TMC Research

Evaluation of the 2022 Peer Circle experiment at the Alexander von Humboldt Foundation

Peter van den Besselaar, Charlie Mom, Michelle Herte

TMC Research

Middenweg 203
1098 AN Amsterdam
The Netherlands

Email: info@teresamom.com
Phone: +31 20 663 2184

(empty page)

Preface

This report is a result of the evaluation study that accompanied the Peer Circle experiment organized at the Alexander von Humboldt Foundation in 2022. Review models and their effects are an important topic, and the findings of this report may contribute to the understanding of how grant review and grant selection processes function, and how these can be improved.

The project started in spring 2021 with the development of the research plan, which was adopted in early 2022. The Peer Circle experiment took place in two rounds, namely the summer round from March to June 2022, and the fall round between mid-August and mid-November 2022.

The authors would like to thank the members of the Peer Circle (the reviewers) for their collaboration in the interviews and in two surveys. We are also grateful to committee members and AvH staff members for participating in interviews and for allowing us to observe the committee meetings. The Alexander von Humboldt Foundation has supported the study financially, has provided data, and has made data collection possible.

Dr. Michelle Herte (AvH) wrote the description of the HFST funding instrument and the evaluation procedures, which are included in this report as Chapters 2 and 3. She also performed the role of liaison with the AvH Foundation in a fast and accurate manner.

Amsterdam, February 12, 2023

(empty page)

Management summary

Recent discussions about the quality and sustainability of (grant) peer review have led to a series of proposals to improve the peer review system. The main problem is that it is increasingly difficult to attract peer reviewers because many potential reviewers lack time. The Alexander von Humboldt Foundation (AvH) alone requires around 5,000 external reviews for around 3,200 applications per year, in addition to reviews supplied by members of the selection committees. Including host statements and references, AvH selection procedures require around 12,000 reviews and statements by the scientific community every year. Data monitored by the AvH shows that around 60% of review requests are rejected, most of them due to lack of time. Because of this strain on the system, alternative approaches need to be considered. Are there other models for organizing the review process in a less time-consuming manner that would solve this problem? Can this be done without compromising the quality of assessments? And would such an alternative model be accepted by the scientific community?

In the HFST program – which is the focus of this report – the goal is to have two external reviews per application. In the conventional peer review model, reviewers are identified and appointed individually for each application. The Peer Circle (PC) model is an attempt to improve the peer review system by creating a collaborative procedure. In this model, a Peer Circle consists of five to ten reviewers plus committee members from the same field, who together review between 15 and 30 applications in two rounds. For each application, one or two of the reviewers are designated as "first reviewers", who are expected to start the review process with a structured but comprehensive assessment. Other reviewers are expected to review those parts of the applications that they believe they can review given their experience and expertise. Finally, the reviewers are expected to comment on one another's (partial) reviews. Taken together, this should produce a high-quality and complete assessment of the application.

The Peer Circle experiment was conducted in 2022 during the summer and fall rounds of the HFST program in four fields: Inorganic Chemistry, Materials Science, Zoology, and Modern History. For the evaluation of the experiment, these fields were compared with the same fields in 2021 (the year before the PC was introduced) and with four other fields that are close to the four experimental fields in terms of disciplinary culture: Solid State Chemistry, Materials Engineering, Plant Science, and Ancient History.

The evaluation aims to answer the following questions about the Peer Circle:

1. Is the quality of the resulting reviews at least equivalent to that of conventional reviews?
2. Does the interaction within the PC result in premature consensus, hindering critical consideration of the applications?
3. Do the reviewers and committee members perceive the PC as a better alternative?
4. Is the PC more efficient than the conventional approach?
5. Does the PC committee select the best applicants?
6. Do the PC reviews influence the gender balance in the selection outcome?

Data and methodology

The evaluation is based on several datasets. Interviews were conducted with all PC reviewers and with all PC committee members who participated in the experiment, and with involved AvH staff. In addition, reviewers were asked to complete two surveys, one after the first round and one after the second round. These data included Peer Circle participants' opinions and self-reported review behavior. The review process was largely automated on an online platform, and therefore logfiles from the platform could be used to track review behavior in terms of intensity and distribution over time of the activities. The review texts and application texts were analyzed, and administrative data files from the AvH provided information about the applicant (age, academic age, nationality, institutional affiliation, etc.), as well as information about the grading. Finally, bibliometric data were collected for two fields to estimate the academic performance of granted and non-granted applicants.

Summary of the findings:

1. Quality and comprehensiveness of PC reviews

Almost all PC members find that the PC produces comprehensive reviews, and when specific expertise was lacking, it was easy to add an outside expert to the PC for specific issues. The PC consists of a diverse set of reviewers, bringing more expertise to the process than can be done in conventional reviews. This leads to even better reviews, as more perspectives on an application lead to a more comprehensive review, to more complete coverage of the various review criteria, and to a more transparent, more objective, and less biased outcome. The assessment is less dependent on the opinions of one or two reviewers only. Another advantage of the PC is that the reviewers have an overview of all applications in the PC and therefore tend to assess the applications relative to one another. This makes grading easier and would lead to more differentiated and realistic grades.

Most committee members also see advantages of the PC over conventional reviews. Comments from a larger number of reviewers allow committee members to have more confidence in the results of the assessment. In addition, reviewers' arguments are clearer (compared to conventional review), and when something is unclear, it is easy to go back to the reviewer for clarification. The PC also makes it easier for committee members, since in cases where the PC arrives at a clear "fund" or "reject" decision, that decision can simply be accepted, allowing committee members to focus on less clear cases, informed by the critical points identified in the PC assessment. This can work because more views in the PC lead to more certainty for committee members. Finally, the Peer Circle has the advantage of solving the problem of late reviews.

In summary, most participants in the experiment found the PC to produce better reviews than the conventional review process, and only a few were unsure or preferred the conventional approach.

2. Consensus

Almost all the reviewers interviewed were aware of the possibility of being influenced in a group process such as the PC, and most reviewers indicated that they were influenced to

some extent by others. However, influence of others helps one to reflect and develop one's own point of view, and that is a positive since the goal of the evaluation process is to reach some consensus on who should be funded.

Influence is not the same as premature consensus with the risk of suppressing possible conflicting views. The interviewees (reviewers, committee members, department heads) were all very aware of this risk, but almost none felt it had occurred. When it did, it was explicitly addressed within the PC.

3. Acceptance

Participants in the Peer Circle experiment were not randomly selected, and the number is far too low to generalize findings. Nevertheless, they came from different research fields, thus creating adequate disciplinary diversity. With these limitations in mind, one can conclude that the PC is generally appreciated and perceived as a promising way forward. In the surveys and in the interviews, the majority of PC members expressed positive views regarding the PC as an alternative to the conventional peer review model, stating that it is the way forward. Reviewers also noted that they can advance both their reviewing and application skills from reading the assessments of other reviewers. Only three (of the ten) members of the Modern History PC preferred the conventional review model. The vast majority of committee members shared the reviewers' positive opinion, with Modern History representatives again being more critical.

4. Efficiency of the PC process

Most PC members reported that the PC saved time and that the time spent per application was lower because writing was more efficient. Despite more reading work, the overall time investment for the PC is perceived as reasonable. For the individual reviewer, the PC may take (slightly) longer than a single review using the conventional approach, whereas for the community as a whole it saves a lot of time. For example, 89 applications were processed by 27 PC members (reviewers), which would have required 178 reviewers using the conventional procedure (if one follows the norm of two reviewers per application).

The AvH was able to recruit PC members from a larger population of potential reviewers, and consequently the PC also included relatively young reviewers. Interviews and logfiles show that the number and quality of review contributions are not related to age.

Applications that were on the agenda of the PC committee led on average to less discussion than in the other committees, suggesting that the PC review process reduces uncertainty and makes committee meetings at least as efficient.

5. Selection results

It is known from the literature that the way in which review and selection processes are organized affects the outcome. Is the quality of the successful applications in the PC equivalent to that of applications that were successful under the conventional approach? Two aspects were analyzed. First, the success rates and average scores given by the selection

committee were compared between the PC and the control fields. Second, for the PC and control fields in Chemistry only, bibliometric indicators were compared. The results for both comparisons indicate no significant difference between PC and conventional peer review, suggesting that the PC leads to an equally strict decision-making process.

For both Inorganic Chemistry (PC) and Solid State Chemistry (control group), bibliometric analysis shows the overall scores of granted applicants to be no higher than those of rejected applicants, suggesting that publication performance is less important in the decision-making process than expected from review reports and committee meeting observations. Due to the low number of applications analyzed, this would need to be confirmed in a larger study.

6. Gender balance

Does the PC affect the gender balance in the outcomes of the selection process? The success rate of women among applicants assessed in the Peer Circle is similar to the success rate of women assessed in the conventional manner.

However, this comparison is rather aggregate, and at the level of individual research fields results vary widely by field and year. These may be (random) fluctuations related to the low numbers and the varying quality of applications. The differences in success rates may also indicate gender bias, but this cannot yet be determined. Further analysis is needed to include larger samples and data on applicant performance. For now, it can be concluded that the Peer Circle is neutral in terms of the success chances for women, but this issue needs further investigation.

Table of contents

Part I – Case and methodology

1. Introduction

There is a need to experiment with new models for grant peer review because the functioning of the conventional peer review system is problematic and it is becoming increasingly difficult to find enough reviewers. In this research report, we evaluate the Peer Circle project, an experiment with a new form of peer review that took place between March and November 2022 at the Alexander von Humboldt Foundation (AvH).

This evaluation was designed as a multi-method research project and compares the Peer Circle with the conventional approach for peer review of grant applications. Multi-method in this context means that we used a wider variety of data and methods than is commonly done, and integrated usual data sources such as documents, interviews, and a survey, with other data sources, such as bibliometric data, textual data, and the logfiles of the online platform on which the review activities take place.

The interviews lead to rich inputs, which sometimes point in the same direction, but are sometimes (partially) contradictory. We have tried to reflect this as well as possible, but of course we cannot include everything that was said – among other things because this would make it easy to link statements to the interviewed person, which needs to be avoided.

The structure of the report is as follows: Part I of the report consists of three chapters in addition to this introduction. Chapter 2 gives a description of the HFST funding scheme under which the Peer Circle experiment took place. Chapter 3 describes the conventional review format and the Peer Circle in some detail. These two chapters were written by Michelle Herte (AvH). Chapter 4 discusses the approach used in the study, including the types of data used.

Part II consists of four chapters outlining the findings of the study, namely the quality of the reviews (Chapter 5), the decision-making process and the resulting selection (Chapter 6), the activities within the Peer Circles (Chapter 7), and the design and implementation of the Peer Circle procedures (Chapter 8).

Part III consists of the conclusions (Chapter 9) and recommendations (Chapter 10).

A separate paper will contain the annexes, one providing more details about the data and the methodology used, and another containing recommendations for the online platform.

2. Selection under the Humboldt Research Fellowship Program

2.1 Program and selection criteria

We will start by summarizing the regular Humboldt Research Fellowship (HFST) selection procedure for better contextualization of the Peer Circle pilot and the results of its evaluation.

In providing Humboldt Research Fellowships, the Alexander von Humboldt Foundation enables highly qualified scientists and scholars from abroad to spend extended periods of research in Germany. Scientists and scholars from all disciplines and countries may apply. The program is designed to sponsor the outstanding scientific talent of excellent researchers early in their career. Thus, it addresses postdocs (less than four years after completing their doctorates) and experienced researchers (less than 12 years after completing their doctorates) with above-average qualifications.

A candidate's academic qualification is assessed based on the following *selection criteria*:

- Academic record and academic performance to date (mobility, determination, breadth of research activities, academic productivity).
- Quality of key publications specified in the application (originality, degree of innovation, applicant's contribution in cases of multiple authorship).
- Originality and innovation potential of the suggested research outline (significance for the advancement of the discipline, convincing choice of academic methodologies, prospects for academic development of the applicant, clear focus and feasibility of realization within the requested funding period and at the chosen host institution).
- The applicant's future potential (academic potential, further academic development, career prospects).

2.2 Conventional selection procedure

The selection procedure is fully digitized. Applications can be submitted online at any time and decisions are made thereon four to seven months later during one of three annual selection committee meetings in March, July, and November.

The AvH checks all applications for eligibility and completeness before initiating peer review. During the *review phase*, each application is assessed by two independent expert researchers, who are appointed by the AvH according to their specialization, excluding those researchers who might have potential conflicts of interest according to the Foundation's impartiality guidelines. Reviews are submitted online using a standardized form that comprises the following elements:

- Four key questions reflecting the selection criteria (see above), to be answered both in open form and by grading the applicant's level on a scale from "below average", to "average", to "well above average (top 15%)", to "leaders (top 5%)".
- A summary of the positive and negative aspects of the application.
- The reviewer's overall recommendation to the selection committee, ranging from "rejection" to "borderline case" to "research fellowship".
- A question about whether there are any aspects which militate against sponsorship (e.g., potential conflicts with legally-binding principles of scientific ethics, danger of an

armaments-related technology transfer in terms of legal regulations, etc.), to be specified if answered positively.

Following peer review, the AvH adds a written executive summary of each application based on its key characteristics and the reviews, and suggests the applications for fellowship or rejection. The summary and expert reviews are included with the application documents that are provided to the *interdisciplinary selection committee* that is composed of:

- specialist members: recognized researchers of various disciplines (= specialist selection committee members).
- non-specialist members: voting representatives of public and private funders; non-voting representatives of other science (funding) organizations.

