" first CR can help ensure more efficient processing of consecutive CRs, including those with large divergences of views.

- **Consider uniformly implementing a more active moderation approach**
  While some moderators were observed to take a quite active approach to moderation, others had a more withdrawn style, leaving moderation to be handled predominantly by the rapporteur. While both approaches overall worked well, the more active moderation style appeared more effective in ensuring swifter discussions and also appeared superior in terms of task-sharing. Moderators are experts on the process and the appropriate understanding of criteria and use of scoring scale and are better placed than external recorders to effectively moderate discussions pertaining to these issues.

- **Consider more frequent breaks in order to tailor working conditions to the online format**
  Discussing online is more tiring than discussing onsite, and while overall discussions in this evaluation were quite brief, some were observed to go on for up to three hours. A standardized approach to taking breaks could usefully be considered; short five minute breaks per hour for example, in order to ensure experts stay focused.

- **Consider a more restrictive approach to the use of email for exchanging evaluation information**
  Panels used word as the main working tool to develop their CRs and receive feedback on these from QC. This worked very well and is superior to the use of SEP in this respect, as SEP functionality for cooperative working on documents is poor. However, this practice is more vulnerable with respect to confidentiality as email was used to exchange documents back and forth. For the future, a shared drive could be considered as a more secure option for such communication.

- **Consider improved information to experts on IT security**
  Given that the evaluation in its entirety is now carried out online, guidance to experts could usefully include some general information on the importance of IT security in order to secure confidentiality.

- **Consider how assessments to a greater degree can focus on the strong points of proposals.**
  There is a risk that an excessive focus on the negative aspects of proposals can lead to an anti-innovation bias in assessments. Means of better balancing the assessment of the strong points of proposals against shortcomings should thus be sought. Equal attention should be paid to correctly qualify the strong points of a proposal as is currently paid to qualifying the shortcomings. While all panels in this evaluation were observed to focus extensively on qualifying the shortcomings and weaknesses of a proposal in accordance with the scoring scheme, labelling them minor shortcomings, shortcomings or significant weaknesses, very limited attention was paid to the qualification of strong points. Scores were predominantly set based on a count of shortcomings. Efforts should be made to also label the strong points in accordance with the scoring scheme – assessing if they are excellent, very good, good, etc., and base the decision on final scoring in equal measure on both the strong and weak points of applications.

- **Consider collecting experts' view on the assessment process more systematically**
  While experts had the option to contact the observer to give their views on the evaluation process, it might prove an added value if experts' feedback is also collected in a more systematic manner. In other EU-evaluations observed, experts routinely receive an online questionnaire soliciting their views on the evaluation process, and this might also be trialled by EU-RAIL for the future in order to in order to enable even more systematic work on continuous improvements to the evaluation exercise.