Committee meetings are separated into interdisciplinary groups of similar size to discuss and vote on the applications assigned to their group. Before each committee meeting, the AvH assigns each application to the *specialist member* closest to the applicant's research field and without any conflict of interest.[1] Specialist members are asked to provide a written statement for each application and to classify them into one of three categories:

(1) clearly negative cases to be rejected without discussion ("A" cases, for German "Ablehnung").
(2) cases worth being discussed by the committee ("D" cases, for "Diskussion").
(3) clearly positive cases to be approved without discussion ("S" cases, for "Stipendium").

All committee members are notified of the classifications (A, D, S) before the selection meeting. If a member objects to an A or S classification, the application in question will be handled as a D case to be discussed during the committee meeting. All A and S cases are decided *en bloc* at the beginning of the meeting. Subsequently, specialist committee members present their assigned D cases, which are then discussed by the committee. Finally, committee members vote for each D case by allocating 2 points (approved with high priority), 1 point (approved), or 0 points (not approved). The AvH combines the voting results of all groups into one ranking list. Applications are approved if they receive a positive vote (1 or 2 points) from more than half of the voters. If more applications are approved than there are fellowships available, fellowships will be awarded to the highest-ranking applications first.

2.3 Peer Circle selection procedure

The Peer Circle is a new approach for organizing the review phase of the selection procedure, designed to address the problem that it is becoming increasingly difficult to find and retain peer reviewers. The idea behind the procedure is to make reviewing a collective task. Reviewers are appointed not for a single application at a time but for an extended period, during which they are invited to review all applications from their research field together with other experts. All reviewers from the same field plus the corresponding specialist committee members form a Peer Circle. The size of Peer Circles can vary from around seven to ten members, depending on the research field. When appointing expert reviewers based on their

[1] If a committee member's specialization closely matches the topic of the application, they can additionally be asked to provide one of the two expert reviews.

expertise, the AvH strives for diversity within Peer Circles. Ideally, the members of one Peer Circle cover a variety of specializations and academic ages (from senior postdoctoral level onwards), and are balanced in terms of gender.

For the 2022 pilot, four research areas across all scientific disciplines were chosen in order to evaluate the Peer Circle approach in the Humboldt Research Fellowship program (HFST): (1) Modern History, (2) Zoology, (3) Inorganic Chemistry, (4) and Materials Science. The overall selection procedure was kept as similar as possible to the program's regular procedure. Thus, submitted applications from the chosen fields were processed for eligibility and completeness, reviewed by independent expert researchers (here: the Peer Circle members), and decided upon by an interdisciplinary selection committee, with one specialist committee member assigned to each application.

HFST committee members involved in the Peer Circle pilot were grouped together to meet in a session separate from the regular selection meeting. In addition to the four active specialist committee members for the four research fields involved, another member for each field was recruited to enable expert discussion within the otherwise interdisciplinary committee. With two members from each field, the Peer Circle group of the HFST committee comprised eight members in total. Thus, it resembled in size and interdisciplinarity the other, regular HFST committee groups.

Specialist committee members were asked to categorize their assigned applications only into clearly negative cases (A) and cases worth discussing (D). The committee voting system, the selection criteria, and the quality standards of the HFST program were to be applied without any changes. The success rate was aligned with that of the remaining applications in the HFST program, based on the total number of fellowships available in the selection round. After the selection meetings of the Peer Circle group and the remaining HFST groups, voting results were consolidated into one ranking for all applications.

3. The Peer Circle pilot

3.1 The Peer Circle process

The Peer Circle is a collaborative review procedure that takes place on an interactive online platform. The platform allows reviewers to access all relevant documents relating to applications from their research fields. For individual applications, access can be restricted for reviewers with potential conflicts of interest, according to the AvH impartiality guidelines. Peer Circle reviewers provide their assessments by posting comments directly to the applications. The members can see one another's comments, refer to them, ask questions, and so on. Specialist committee members can also access applications in their fields when the review phase starts and can, if they wish, ask the group of reviewers for opinions on certain aspects, clarifications and, according to their own specific expertise, participate in the review themselves.

To guarantee anonymity during discussions, unique, non-disclosed pseudonyms are assigned to all reviewers. During the review phase, the reviewers can only be identified by staff of the AvH. After completion of all reviews, reviewer access is disabled, and pseudonyms are deactivated, so that committee members also know reviewer identities, as is the case in the regular HFST procedure.

Peer Circle members can comment on any aspect of the application but are asked to answer key questions provided by the AvH as comments. These questions are based on HFST selection criteria and correspond to the key questions that are asked during the conventional review process. However, statements can be made in comparison to other applications of the same selection round. The grading that is requested in the conventional process for each question is only given with a reviewer's overall assessment, rating an application from 1 (below average), to 2 (average), to 3 (top 15 % worldwide), to 4 (top 5 % worldwide). As Peer Circle reviewers assess several applications at a time, they are encouraged to base their overall ranking on a comparison of the cohort.

To kickstart an application's assessment, the AvH appoints a Peer Circle member most closely matching the application's specialization as *first reviewer*. The aim of assigning someone specifically to initiate the review process is to facilitate the launch of the review. The AvH aims to distribute the number of times reviewers are asked to act as first reviewer evenly among the members of one Peer Circle. All Peer Circle members are invited to review all applications from their research area, and to discuss them with other reviewers. In general, they were also free to begin an application's assessment before the first reviewer assigned so as to accommodate personal schedules and preferences.

In rare cases of highly interdisciplinary applications, or applications with an unusual specialization not adequately covered by the Peer Circle, the AvH can appoint a member of another Peer Circle, or an external expert reviewer to also join in the review and, thus, close the existing gap.

During the whole review phase, two AvH staff members per Peer Circle monitor and moderate the process as necessary. Their tasks include answering any (formal) questions regarding, e.g., applications or selection criteria, asking reviewers to add to or clarify incomplete or indistinct statements, monitoring completeness (i.e., whether all key aspects

have been adequately assessed by the Peer Circle), and, if needed, appointing additional reviewers. At the end of the review phase, the research area specialist among the AvH staff adds a comment to summarize key aspects of the reviews, before inviting the assigned specialist committee member to add their final assessment and recommendation.

3.2 The pilot's timeframe and technical implementation

The Peer Circle procedure was piloted in the Humboldt Research Fellowship program (HFST) during two selection rounds: summer and fall 2022, ranging from March to November 2022. The Peer Circle committee meetings each took place one week before the regular HFST committee meeting: on June 29, and November 3, respectively. Deadlines (for eligibility checks, reviews, specialist committee member statements etc.) were based on regular program deadlines with slight adaptations.

With Peer Circle review phases aligned to those of the regular procedure, the overall duration of the phases during which reviewers were able to access and comment on applications was up to 11 weeks in the 2022 pilot. For each selection round, two interactive review phases of two weeks were defined, taking place directly after the start, and directly before the end of the whole review phase, respectively. The interactive review phases were intended to harmonize Peer Circle members' review activities in order to facilitate communication on the online platform by potentially reducing response times.

The Peer Circle procedure was tested using commercial online proofing software hosted in Germany. Implementation of the Peer Circle method was subject to the platform's technical constraints and therefore required compromise or workaround solutions for some details of the procedure, most of which did not affect Peer Circle members. Compromises on the following aspects were required:

- Pseudonyms: Reviewer pseudonyms, rather than being meaningful handles, were alphanumerical codes automatically assigned by the system.
- Tagging: Tagging users by name is a system feature used to notify these users of comments or questions directly addressed to them. During the pilot, it was not possible to tag users by pseudonym directly, so reviewers were not able to tag one another. They were able, however, to refer to the pseudonym, or tag AvH staff, who in turn could tag and thereby notify the intended Peer Circle member.

3.3 Summary: Peer review and Peer Circle review roles

Overall, the participants involved in both procedures are generally the same, although their roles differ in a few aspects as is summarized below.

Role	*In Peer Circle procedure*	*In conventional procedure*
Applicant	• Submits application	• Submits application
Independent peer reviewer	• Peer Circle member • Is appointed for an extended period • Reviews several applications per round • Acts as first reviewer for applications closest in specialization • Can read all Peer Circle reviewers' comments and interact • Cannot identify Peer Circle members	• Is appointed for a single application • Does not know other reviewers, or their assessments
Specialist committee member	• Can participate in the review and Peer Circle discussions (optional) • Gives final assessment and recommendation (rejection or discussion) for assigned applications • Presents assigned applications to the interdisciplinary committee • Knows the identity of all reviewers (after conclusion of the review phase)	• No participation in the review (unless close in specialization) • Gives final assessment and recommendation (rejection, discussion, or fellowship) for assigned applications • Presents assigned applications to the interdisciplinary committee • Knows the identity of all reviewers
AvH staff	• Appoints expert reviewers for an extended period (6-9 per Peer Circle) • Processes applications • Moderates Peer Circle in review phase • Summarizes review results • Assigns applications to specialist committee members	• Processes applications • Appoints expert reviewers individually (two per application) • Summarizes and assesses applications and review results • Assigns applications to specialist committee members

4. The evaluation approach

4.1 Aim and focus

The Peer Circle is a new way of reviewing grant applications by a group of reviewers, as outlined in Chapters 2 and 3. The objective of this report is to evaluate the Peer Circle (PC) in comparison to the conventional review (CR) process. The PC aims to solve the problem of engaging and retaining reviewers, which has become increasingly difficult. One of the main reasons that researchers turn down review invitations is a lack of time, so the PC should save time compared to the conventional review procedure. In addition, the PC is designed in such a way that less experienced researchers can also participate in the review process, increasing the number of potential reviewers considerably. Finally, the scientific community should accept the Peer Circle approach as an improvement over the conventional approach to peer review of grant applications. These goals should be realized without compromising on the quality of the review process.

One cannot estimate the long-term willingness to participate in the PC on the basis of a short-term experiment, but it proved relatively easy to attract PC members, and most of them continue for a second year. But one can find out whether PC members are positive about the Peer Circle review procedure, and whether they prefer it to the conventional review method. This leads to several questions for the evaluation of the Peer Circle process and its outcomes:

1. Are the resulting PC reviews of at least the same quality as conventional reviews?
2. Does the interaction within the PC result in premature consensus, hindering critical consideration of the applications?
3. Do the reviewers and committee members perceive the PC as a better alternative?
4. Is the PC more efficient than the conventional approach in the various aspects, such as the amount of time it takes?

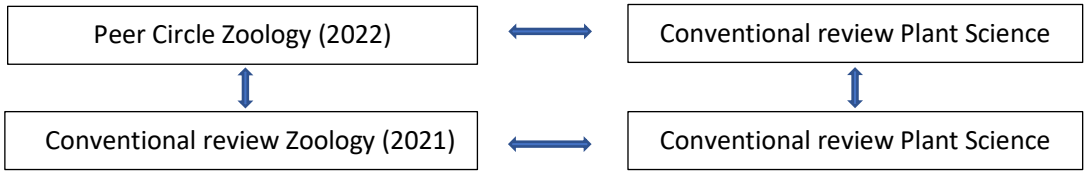
Finally, there are questions about the decision-making process after the PC, which may be affected by the different review procedures:

5. Does the PC committee select the best applicants?
6. Do the PC reviews influence the gender balance in the selection outcome?

4.2 Approach

To answer these questions, the Peer Circle was used for four different fields in two rounds in 2022. The PC will be compared with the conventional peer review process in the same fields in 2021, and with similar fields in 2022 and 2021 that use the conventional review process. Similarity refers here to the disciplinary culture, especially research style and evaluation style.

The following fields were selected for testing the Peer Circle: Inorganic Chemistry, Materials Science, Zoology (biodiversity), and Modern History. Four fields (control groups) were selected, as these are most similar to the experimental fields: Solid State Chemistry, Materials Engineering, Plant Science, and Ancient History. For the Peer Circle in Zoology, the comparison looks as follows:



No social science field was included. The study includes 325 cases and 28 Peer Circle members. The overview below shows the distribution of the cases over the eight fields and the two years.

	2021	2022
Experimental fields	Pre Peer Circle	Peer Circle
Inorganic Chemistry	29	20
Materials Science	13	19
Zoology (biodiversity)	24	30
Modern & Contemporary History	32	20
Control fields		
Solid State Chemistry	21	14
Materials Engineering	16	21
Plant Science	15	19
Ancient History	17	15

Two rounds: For several reasons, the study needed to have at least two rounds of Peer Circles. Firstly, more rounds with the same Peer Circle members avoids the situation where we only observe the effect of "newness", which may be biased in a positive or negative way. Secondly, more rounds also enable the Peer Circle members to get used to the tools and the way of working. This avoids that the results of the evaluation reflect the learning curve more than the "stable" use of the Peer Circle method. Thirdly, the number of applications per field is rather low in each round (10 to 15), so having more rounds of the Peer Circle creates a larger dataset that allows for more robust results. The decision was to use two rounds: the summer and fall rounds of 2022.

Field differences: The evaluation included four different fields, as the use and acceptance of new forms of peer review may differ between the disciplinary cultures. We will check whether there are any striking differences between the fields in the interviews, surveys, and other data that would indicate such differences.

4.3 Data collection methods

Several data collection methods were used, with some focusing on opinions of the Peer Circle members, the committee members, and AvH staff (interviews, surveys), and others involving behavioral data (logfiles, observations, written comments/reviews, bibliometric data).

Observations: The Peer Circle process was observed using the *logfiles* of the online platform. These data show the number of interventions by AvH staff, the number of contributions by PC members and committee members, the level of mutual interaction, and the distribution of the activities over time. The time spent on the review process can be roughly estimated from the logfiles.

The two PC committee meetings and the two regular HFST committee meetings were observed to find out whether there were any differences in the way the meetings proceeded, which may be related to the different review models, namely the PC and the conventional approach.

Interviews: Peer Circle members, the involved committee members, and a selection of AvH staff members were interviewed about their experiences with the Peer Circle, about the associated advantages and the disadvantages, and about whether they are satisfied with the new model compared to the conventional form of peer review, in terms of the process and the quality of the assessments. The interviewees were also asked to compare their Peer Circle experiences with experiences they have had with conventional grant peer review. It is important to note that the conventional peer review method for the AvH-HFST is more structured than is generally the case, as the description in Chapter 2 illustrates. In other words, it has a question-answer format. However, the interviewees and the survey respondents compare the PC with the common forms of peer review, where the reviewers can chose their own format for the review.

We interviewed half of the 28 reviewers after the first round and the other half after the second round. We interviewed the eight committee members and the four subject group managers twice (after each round), and four case handlers after the first round. In total 56 interviews were held. A list of topics was drawn up to guide the interviews. Committee members were also involved in a group discussion after the committee meeting in the first round. AvH staff members were interviewed about their experiences with the Peer Circle compared to the conventional approach, i.e., what went well, what problems came up?

Survey: The results from the first round of interviews were used to draw up the survey, which covered the same topics as the interviews. In the first round, we surveyed those PC members who were not interviewed in that round, but in the second round of the Peer Circle we surveyed all members with a slightly altered questionnaire. The response was 25 (out of 28). For those who participated in the survey twice, we were able to compare survey results. This provides some information about potentially changing opinions as a result of becoming familiar with the PC.

Reviews and comments: The review reports of PC members consist of answers to a list of questions about the application, plus comments on the reviews provided by other PC members. The length, the level of detail, and the coverage of evaluation dimensions are compared with the conventional review reports. The comparison is done using the interviews and the surveys (see above) and by a *text analysis* of the reviews, which may give a more objective analysis of the content and style of the texts. More specifically, the words used in the reviews are assigned to linguistic categories, and the frequency of use of these linguistic categories enables the relevant characteristics of the review texts to be identified, such as writing style and focus of the review.[2]

The decision data: The outcome of the selection process is a rank order of the applicants, and a group of selected applicants and a group of non-selected applicants. Since publication productivity and impact of journal articles are accepted selection criteria in many of the natural sciences, one may expect that the successful applicants *on average* perform better in

[2] For more details see 5.2.

terms of publications and impact. The evaluation of the applicants is not only based on the applicants' past publication performance, but also on the quality of the proposal, on the CV, on mobility, and on the future potential of the applicants. But one would still expect a positive correlation between past performance and the perceived quality of the applicant.[3] We collected *bibliometric data,* calculated the indicators[4], and compare the average past performance levels of the rejected and accepted applications. For the Chemistry fields, this is done for four groups: the Inorganic Chemistry Peer Circle and the three control groups Inorganic Chemistry 2021, and Solid State Chemistry 2021 and 2022. The aim is to give an initial insight into the quality of the applicants and the selected applications, and to assess the gender balance in the outcome.

4.4 Data

To summarize, the following datasets were used:

1. Coded interviews. These include opinions on a range of items, and self-reported behaviors. Added to this are data from a group discussion report of the committee members after the first round.
2. Results of the survey among the PC members, covering self-reported behavior and opinions on a series of items.
3. Logfile dataset: Number of logins, time online, time active, number and date of review statements and of comments.
4. Observational data from the four observed committee meetings: Time spent for each application, number of discussants for each application.
5. Review dataset: Textual data from the reviews and comments for each reviewer and each review dimension (the four key questions; the summary) plus data about the scores given by the reviewers.

These datasets were enriched with data about the applicants:

6. The different committee scores the applicants received, and the decision (from the administrative file with the rank order that resulted from the voting in the committees).
7. Personal characteristics (from the application files): Age, nationality, gender, residence, institutional affiliation, host, current position.
8. Bibliometric data measuring productivity and impact of the applicants in Chemistry.

4.5 Terminology

The following types of participants in the Peer Circles can be distinguished:

(i) The reviewers, in German the *Fachgutachter*. For this group we use the term *PC members,* or *reviewers.*
(ii) The specialist committee members, in German the *Fachvertreter*. For these we use the term *committee members*.

[3] For an example of a comprehensive analysis: Van den Besselaar P & Mom C (preprint 2020), *Gender bias and grant allocation – a mixed picture*.

[4] We use the number of publications, the fractionally counted number of publications, total source normalized impact per paper, and the number of top-cited papers (top 1% and top 5%). Fractional counting means that the number of co-authors is taken into account: A paper with four authors counts for ¼. For more details see 6.2.

(iii) Staff members of the Alexander von Humboldt Foundation, the *subject group managers* (in German the *Fachgruppenleiter*) and the *case handlers* (the *Sachbearbeiter*). When referring to both groups together, the term *AvH staff* is used.

Finally, when using personal pronouns, these are intended to include all gender identities.

Part II – Evaluation results

5. Review quality

What counts as a high-quality grant review depends on the aims of the grant scheme. As there are very different types of grants, the criteria to apply are also different. Nevertheless, in conventional grant *peer* review, one generally searches for reviewers who have core expertise in the specific topic of the grant application. The PC members raised the issue of appropriate criteria during the interviews. Having a topical specialist reviewing an application focuses the review often on the content of the research plan, more than on other relevant aspects. The question is then whether detailed technical comments – which may be useful as feedback – are also very relevant for the grant selection process. This issue is relevant here because the emphasis in the HFST program is more on the achievements and the future potential of the applicant, rather than on the specific project proposed, which should be good. It was regularly argued that the PC may be a better way of evaluating HFST applications, as the broader PC would more naturally focus on other aspects than the technical content of the proposed project.

To what extent do PC reviews differ from conventional reviews? It is worth pointing out that the quality of conventional reviews varies considerably and therefore "the" quality of conventional reviews cannot be a standard by which to evaluate PC reviews. We therefore focus on the opinions of participants in the PC, who were interviewed and surveyed about the quality of the PC reviews. Since most PC members have been involved in reviews in different contexts, they are a useful source for making a comparison between the PC and conventional reviews. To this is added a textual analysis of the reviews, as a first step to explore how it can support the evaluation of the review procedures and quality.

5.1 Comprehensiveness

In a conventional review, reviewers should address all aspects of the application. In the PC, one or two members whose expertise is close to an application are designated as *first reviewer* for that application, and they are expected to address all evaluation dimensions. The other PC members can add their comments on parts of the application and can react to comments of the other members. In this way, the PC evaluation should be comprehensive. Indeed, the PC members feel that together the answers to the review questions and the comments add up to a comprehensive review, and only two out of the 23 PC members who answered this question were slightly negative (Fig. 1).

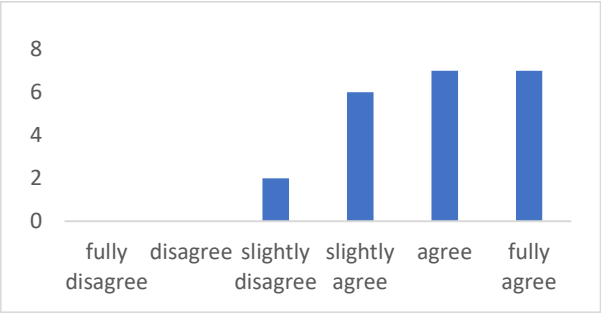


Figure 1. Do the comments together form a comprehensive review?
N= 23

The interviews with AvH staff members also indicate this positive view. In the conventional review model, two reviewers are selected who are as close as possible to the proposed project. In the PC this is not the case, as the PC members are selected before the applications come in. This may affect the comprehensiveness of the PC review, as expertise may be lacking in the PC. The interviews show that PC members see this as a potential issue, but at the same time they report that this problem did not emerge.

- Firstly, in a few cases an external reviewer was included in the PC because of specific expertise, and this worked well.
- Secondly, the quality of the applicant, their CV and future perspectives are the main criteria in the HFST program, and not only or mainly the research project. As interviewees emphasized, when reviewers are very close to the research topic, there is a risk that they will focus too much on technical details, which are often not very relevant for the decision-making process.
- Thirdly, the PC members together create a large pool of collective knowledge. This not only relates to the technical aspects of the research plan, but also to the countries of origin of the applicants. In a PC, the probability is much larger that one of the members will have some experience in the country of origin and is therefore better able to assess the CV and the career of the applicants in the local context.
- Finally, some of the reviewers and committee members interviewed pointed out that the distance from the technical content of an application is not an issue specific for the PC, as under the conventional approach "scientific proximity" is also not guaranteed for the reviewers.

5.2 Level of depth and detail

In the interviews, many reviewers expressed the opinion that the level of depth and detail was at least similar to the conventional peer review. That may not be the case for all comments individually, but the larger number of comments compensates for that. And the more people who view an application, the more objective the evaluation becomes and the better the overall judgement.

Level of depth and detail may relate to the length of the reviews, which differs between the conventional review and the PC review. In Table 1, the length of Peer Circle and conventional reviews is compared.[5] The review texts relating to an application were combined, and are restricted here to the texts addressing the CV, the core publications, the proposed research project and future potential. Table 1 shows the averages and the coefficient of variance (CoV), a measure for the dispersion.

Overall, conventional reviews are longer than the PC output, and interviewees suggested that this may be because several of the standard parts of the conventional review are lacking here, such as a summary of the proposed project and courtesy formulations. Some interviewees mentioned that conventional reviews are becoming shorter too.

The expectation was that the PC would focus less on the project proposal than in a conventional review, so one would expect that in the PC a smaller part of the text would be devoted to the planned project than in a conventional review. This, however, is not the case.

[5] We do not differentiate here between the two rounds, as the length was about the same in the two rounds.

In the conventional reviews, 34% of the text is devoted to the project proposal, in the Peer Circle this is 41%. One explanation for this could be that in the PC, many words are used in order to explain the project to those PC members who are less familiar with the topic.

Furthermore, the PC devotes fewer words to the core publications, and about the same share of the text on discussing the CV and future potential.

Table 1. Length of the review texts[6]

		Total*	CV	Core publications	Project proposal	Future potential
Conventional	Average	1,253	347	292	422	192
	CoV**		0.48	0.59	0.52	0.52
Peer Circle	Average	751	200	141	311	99
	CoV		0.67	0.85	0.59	0.70
Ratio	Conventional/PC	1.67	1.74	2.07	1.36	1.94

* Sum of the number of words in the four core parts of the review. We did not include the summary texts and the profile texts, as these do not exist in the PC procedure.
** Coefficient of variance

Some interviewees mentioned that the *writing style*[7] in the PC reviews and comments is more informal than in conventional reviews, including the structured form of the conventional HFST review. A first analysis of the review texts confirms this:

- Reviews normally have a strong *analytical style* (measured on a scale from 0 to 100) as opposed to a more narrative style, and this also holds for the reviews in the HFST case. As expected, the PC reviews score somewhat lower on analytical writing: 87 versus 97, and the difference is significant: $F(1, 322) = 89.87$, $p < .000$.
- Another difference is the score on *clout* where the higher the score, the higher the reviewer emphasizes expertise and confidence. The PC scores substantially lower, 46 versus 61, and this is also significant: $F(1, 322) = 159.30$, $p < .000$.
- The opposite is the case for the score on *authentic*, measuring an open versus a guarded text: 48 for the PC versus 26 for the conventional reviews: $F(1, 322) = 99.14$, $p < .000$.

Obviously, the evaluation in the PC uses a different writing style. As one interviewee noted, "the comments are open and honest", and another called it "a less curated review, less effort to make a coherent argument that leads to the selected verdict".

Another question that can be tentatively answered by analyzing the review texts is whether conventional reviews have similar emphasis on the various evaluation criteria as in the Peer Circle. Using a dictionary for the terms that relate to the evaluation criteria (related to career, mobility, independence, publications, school and university performance, the project, quality of the host, and excellence), we compared the review texts of the PC with those of the

[6] In our comparison fields, using the conventional peer review approach, there were 235 applicants and about 100 had only one external reviewer. This illustrates the problem of finding enough reviewers. In some of those cases, when the committee member had the appropriate specialization, they were appointed as second reviewer. For the other cases we used the statements of the committee member for the analysis, even if they were not a formal reviewer. The summaries by AvH staff were not used in the text analysis.

[7] Van den Besselaar P, Sandström U, Schiffbaenker H (2018), Using linguistic analysis of peer review reports to study panel processes. *Scientometrics* 117, 313-329; Van den Besselaar P, & Mom C (2022), The effect of writing style on success in grant applications. *Journal of Informetrics* 16(2):101257

conventional review approach. There are some differences, as Table 2 shows. First of all, the PC reviews have a significantly higher share of *common words*, indicating that there are less technical terms in those reviews. This is in line with the expectation that the PC would focus less on the technical content, and more on the other evaluation criteria, something that came out of the interviews. This is also consistent with the observation that the conventional review reports contain significantly more project-related words. In the conventional review, the reports focus more on performance than in the PC review, but the PC reviews focus more on the host and the independence of the applicant.

One of the criteria is the quality of key publications. Interviews with several PC members indicate that they scan the core publications rather than read them, and subject group managers have also raised this issue. Publications are discussed more in terms of numbers, journal impact, and citations than in terms of content and findings. Is this different from the conventional review process where specialist reviewers are selected for each application? Our data suggest not, and the frequency of the use of "bibliometric" terms in conventional reviews is similar to that in PC reviews (Table 2). Observation of discussions in the selection committees suggests the same, and there too the quantitative properties of the publication lists are mentioned often.

Table 2. Differences in evaluation emphasis between the PC review and the conventional review

	Peer Circle review	Conventional review	Oneway ANOVA
Common (non-technical) words	80.0%	76.0%	$F(1, 322) = 46.03$, $p = .0000$
Career (incl. mobility)	0.93%	0.96%	n.s.
Performance (incl. school & university)	0.60%	0.72%	$F(1, 322) = 5.03$, $p = .0257$
Publications (bibliometrics)	1.26%	1.17%	n.s.
Proposed project	1.04%	1.24%	$F(1, 322) = 9.14$, $p = .0027$
Independence	0.07%	0.02%	$F(1,322) = 31.52$, $p = .0000$
Excellence	1.39%	1.47%	n.s.
Host	0.23%	0.17%	$F(1, 322) = 7.00$, $p=.0086$
Final score by the committee	0.55%	0.58%	n.s.

n.s = non-significant

In general, the interviewees feel that the PC reviews have good depth and scope, and that only the technical content of the applications is covered more in the conventional reviews. Several interviewees welcome the reduced focus on technical content, as it prevents that technical details become too important in the review and the selection process. This is supported by the survey, as almost all PC members tend to find the overall review result of high quality (Fig. 2), and furthermore, 17 out of 20 find it at least as good as the conventional review procedure (Fig. 3).

Ten PC members consider their own contributions to be (slightly) worse than in the conventional procedure, while only seven consider them to be (somewhat) superior (Fig. 4). It should be noted that six PC members did not answer this question, probably because they had little previous review experience and are therefore not able to compare. However, as one interviewee stated, the PC comments together may still be better ("crowd intelligence") even if the individual assessment contributions are not as good as in the conventional procedure.

To summarize, PC members view the reviews/comments to be on average at least at the level of the conventional procedure. This opinion is also shared by the subject group managers, who are also positive about scope, depth, and detail. The same is true for most committee members, who share the opinions of the reviewers.

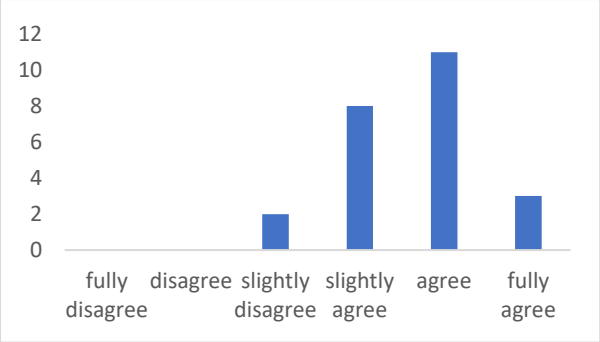


Figure 2. The comments and assessments are of high quality
N = 24

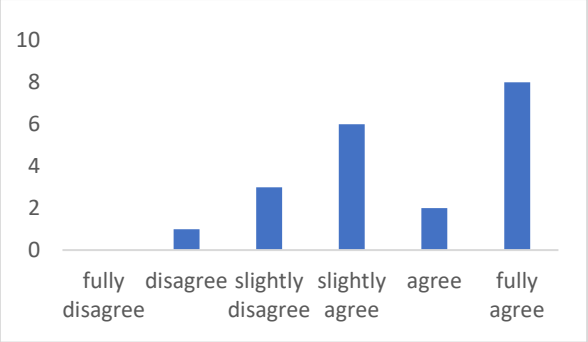


Figure 3. The quality of the PC evaluation is at least equivalent to the conventional peer review
N = 21

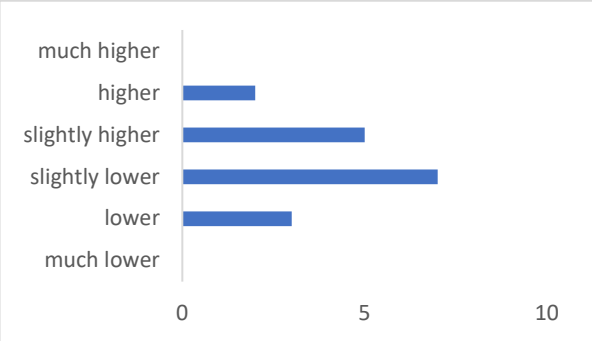


Figure 4. How is the quality of my own contributions compared to a conventional review?
N = 17

Does experience matter?

Another issue is the heterogeneity in age of the PC members in relation to the quality of the reviews. A variety of opinions were expressed in the interviews, with one interviewee stating, for example, that only senior researchers should prepare reviews. Some other interviewees felt that younger researchers were somewhat more insecure than experienced researchers. But the prevailing opinion is that the number and quality of review contributions are not related to age because in all age groups there are more and less active reviewers, long and short contributions, and detailed and general contributions. The logfiles support this: We calculated the correlation between age of the reviewers and the time active on the online platform, and that correlation was about zero. The correlation between reviewer's age and the number of comments was low and not statistically significant ($r = 0.19$; $p = 0.34$).

5.3 Number of reviewers

A notable feature of the PC is that a much larger group of peers contributes to the review of the applications than the two or one reviewers in the conventional procedure. This was seen almost uniformly by the interviewees as a major advantage. The following advantages were mentioned:

- In this way, the review process becomes more transparent, objective, and less biased, which may lead to a better decision and may have a positive effect on the acceptance of the outcomes.

- It makes the assessment less dependent on selection of the reviewers.
- It brings more expertise into the assessment. For example, in the funding schemes of the AvH applicants come from many countries and assessing their quality and career prospects (using the CV) requires knowledge about the country the applicants come from. This knowledge is more likely to be available in a 10-person PC than with two reviewers only.
- The larger number of reviewers also leads to sufficient coverage of the research topics that feature in the applications.
- It provides a self-correction mechanism: Less likely that wrong assessments go unnoticed.
- Some committee members mentioned that having comments from a larger set of reviewers increases the confidence one can have in the result of the review procedure.

5.4 Grading

Several interviewees stated that in the conventional review grades tend to be high even when the review texts point out deficiencies because reviewers generally strive to support applicants. The way the Peer Circle is organized encourages *comparative* grading, which would tend toward more differentiated and lower average grades. This would be a potential improvement in the decision-making process, but as we will discuss in 8.2, the Peer Circle currently lacks a systematic grading procedure.

6. Decision-making and selection results

The previous chapter focused on the quality of the reviews produced in the Peer Circle. This chapter focuses on the post-PC phase, where committee members formulate a proposal for the committee, followed by committee discussion and decision.

6.1 Preparation for decision-making

The role of committee members varied from PC to PC and person to person. Three of them began their work after the Peer Circles had finished reviewing and commenting, and limited themselves to the typical committee tasks. The other five also acted as reviewers, with some of them even emphasizing that they read and reviewed *all* applications. This is not dissimilar to the conventional review approach, where in many cases it proved impossible to find two external reviewers. In a minority of those cases – when the application was close to the committee member's expertise – they were appointed as second reviewer.[8]

Most committee members described in the interviews that they depend on the results of the Peer Circle. Many of the arguments relate to the fact that, in the Peer Circle, more reviewers look at the applications:

- In the PC it is easier to follow the arguments of reviewers (compared to the conventional review procedures), and then rank the proposals.
- In cases where a comment is not clear, one can ask the PC members for clarification, which is not so easy in the conventional procedure.
- The comments made in the PC make it easy to find the critical points in the applications.
- When there is consensus about funding or about rejecting, then that verdict can be accepted. The focus can then be only on the cases with contradicting assessments in the Peer Circle, and those that fall in between reject and accept.
- The problem remains in the grey area, where selection objectivity is not achievable. Then the PC works better because more people review.
- More views lead to more certainty for the committee member: "If I am uncertain about something, the PC comments are very helpful to shape my opinion", and "The PC reviews make me more certain about my own ranking".
- The conventional procedure suffers from reviews that are handed in too late for decision-making, meaning that applications are rescheduled to the next round. The Peer Circle solves this problem because it is based on group activity which delivers even if one of the members does not deliver.

One or two committee members have different opinions about how the review process in the Peer Circle contributes to their tasks. Their arguments are:

- The reviews and comments of the non-specialists do not add much.
- Converting the PC comments into a proposal is more work (two respondents), but they may at least lead to slightly better decisions (one respondent).

[8] When the committee member finds the arguments of the single review clear and complete enough to make a decision about the applicant, the usual statement is written for the committee meeting.

- The applications and the proposed ranking must be presented in an interdisciplinary committee, which requires different types of arguments then those generated by the circle.

One of the committee members suggested it might be a good idea for a committee member to share their final ranking with the Peer Circle members and ask for feedback. In that way, the PC would be more directly involved in translating the review results into the proposal for the decision-making process. Another committee member had the opposite view, and emphasized examples where they contradicted the PC consensus. They emphasized their independence from the Peer Circle, describing committee members as the "reviewer of the reviewers".

A few of the committee members interviewed reflected on the final results, stating that the outcome would not have been different if the applications had been reviewed using the conventional approach. But did the different preparation influence decision-making? Does the Peer Circle lead to more agreement and greater certainty (as some committee members suggested) and thus to fewer committee discussions[9], improving the efficiency of the process? While observing the committee meetings, one of the things recorded was the number of discussants for each of the discussed applications. About 31% of the applications presented in the conventional committee were not further discussed, and 37% in the PC committee. The striking difference is the long tail in the conventional approach (Fig. 5).

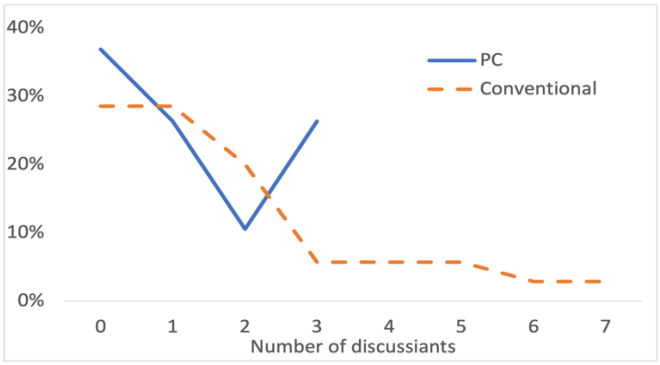


Figure 5. Discussants per application (summer round)

The average number of discussants per application was lower in the PC than with the conventional approach (1.26 and 1.74 respectively). The two committees were different in size (8 versus 13 members) and more participants means more potential discussants. However, participants in larger meetings may be less inclined to contribute to the discussion, but not all members were present, and not all members were present throughout the whole meeting. There were many more applications to discuss in the conventional committee, creating time constraints, which also may prevent participants from engaging in the discussions. These effects may offset each other, and one can tentatively conclude that applications in the PC committee lead to less discussion on average than in the conventional committees. We therefore tend to conclude that the PC review procedure can reduce uncertainty, resulting in committee meetings that are at least equally efficient.

[9] Earlier research has shown that if there is disagreement in a group, e.g., the quality of a grant application, there is more discussion about it (Festinger, Informal social communication. In: *Psychological Review* 57 (1950) 271-280).

6.2 The success rates

The average success rate in both the experimental and control fields was much lower in 2022 than in 2021 (Table 3). The table also shows that the average success rate in the PC in 2022 is slightly higher than in the control fields (23.6% versus 22.7%). The lower success rates in 2022 are at least partly due to the decline in the overall ratio of total funding to total applications.

Table 3. Success rate in PC fields and conventional reviewed fields, pre-PC and PC

	2021 pre-PC success rate	2022 PC period success rate
PC fields	28.9%	23.6%
Control fields	33.3%	22.7%
All other fields	33.7%	27.6%

N= 321

Why is the decline stronger in the control fields than in the experimental fields? For final decisions, the rankings from all committees are combined. So, even if the average score for a field remains the same, the success rate will decline if other fields show on average increasing scores. Table 4 shows the average field scores for the four conditions. In the PC fields, the average committee score remains the same (0.54). In the control fields, the average score declined considerably by 18% (from 0.67 to 0.55).

Table 4. Average committee score pre-PC and PC

	2021 pre-PC committee score			2022 PC period committee score		
	N	Mean	Std. Dev.	N	Mean	Std. Dev.
PC fields	97	0.54	0.79	89	0.54	0.72
Control fields	69	0.67	0.84	70	0.55	0.78
All other fields	926	0.65	0.83	822	0.61	0.81

The S cases were given a score of 2 and the A cases a score of 0.

Since the average scores of the PC fields and of all other fields remain about the same, it is not the scores but funding levels that explain the decline in the success rate for the PC fields. In the control fields, the mean score decreased significantly, indicating that the success rate of those fields decreased much more than the success rate of the experimental fields.

6.3 Quality of the applications

To measure and compare the quality of applications, more objective data are needed in addition to committees' scores on the assessment criteria. Various factual data may be derived by combining the CV with bibliometric data, such as about mobility, the quality of the applicants' network, and their career, but this is beyond the scope of the study. Here we are limiting ourselves to the past research performance of the applicants because interviews and the observations of the committee discussions clearly showed that productivity and impact (status) of the journals do matter. Bibliometric data were collected, and performance indicators were calculated. Productivity is measured in two ways:

- The number of published papers in international journals included in the Scopus database.
- The number of fractionally counted papers. The latter indicator takes the number of co-authors into account: An author only gets a fraction of each paper, and the fraction

is calculated by dividing a paper equally among the authors. If an applicant has a paper with 5 co-authors, only 1/6 of that paper is assigned to the applicant.

The impact of an applicant's papers is measured by:

- The share and number of top 1% highly cited papers, normalized by field, by publication type, and by year of publication.
- The share and number of top 5% highly cited papers, normalized in the same way.
- The sum of the SNIP scores. SNIP score measures the impact of a journal, so each paper is given that score, and for calculating the total impact of the applicant the SNIP scores are added together.

We are limiting the analysis to the two chemistry fields for time reasons. Chemistry is a good choice for this because journal publications are the standard form of communication within chemistry. We are not taking senior applicants into account because there are too few of them in the sample. Also the number of postdoc applicants is not high and, therefore, findings should be interpreted with caution.[10] Table 5 shows the results for the two chemistry fields in 2021 and 2022.

Table 5. Average scores on bibliometric performance indicators, two chemistry fields*

	2021 inorganic chemistry (N=28)			2021 solid state chemistry (N=18)		
	Mean	Median	Std. Dev.	Mean	Median	Std. Dev.
full_P	19.9	13.5	29.7	16.6	17.5	6.4
frac_P	3.5	2.8	4.1	2.8	2.6	1.4
Sum Snip	28.5	15.8	50.1	29.4	26.8	18.9
Share top 1%	2.9%	0.0%	0.1	6.0%	0.0%	0.1
Share top 5%	14.2%	14.0%	0.1	20.6%	11.8%	0.2
	2022 inorganic chemistry (N=17)			2022 solid state chemistry (N=9)		
	Mean	Median	Std. Dev.	Mean	Median	Std. Dev.
full_P	18.4	9	18.6	18.67	14	13.5
frac_P	3.3	1.8	3.8	3.6	2.6	2.1
Sum Snip	22.1	12.3	27.2	18.5	13.9	13.1
Share top 1%	1.9%	0.0%	0.1	1.7%	0.0%	0.0
Share top 5%	11.3%	7.7%	0.1	16.0%	10.0%	0.2

* Only the postdocs

The table shows that the applicants score quite well on average, which is especially visible in the scores for the share of top 1% and top 5% highly cited papers. These are substantially higher than the expected value (which is of course 1% and 5% respectively). For example, the applicants in 2021 in Inorganic chemistry (the pre-PC set) have on average 2.9% papers in the top 1% cited class, and 14.2% in the top 5% cited class, which is for both about three times more than expected. The average scores on most of the performance indicators for Inorganic Chemistry are somewhat lower in 2022 than in 2021, and this also applies to (for the impact indicators) Solid State Chemistry.

[10] We include here granted and non-granted applicants in these two fields in 2021 and 2022. The numbers therefore cannot easily be compared to the bibliometric data presented in the recent evaluation of the HFST program, as that study only includes granted applicants (Geyer, Grasenick, Kleinberger-Pierer, Gorriaz (2021), *Evaluation des Humboldt-Forschungsstipendien-Programms (HFST) der Alexander von Humboldt- Stiftung*. Endbericht. Wien: Inspire).

The committee scores show a similar pattern because the average score for Inorganic Chemistry declined from 0.59 to 0.45, suggesting an overall lower quality of the applications in 2022. The average score for Solid State Chemistry drops about the same (from 0.81 to 0.63) and the standard deviations indicate for both fields quite a large variation. The lower average bibliometric performance is reflected in the lower average committee scores, so the findings indicate that the committee scores are quite supported by the more objective indicators.

Table 6. Average committee scores for Inorganic Chemistry and Solid State Chemistry

	2021 pre-PC mean committee score		2022 post-PC mean committee score	
	mean	standard deviation	mean	standard deviation
Inorganic chemistry	0.59	0.80	0.45	0.69
Solid state chemistry	0.84	0.91	0.63	0.95

The S cases were given a score of 2 and the A cases a score 0.
N= 72. Only postdocs

The review reports and the committee meetings suggest that publication output is an important factor in assessing applications. Terms such as "length of publication list", "top papers", and "top journals" are used frequently, more often than terms referring to almost all other assessment dimensions (Table 2). Although bibliometric indicators are not used in the review process, one would expect the granted applicants on average to have a higher score on those indicators than the rejected applications.[11] The term average is important since the use of these indicators at the individual level is disputed, but they are useful at the group level.[12]

Do those who receive the fellowship perform better on average than those who are not successful? Table 7 shows where the granted applicants score higher than non-granted applicants, where it is the other way around, and where the two groups score about equal. The table shows a mixed picture, which again may have been influenced by the low numbers. In Inorganic Chemistry in 2021, the non-granted applicants have a better score on five out of ten indicators, the granted have a better score on three indicators, and on two indicators both groups score about equal. In the Peer Circle (Inorganic chemistry 2022), the grantees score better on five indicators, the non-grantees on three, and again the scores are similar on two. Interestingly, in the PC (2022), the grantees score lower on the publication counts and on journal impact, but higher on the numbers of top cited papers. The findings for Solid State Chemistry are somewhat different: In 2021, the granted applicants score better on all ten indicators, but in 2022 the picture is the same as for Inorganic Chemistry. These findings suggest that the PC outcomes are not different from the other outcomes.

[11] As we are using all rejected applicants here, the average of the group of rejected applicants can be low due to some very low scoring applicants. If one were to restrict the analysis to the better performing rejected applicants only, the picture may become different. See e.g., Peter Van den Besselaar, Loet Leydesdorff, Past performance, peer review, and project selection, *Research Evaluation* 18 (2009) Issue 4, 273-288.

[12] Performance indicators also vary strikingly for applicants from different countries/regions. Due to the small numbers, we cannot include such a differentiation here.

Table 7. Comparing the bibliometric scores for the granted applicants and the rejected applicants, two chemistry fields

		2021		2022	
		Median	Average	Median	Average
Inorganic chemistry	# Publications	Granted*	Rejected	Equal	Rejected
	# Fractional publ[&]	Equal	Rejected	Granted	Rejected
	# Sum SNIP	Granted	Rejected	Granted	Rejected
	# Top 5% highly cited	Granted	Rejected	Granted	Granted
	# Top 1% highly cited	Equal	Rejected	Equal	Granted
	# Granted	3**		5	
	# Neutral	2		2	
	# Rejected	5		3	
Solid state chemistry	# Publications	Granted	Granted	Granted	Equal
	# Fractional Publ[&]	Granted	Granted	Rejected	Rejected
	# Sum Snip	Granted	Granted	Granted	Granted
	# Top 5% highly cited	Granted	Granted	Granted	Rejected
	# Top 1% highly cited	Granted	Granted	Equal	Granted
	# Granted	10		5	
	# Neutral	0		2	
	# Rejected	0		3	

N = 72. Only the postdoc applicants.

N per group (IC = Inorganic Chemistry; SC = Solid State Chemistry);

IC 2021 granted = 9; IC 2021 rejected = 19; IC 2022 granted = 3; IC 2022 rejected = 14;

SC 2021 granted = 8; SC 2021 rejected = 10; SC 2022 granted = 3; SC 2024 rejected = 6.

* These cells indicate which group scores higher. Equal means about the same score.

** In Inorganic Chemistry in 2021, the granted applicants scored better on three indicators than the rejected applicants, five times scored worse, and two times about equal.

[&] Fractionally counted publications = if a publication has five authors, an author gets 20% of that publication

In summary: (i) the overall quality of the chemistry applicants is high, although somewhat less so in 2022 than in 2021; (ii) the average scores that applicants received from the committees reflect the average bibliometric scores; (iii) the bibliometric scores of the granted applicants are not overall higher than those of the non-granted applicants, which is unexpected. This indicates that productivity, impact, and the quality of journals seem less important in the decision-making process than the review reports and the observations of the committee meetings suggest. It would be good to repeat this analysis with more Peer Circles and control groups, to overcome the limitations of the low number of cases.

6.4 Diversity of selected applications

An important question is whether the way in which the selection process is organized affects the outcome in terms of gender (in)equality and gender bias. Gender inequality is defined as a different success rate, while bias occurs when such a difference is not based on merit. We can present the success rates by gender but cannot control whether these are based on differences in past performance. This is because the numbers of cases becomes too small: In the Inorganic Chemistry Peer Circle (2022) there are only six women; in the Inorganic Chemistry group in 2021, and in the Solid State groups in 2022 and 2021 there are eight, eight, and three women respectively.

Overall, men and women have the same probability of receiving an HFST grant and for both genders the overall success rate in our sample of 321 applicants is 27%. Broken down by year

and review model, we find that in 2021 the PC fields show a much higher success rate for men in 2021, which changed to a slightly higher success rate for women in 2022. The control fields show a much higher success rate for women in 2021, and a more modest difference in success rates for 2022 (Table 8).

Table 8. Success rates for men and women in 2021 and 2022, two chemistry fields

		2021 pre-PC success rate	2022 post-PC success rate
PC fields	women	20.0%	24.1%
	men	32.8%	23.3%
	total	28.9%	23.6%
Control fields	women	41.7%	25.0%
	men	28.9%	21.4%
	total	33.3%	22.7%

N=321

These are aggregated success rates, with very different underlying success rates at the field level. Table 9 –only postdocs – shows that the fluctuations are large, which is also because one more or one less successful woman can change the imbalance to the other direction. In six out of eight fields, the imbalance changes between the two years, with women underrepresented among grantees in one year and equal or overrepresented in the other year. Only in Modern History (women underrepresented) and in Materials Engineering (women overrepresented) did the gender imbalance not change. It would be interesting to compare the gender differences at the field level with performance differences, as that would say more about the (absence) of bias. But that is beyond the scope of this report.

Table 9. Success rate by gender, field and year*

	2021		2022		Total	
	N	share	N	Share	N	share
Women						
Inorganic Chemistry	8	12.5%	6	33.3%	14	21.4%
Materials Science	2	50.0%	4	25.0%	6	33.3%
Modern History	3	0.0%	2	0.0%	5	0.0%
Zoology	11	18.2%	10	20.0%	21	19.0%
Ancient History	4	50.0%	3	33.3%	7	42.9%
Plant Science	7	57.1%	6	16.7%	13	38.5%
Solid State Chemistry	8	37.5%	3	33.3%	11	36.4%
Materials Engineering	3	33.3%	4	25.0%	7	28.6%
Men						
Inorganic Chemistry	20	40.0%	11	9.1%	31	29.0%
Materials Science	9	33.3%	13	30.8%	22	31.8%
Modern History	16	31.3%	8	37.5%	24	33.3%
Zoology	11	45.5%	14	21.4%	25	32.0%
Ancient History	10	20.0%	5	40.0%	15	26.7%
Plant Science	7	14.3%	7	28.6%	14	21.4%
Solid State Chemistry	10	50.0%	6	33.3%	16	43.8%
Materials Engineering	10	20.0%	11	0.0%	21	9.5%

* Only postdocs.

Red/blue: Women have a lower/higher success rate than men.

To summarize, these findings suggest that the gender balance at the field level is in flux but that these changes disappear at higher levels of aggregation and lead to an overall balanced picture. A caveat is that the numbers are small, and therefore the observed inequalities can as easily be the effect of random variation as of gender bias. As is often the case, more data would be the solution. Specific conclusions about the Peer Circle cannot be drawn.

7. Reviewer activity

7.1 Field differences

As reviewer activities and styles may be field dependent, we first address the question whether there are field differences in the way in which the PC is appreciated. The average scores on the various survey items do not differ much between the four experimental fields. Only for a few items is the difference between the highest and lowest score more than 40%, and in all but one of those cases it is *Materials Science* that has a much lower score than the other fields. All these items relate to how the PC worked (see 8.1) and the field differences are small for the more evaluative survey items relating to the quality of the Peer Circle. The number of PC members is restricted, and one should not generalize the findings beyond the sample.

7.2 Peer Circle activities over time

We analyzed the logfiles of the online platform to map the pattern and intensity of activities in the PC. Table 8 shows the level of review activities in each of the Peer Circles in the first evaluation round (summer 2022) and the second evaluation round (fall 2022). In the summer round, reviewers on average contributed 35 comments, the committee members 25, and AvH staff members around 45[13]. The differences between the participants are large, ranging from only a few comments to more than 70.

Table 10. Number of comments by field and role

	Average		Per application		Minimum		Maximum	
	Summer	Fall	Summer	Fall	Summer	Fall	Summer	Fall
Inorganic Chemistry – Reviewers	37	30	3.7	3.0	11	6	59	47
Inorganic Chemistry – Committee members	43	23	4.3	2.3	15	13	70	33
Inorganic Chemistry – AvH staff	36	37	3.6	3.7	2	5	70	68
Materials Science – Reviewers	31	20	3.1	2.2	15	4	51	37
Materials Science – Committee members	16	5	1.6	0.6	16	3	16	6
Materials Science – AvH staff	41	40	4.1	4.4	3	7	79	72
Zoology – Reviewers	36	31	2.3	3.4	15	12	65	52
Zoology – Committee members	37	79	2.3	8.8	37	79	37	79
Zoology – AvH staff	59	49	3.7	5.4	29	10	88	88
Modern History – Reviewers	38	22	2.7	1.6	8	11	71	36
Modern History – Committee members	11	5	0.8	0.4	7	5	14	5
Modern History – AvH staff	44	31	3.1	2.2	26	1	61	60

Source: Compiled from the platform logfiles

Another difference is the involvement of the committee members. In the conventional procedure, committee members can also act as reviewer when their specialization is closely related to the application. In the PC, this is also the case, and some but not all committee members are active in the review process. In Zoology, the committee members were very active (together 6.2 hours in round 1 and 7.8 hours in round 2). Next is Inorganic Chemistry

[13] This includes coordinating (not substantive) comments, such as asking reviewers to address a certain issue.

(5.9 hours; 4.3 hours), followed by Materials Science (2.5 hours; 4.5 hours) and History (0.9 hours; 1.2 hours).

In terms of change, in all PCs the number of comments declined for reviewers, and for committee members, apart from Zoology where they increased. For AvH staff the picture is mixed. This overall decline is not due to a decline in the number of applications, as the "per application" column (Table 10) shows.

The decrease in the number of comments per application was not offset by longer comments: the average length of the comments decreased slightly between round 1 and round 2, from 783 to 712 words. As we will see below, the time spent reviewing the applications also decreased in round 2 compared to round 1. The reviewers attributed this at least in part to the planning of the second round: Starting mid-August, the first half of the second round coincided for many PC members with the summer holiday (August), with the period where many scholarly conferences are organized (September), and the start of teaching.

The distribution of PC activities over time in the first round differed somewhat between the Peer Circles (Map 1). In Inorganic Chemistry, some reviewers started early, but then there was a long period with hardly any activities, until the PC became active at the end of the period. The same is true for Materials Science. In Modern History, activities were concentrated at the beginning and again at the end, and in the Zoology PC, there was also a concentration of activities at the beginning, but then the activities were distributed over the second half of the available period. We did the same analysis for the fall round. The pattern is somewhat different: In Zoology, activities were distributed over a long period, with some concentration in the middle, and in Materials Science, there was no activity at all at the beginning, and all activities took place in the second half of the review period. To summarize, in some but not all panels, the activities were concentrated in or around the interactive period. In the second period this was less the case than in the first.

Legend: Maps 1 and 2 – next two pages

The tables on the next two pages (based on the platform logfiles):

- The top line indicates the weeks: week 1, week 2 etc.
- Then the three groups by Peer Circle:
    - the white lines are the reviewers,
    - the green lines the committee members,
    - the yellow lines are the AvH staff members
- This repeats for each of the four fields
- The number in a cell shows how many review statements or comments were given in that week by the PC member
- The red ovals indicate the periods with dense activity
- The grey strips indicate the 'interactive periods'

Map 1. Activity over time – summer round

		04/	05/	08/	09/	10/	15/	16/	17/	18/	19/	22/	23/	24/	25/	26/	27/	28/	29/	30/	31/	01/	02/	05/	06/	07/	08/	09/	11/	12/	13/	14/	15/	16/	17/	18/	19/	20/	21/	22/	23/	25/	26/	27/	28/	29/	30/	01/	02/	03/	04/	05/	06/	07/	10/	11/	12/	13/	14/	17/	23/	24/	25/	26/	27/Oct	
PC	role	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60	61	62	63	64	tota
chem	FG																																																			38	6			3										47
chem	FG												10			2	5	4	3									6			1																							10												41
chem	FG																			1							12	5			3									2												8		1												32
chem	FG																6			4																										6								8												24
chem	FG														5						1																																													6
chem	2 FV																																																									6	7							13
chem	2 AvH						1	1	2	1		7	1	1					3	3	1				1						3			2			3			3	2						1				1	1	10	4	8	10	2	1								73
mater	FG																																											14	22									1												37
mater	FG																																		10							10								8		4	1													33
mater	FG																																			2	9		2		4		3			1				3																24
mater	FG																																																		6		8	4												18
mater	FG																																																				16													16
mater	FG																																										4	5																						9
mater	FG																																																		4															4
mater	2 FV																																																													3		6		9
mater	2 AvH									1		10														1				1													1		1	5	1			2				12		28	6			3						72
zool	FG																								11		11											4													26															52
zool	FG																		2								4		6									11	5																						8	1	2			39
zool	FG																6	2		6							4				5						3						2			3	3																			34
zool	FG																									15							9	5				4																												33
zool	FG											2								10				1							5	10																																		28
zool	FG																												9											7			5																							21
zool	FV																					7													47			2		1	1			1		11		1										8								79
zool	2 AvH			1		3				3		11		1						3				8			6	1				1		1			2	3		5	4			5	3	1	5					1	1		14		15									98
hist	FG											11								16	9																																													36
hist	FG												1	5			7		2				6																				1			4	4	3																		33
hist	FG																			12	8																															5	7													32
hist	FG											1	8				10		5																																	4														28
hist	FG											6				10																																			9															25
hist	FG											1	4	1		5																																				3	1													15
hist	FG																													6								6																												12
hist	FG												1					5		1								4																																						11
hist	FG																																													4		3	4																	11
hist	FG																	1	3																								1			5						1														11
hist	2 FV																																																														2	8		10
hist	2 AvH								10		1	6			6					5				2																7	1		3								2				6	3	8							1		61
total		0	0	1	0	3	1	1	12	5	1	55	25	8	11	17	34	12	18	61	19	7	6	11	12	16	37	16	15	7	17	11	9	8	57	2	17	30	7	25	12	10	20	25	26	40	14	7	4	13	48	65	50	40	28	44	31	15	7	3	8	4	4	15	0	##

Map 2. Activity over time – fall round

7.3 Mode of operation and integration of Peer Circle activities

One important concept underlying the PC approach is that it is a joint review process, where not all PC members need to review all parts of an application. The first reviewer (usually the one whose specialization is closest to the application) is expected to start and answer all the review questions, and the others are expected to answer review questions as far as they feel qualified to do so and comment on the reviews by others. As already described, the first reviewer did not always start early in the process, and that influenced the work of the other PC members. More than half of the reviewers started or tended to start after the assigned first reviewer had begun, while less than half of them tended not to wait for the first reviewer (Fig. 6).

Another concept underpinning the PC is that people whose specialization does not cover the project proposal may still also be able to meaningfully contribute, for example by assessing the CV, but also parts of the technical content such as the methodology, or the (societal) relevance. If a reviewer is not an expert in the topic of the research project, they may have been in the home country of the applicant, and may therefore be able to evaluate the career. PC members did indeed participate in evaluating those applications, by reading parts or the whole application, and adding comments to some of the review questions (Fig. 7).

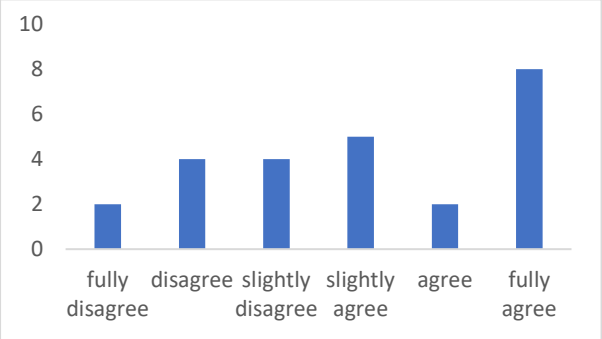


Figure 6. I started my review only after the assigned first reviewer had begun.
N = 25

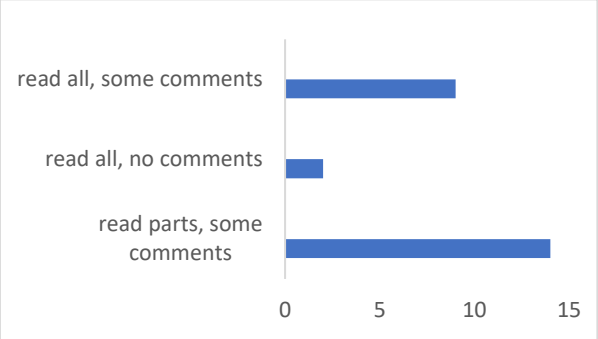


Figure 7. How did you handle applications far from your expertise?
N = 25

The interviewees stated that they read the full application when they were assigned the first reviewer role. However, the PC approach did make it possible to participate in the review process without reading all applications in full. About a quarter of the PC members involved read all applications (Fig. 8). The interviews also indicated that it was easier to evaluate the CV and the future potential than the project proposal – as that required the PC member to be a specialist in the specific specialty.

Finally, the reviewers felt confident to contribute, even without being a real peer. Figure 9 shows that three-quarters of the PC members agreed slightly to fully that it was easy to contribute to the evaluation in such cases. The interviews suggest that some of the more junior PC members felt more uncertain about their role than the experienced members.

In general, one can conclude that the work was distributed among different PC members, related to the differences in expertise – and that is one of the goals of the Peer Circle.

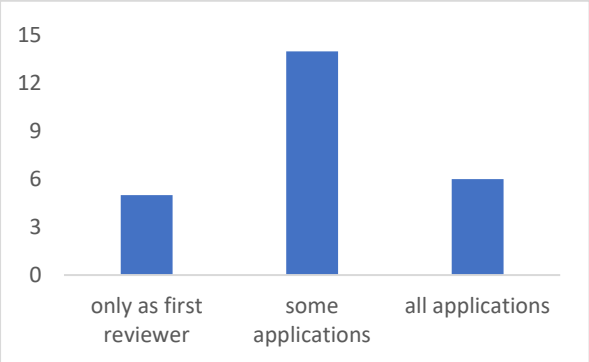


Figure 8. I read the full applications
N = 25

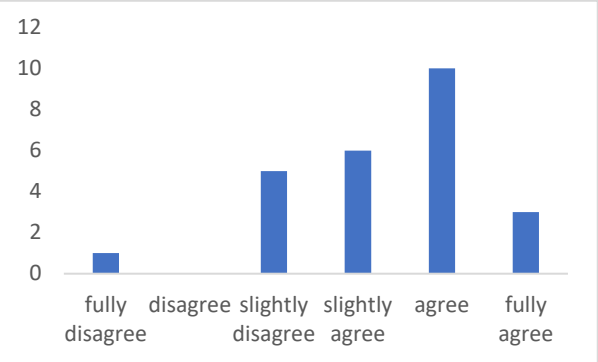


Figure 9. Easy to contribute when I was no real peer, as others could add their view
N = 25

In the conventional review process, a reviewer is fully responsible for the quality of the review, but in the PC there is collective responsibility. Individual PC members may therefore feel less responsible for the result. This is true for slightly more than half of the reviewers, and not for the others (Fig. 10).

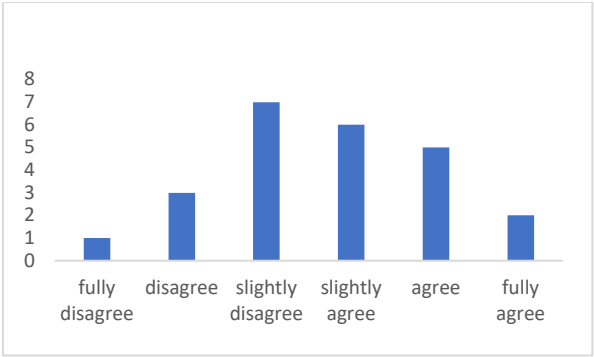


Figure 10. As reviewing is a group activity, I felt less responsible for the results
N = 24

The interviews reveal wide variety when it comes to how PC members plan their activities. Some start immediately, some wait for the interactive periods, and some are deadline workers who do not start until the end of the review period, complicating the ability to comment on one another's contributions. With regard to the latter, integrating the PC activities was especially problematic when one was the assigned first reviewer. Others want to do the work concentrated in several "blocks" and have no time to look in between at the reviews that others have added. Others do the work in between their other activities, making it easy to integrate the PC work.

The interviews in the first round made it clear that starting the review process is not a problem. Commenting on the reviews of others is more difficult because PC members' schedules differ, due to differences in workstyles as described above. Commenting on others requires accessing the online platform regularly, which seems difficult to integrate into everyday activities. For example, to respond to reviews relating to or comments on an application added weeks after a PC member has prepared their own review of that application may require re-reading the application, which is often too time-consuming.

However, the survey after the second round shows that the vast majority find it easier to integrate the PC work into existing schedules, albeit in most cases only modestly (Fig. 11). There seems to be a learning effect: It becomes somewhat easier in the second round. The survey also shows that more than half of the members agree or strongly agree that the PC allows more flexibility than the conventional approach (Fig. 12).

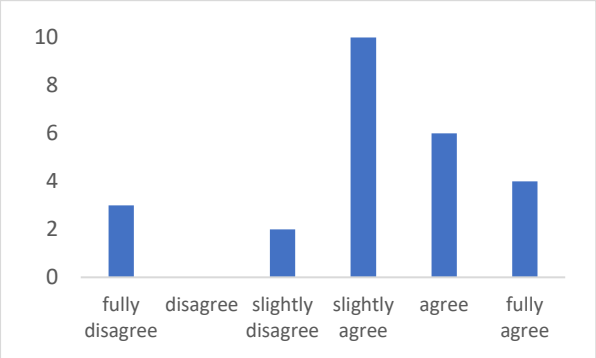


Figure 11. The PC work can easily be integrated in my normal schedule
N = 24

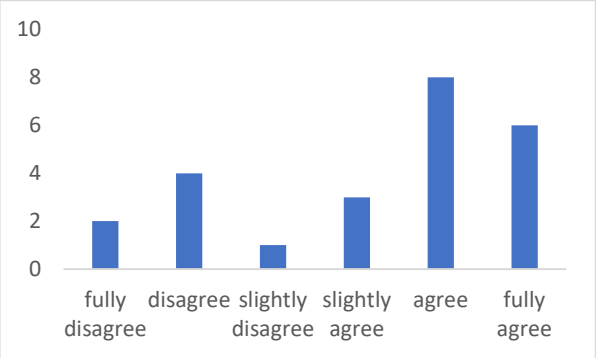


Figure 12. The PC allows more flexibility than than the conventional review
N = 24

One important point that emerged in the interviews is that the period for conducting the reviews is probably too long, and not yet optimally structured. Several interviewees (reviewers, committee members, and AvH staff) suggested shortening it and/or structuring it more. For example, allowing two weeks for the initial reviews and comments, and then two weeks for commenting on one another's initial contributions. Knowing such a structured schedule in advance would allow the PC members to integrate it in their timetables, and it would also make interaction easier. And if there are many applications, such a schedule could be repeated a second time.

The study covers two rounds of the PC, as participants had to get used to the PC, and to develop routines. For example, the role of the first reviewer was not well understood by everybody in the first round, but understood much better in the second. In the interviews with the reviewers and the committee members, we asked how they proceeded in the Peer Circle, and a kind of "average procedure" seemed to emerge from this (Text block 1).

Step 1: Generally, one begins with the applications for which one is assigned as first reviewer. Those applications were reviewed as would be done in the conventional way, but because of the format of the PC (answering questions) there is no need to write a "story".

- Reading the full proposal, and sometimes also (parts of) the papers. The latter are at least skimmed.
- Reading the comments of the other PC members on that application and responding to them.

Step 2: The other applications that are close to the expertise of the reviewer.

- Sometimes a full review as for the assigned application, sometimes less detailed reading, but focusing on crucial issues.
- Only a few Peer Circle members (and some committee members) reviewed/commented on all applications.

Step 3: Finally, the other applications, further away from the reviewer's expertise. This is where the intended division of work occurs.

- Some reviewers go through all applications, others select a few where additional comments seem useful, often using the other PC member's comments as point of reference. For example:
  - neglecting those that has already been reviewed negatively or very positively and focusing on the in-between group; or
  - reading the comments of others and reacting where appropriate.
- These applications are briefly scanned, with a focus on the summary, and on the non-technical parts important for the AvH selection process such as mobility, CV, and productivity.
- Finally, (dis)agreeing with a review/comment. When disagreeing, writing why.

Text block 1. An emerging procedure

7.4 Time used, effort, motivation

Some of the younger PC members reported in the interviews that they had no previous experience with grant reviews and therefore could not assess whether the PC takes more or less time than the conventional approach. Furthermore, time use studies based on self-reporting are notably unreliable. As most PC members did the work online (Fig. 9), an analysis of the logfiles may provide more additional insights.

The interviews show that some reviewers feel that the PC takes more time, mainly because more reading is needed due to the number of applications. Others think it takes less time because of the considerably reduced writing time, which is due to the presentation of the "question structure",[14] the possibility to comment on the contributions of others, the more informal writing style, and because one can (dis)agree with the review and comments of others without repeating the arguments.

The survey gives a differentiated picture. Whether reading other members' comments saves time is not univocal; a small majority agrees, but almost half of the PC members disagree (Fig. 13). However, many of the PC members feel that per application the PC procedure is somewhat more time efficient (Fig. 14) and the same holds for the PC as a whole (Fig. 15). Almost all find the time the PC takes to be within reasonable limits (Fig. 16).

[14] The classical HFST peer review also uses the "question structure", but it presented differently.

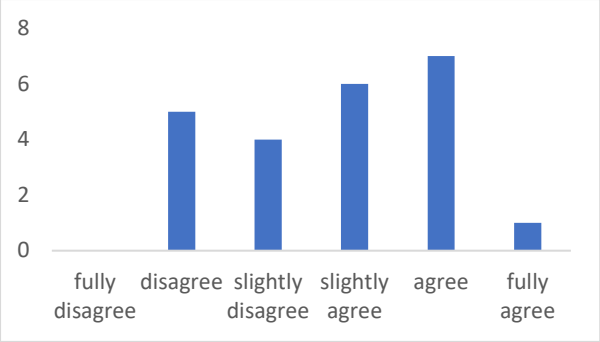


Figure 13. Reading others' comments saves time
N = 23

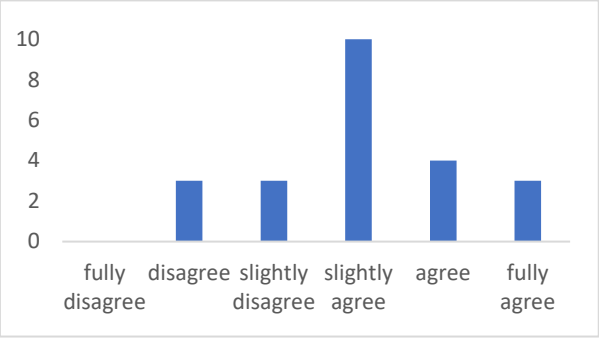


Figure 14. Per application, the PC takes less time
N = 23

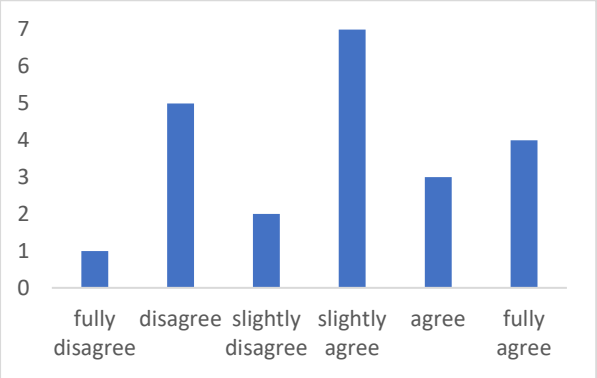


Figure 15. The Peer Circle saves time
N= 22

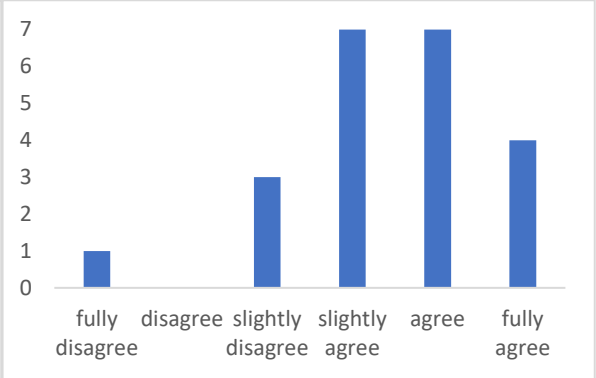


Figure 16. The time the PC takes is reasonable
N = 22

Several of the interviewees who agree that the PC takes less time per application do feel that participating in the PC takes more time overall. In contrast to the conventional review approach where a reviewer assesses only one application, PC members contribute to the review of more applications which takes more time. However, interviewees also argued that it is more efficient for the *community* as a whole: Fewer reviewers are involved than in the conventional approach, and even if PC members were to spend more time individually, for the community as a whole the total time for reviewing becomes lower. In the four experimental fields, 28 PC members reviewed 89 applications, which would have required 178 reviewers in the conventional review procedure.[15]

The logfiles also provide information on the time spent for the PC. As most members did the PC work mainly or fully online, this is a reasonable source, even if activities of some users were not (completely) logged by the system. The *hours logged* in is less reliable, as it also counts the time when one is logged in but is doing something different. The *hours active* is a much better indicator as it counts the time PC members were working on the reviews. Table 11 provides the relevant information at the aggregate level. The average amount of time active online is about 4 hours in the first (summer) round, and about half as much time during the second (fall) round. This may not (always) include the reading time.

[15] This is based on the rule of two independent reviewers per application. This is not met in one third to half of the cases, illustrating the problems with the conventional peer review process. However, in cases where the committee member is close to an application, they may be appointed as second reviewer which reduces the need for reviewers somewhat. When there is only one reviewer, the committee determines whether a decision can be based on that review.

Table 11. Login time in the online platform

All reviewers and committee members	Hours logged in		Hours active	
	Round 1	Round 2	Round 1	Round 2
Average	25.9	14.7	4.2	2.1
Minimum	0.6	0.2	0.9	0.1
Maximum	77.1	49.7	9.4	7.9

Source: Logfiles

In the interviews, a few reviewers tried to estimate the number of days used on the PC. Eight reviewers did so for the first round, and the number of days used varied between 1 and 8 days, with an average of 4 days. The same holds for some of the committee members, who also put this at a few days. These interviewees have a low "use level" in the logfiles, suggesting that the others used more time for the Peer Circle.

Ease of use: The survey suggests that the online platform made it relatively easy to comment on the applications (Fig. 17). The average score is 4.5, which means between slightly agree and agree, a little bit higher than after the first round.

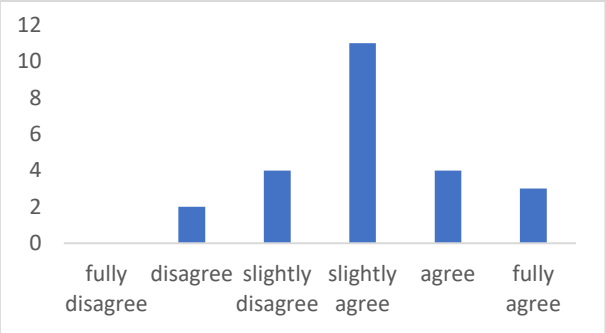


Figure 17. It was easy to comment on the proposals
N = 24

Motivation: Many Peer Circle members were motivated to participate because they had been funded by the AvH in the past, so their participation can be seen as a service to the community and more specifically to the AvH. This became clear in the interviews as well as in the survey. But there were also other motivations:

- Several interviewees wanted to participate in the experiment in order to help improve the peer review system.
- Others, especially the more junior members of the PC, expected to learn about how reviewing works, and what counts as a good application in the eyes of reviewers. Getting exposed to many applications was very useful for them.
- Furthermore, several PC members mentioned benefits like getting new ideas, staying up to date about changes in the field, and improving application skills.

7.5 Interaction, conversation, consensus

Interaction & conversation: The interviews suggest that there was not much interaction, although slightly more in the fall round compared to the summer round. Interviewees felt that

there were some reactions to first review contributions on the platform, but seldom a more continued conversation. The survey shows that almost all PC members observed some or regular replies (Fig. 18).

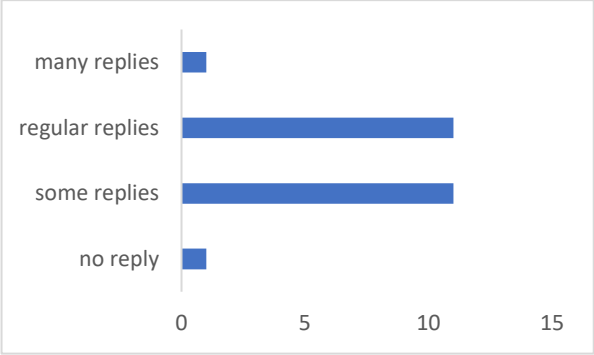


Figure 18. PC members replied on the comments of others
N = 24

Analysis of the textual contributions on the online platform can add to this. The answers to the review questions and the (subsequent) comments were coded with a sequence number indicating the place in the conversation, which is used for counting the number of conversations by length. Content analysis of the review contributions would be useful to clean the data and take out non-substantial contributions, but that is outside the scope of the current report. Table 12 shows the results, and we are including here all contributions to the exchanges, including administrative messages from AvH staff, which is about one third. Taking that into account, the number of long conversations is indeed modest.

Table 12. Length of the conversations

Conversation length*	Number of conversations	Total contributions
1	119	119
2	156	312
3	136	408
4	133	532
5	80	400
6	60	360
7	40	280
8	22	176
9	6	54
10	8	80
11	1	11
12	3	36

* In # contributions
Source: Logfiles of the online platform

Consensus: PC members may be influenced by reading other members' reviews and comments before finishing their own review. In the first round, most PC members tended to read the other members' comments first. This changed in the second round, when most of the reviewers reported not beginning to systematically read the others' reviews and comments before beginning their own (Fig. 19). This suggests that most PC members have their own

assessment that they share through commenting on the reviews of others (Fig. 20) – which was one of the ideas behind the PC concept.

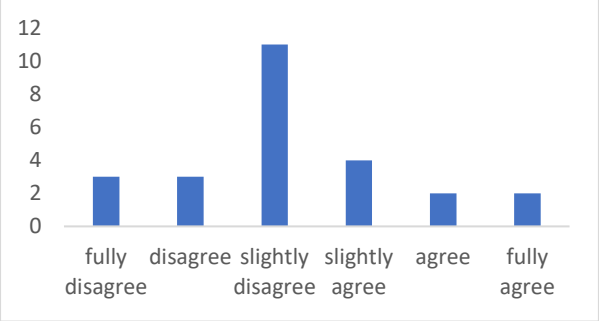


Figure 19. I read the other comments before starting my own
N = 25

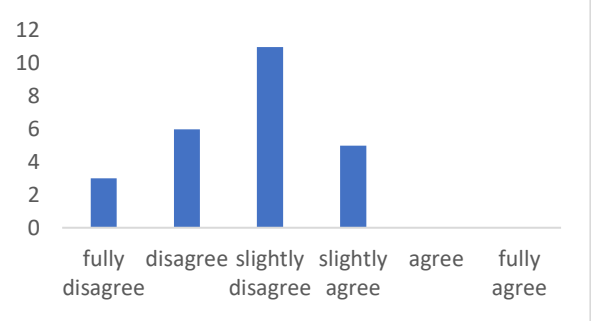


Figure 20. I commented on the assessments of others, instead of presenting my own
N = 25

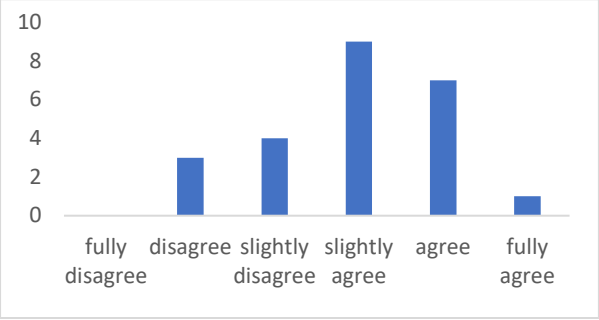


Figure 21. I was influenced by reading the comments of others
N = 24

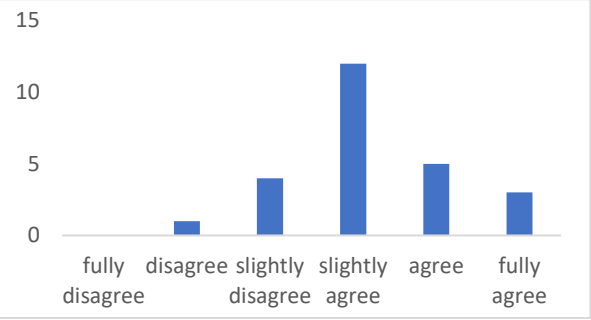


Figure 22. Reading comments of other reviewers helped develop my own opinion
N = 25

Most respondents indicated that they were influenced by reading the comments of other reviewers to some extent (Fig. 21). The interviews indicated the same: Almost all reviewers interviewed were aware of the possibility of being influenced in a group process like the PC, and several mentioned that they were inclined to follow the comments of others, especially those of the assigned first reviewer. However, those interviewees also state they remain aware of this and reflect on whether this occurs.

They also feel that being influenced by others has advantages, because it helps to develop one's own view (Fig. 22), and to reflect, which may ultimately lead to a different assessment. Another advantage of mutual influence is that it helps to achieve consensus. This is a positive as an aim of the review process is to arrive at some consensus about who to fund.

Being influenced is not the same as *premature consensus,* where there is the risk of suppressing possible contradicting views. The interviewees (reviewers, committee members, subject group managers) were all aware of this risk, but almost none of them felt it had occurred. One or two instances were mentioned where premature consensus seemed to emerge, but these were explicitly addressed in the PC.

7.6 Interaction between committee members and reviewers

There was no interaction between reviewers and committee members in the Modern History Peer Circle, as there the committee members only entered the process at the end of the Peer Circle activities. Also, one of the Materials Science committee members entered only after the PC activities were finished. From the perspective of the reviewers, the presence of the committee members in the Peer Circle was not an issue. Even though some of the committee members were very active as reviewers, the other reviewers may not even have been aware of this, as everyone in the PC was anonymous.

Committee members utilize the results of the reviewers and integrate these into a proposal for the committee. Possible tension may emerge here, as it is not only the committee member in the PC who may have developed a good overview of the set of applications, but also some or all of the PC members. To date, it has been the committee member's task to draw up a proposal for the committee, but increasingly this may also be done by the Peer Circle, as the information advantage of the committee member diminishes. Several PC members pointed in that direction by suggesting that the PC may develop its own ranking. One committee member suggested sending the proposed ranking to the PC first to get feedback before sending it to the committee.

7.7 Comparing several applications

One disadvantage of the conventional review procedure is that a reviewer produces only a single review, and they are therefore unable to assess how good that application is compared to other applications. This often leads to relatively high scores for all applications, independently of the amount of praise and criticism expressed in the review. In the PC, the members see all of the applications and therefore they can grade the different applications in relation to one another. Almost all PC members see this as an advantage. The interviews point in the same direction, and many interviewees feel that comparing the applications makes reviewing easier (Fig. 23) and also better. The result is improved input for committee members, who can now rely on a much broader comparative grading of the applications.

Would it be useful if a PC came up with a shared ranking of the applications? Opinions diverge, and slightly more than half of the respondents tend to agree, whereas the smaller other half tend to disagree (Fig. 24). Several interviewees addressed the issue of comparing and ranking the set of applications within the PC and discussed the practical difficulties in organizing the preparation of a ranking. In this respect it was also mentioned that the reviewers should use the whole four-point scale, and not only 2, 3 and 4, because that would help to better differentiate between the applications.

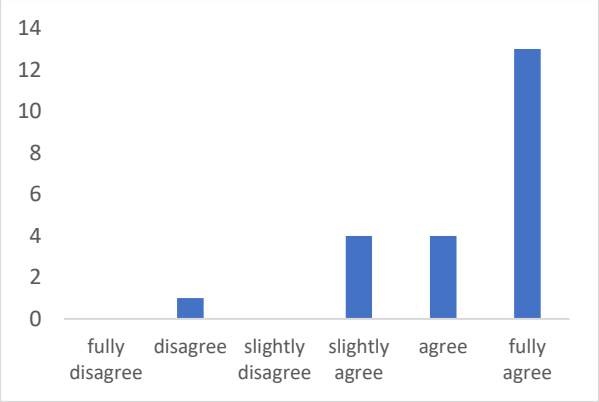


Figure 23. Comparing applications makes reviewing easier
N = 22

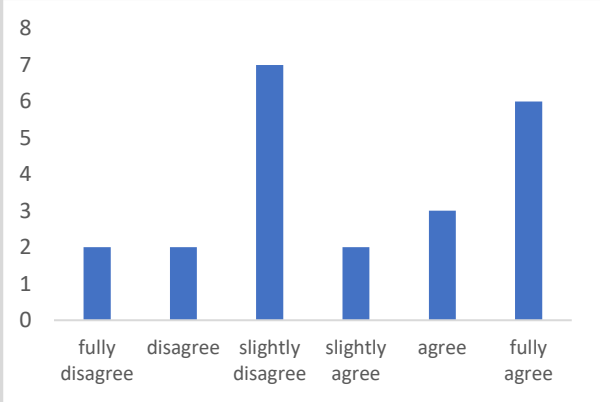


Figure 24. The Peer Circle should come with its own ranking
N = 22

7.8 Acceptance

The experiment shows that the PC could be well received by the scientific community. In the interviews, the majority of PC members shared a positive opinion about the Peer Circle as an alternative to the traditional peer review model, stating that it is the way forward. The survey (Fig. 25) shows that 18 of the 24 respondents were positive or very positive, and another two slightly positive. Only three tended slightly to the negative side. If we differentiate between the four fields, there are some but not large differences. Three PCs scored rather positively: Zoology, Modern History and Inorganic Chemistry are close to one another with average scores of 5.2, 5.1 and 5.0 respectively. Materials Science scores somewhat lower with 4.4 on a six-point scale but is still positive.

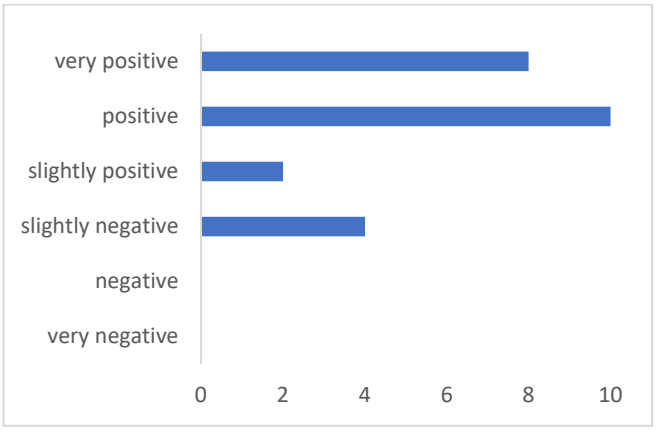


Figure 25. General impression about the PC procedure
N = 24

The interviews revealed a need for new ways to resolve the problems of the peer review system, and most committee members favored the PC concept for the HFST program. A few also mentioned some disadvantages, particularly that it makes it more difficult for a committee member to deviate from consensus in the PC. The only negative assessment came from one Modern History committee member, who believes that reviews by non-specialists are not helpful in the selection process.

8. Context and implementation of the procedure

8.1 Peer Circle membership

One of the aims of the PC is to have a more diverse set of reviewers, including researchers in earlier career stages than professor. Table 13 show the composition of the peer circles, and there are some differences between the four fields. The goal of including more women and more young researchers in the PC has, however, been realized. Almost half of the PC members are women, and there is an even age distribution, with more than a third of the PC members being under 40 years old.

Table 13. Demography of the Peer Circle

	Female	Age 30-39	Age 40-49	Age 50-59	Over 60	Prof	Total
Peer Circle members	43%	36%	32%	31%		46%	28
Committee members	14%			71%	29%	100%	7

How many applications did the 28 PC members review? In the first round it was 50, and in the second round 42. Zoology was the largest subset with 30 applications, followed by Modern History, Inorganic Chemistry and Materials Science with 20, 20 and 19 applications respectively (Table 14).

Table 14. Number of applicants per round

Peer circle	Round 1	Round 2
Inorganic Chemistry	10	10
Materials Science	10	9
Zoology	16	14
Modern History	12	8
Total	48	41

8.2 Anonymity

In the preparation phase, it was discussed whether Peer Circle members should be anonymous. The idea behind this question was that especially the less experienced members without a tenured position may feel less free to give their opinion if that would contradict the evaluation of the senior members of the PC. Before the start of the first round, the PC members and committee members voted by a majority to keep the process anonymous. The anonymity was not complete, as some of the participants met each other (online) during one of the two introductory meetings where the procedure and the online platform were introduced. Once the PC was finished, the committee members would know the identity of the reviewers. Although this was mentioned in the introductory meeting and in information material provided by the AvH, not all reviewers seemed aware of it.

The question of anonymity was addressed in the interviews, and the main arguments in favor of anonymity are (i) that it protects younger reviewers, so that they can speak freely, especially those with temporary positions, (ii) that reviews should not be hindered by status differences, (iii) that it protects confidentiality, and (iv) that it makes it less likely for reviewers to be influenced by others. Some proponents of anonymity argued that during the

PC process everything should remain anonymous, but afterwards the names of the reviewers should be disclosed (which is already the case for the committee members). This would make it easier to weight the comments, and to prevent situations where anonymity causes *conflicts of interest* to go unnoticed. However, according to one AvH staff member, there are checks for CoIs at several moments in the procedure – of which not all PC members may be aware.

The opponents of anonymity use the arguments that (i) also young researchers should be able to defend their views in public (like at conferences), (ii) anonymity may lead to c*onflict of interests going unnoticed*, (iii) anonymity makes interaction more difficult, (iv) anonymity makes it impossible to weight comments and (v) anonymity reduces transparency.

The interviews indicate the majority of the PC members to be in favor of anonymity, and the surveys show that, between the two rounds, the majority became somewhat larger: from two thirds in the first survey to three quarters in the second survey. Those that were against anonymity felt so to a modest degree, i.e., were "somewhat against anonymity" (Fig. 26). Within the group of committee members, a stronger change could be observed. After the first round, three out of seven committee members were against anonymity, after the second interview only one out of eight.

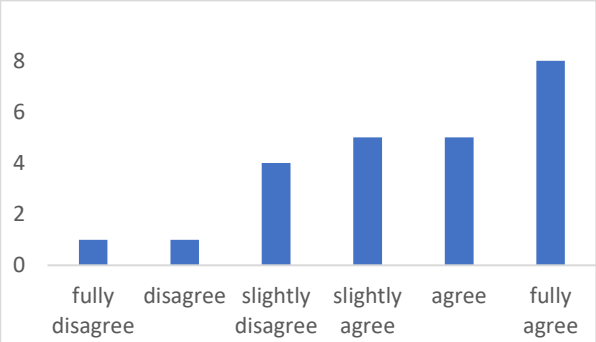


Figure 26. Anonymity is beneficial for the PC
N = 24

8.3 General aspects regarding the procedure

First reviewer: To avoid a situation where everyone waits until someone else has begun reviewing, for each proposal the coordinators asked one or two PC members with expertise close to that proposal to start with the review process. The interviews in the first round indicate that not all PC members had clearly understood the role of the first reviewer, particularly that one was expected to start the review process early, which led to a late start in some PCs. In the second round, the roles were better understood (Fig. 27), and starting early was easy (Fig. 28). Nevertheless, the logfile data show that starting late remained a problem for two of the PCs.

Guest reviewers: Occasionally a project proposal fell outside the expertise of the PC members and a guest reviewer was invited. The subject group managers noted that these were not difficult to find, and the reviewers found this a workable solution. Committee members also pointed out the need to invite an external reviewer in such cases. However, expertise mismatch is not a specific PC problem, and several interviewees noted that it also occurs in the conventional peer review process. In the Peer Circle it is more easily identified than in conventional reviews.

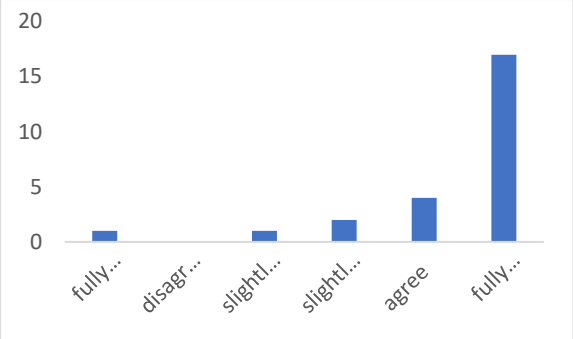


Figure 27. When assigned as first reviewer, I felt stimulated to start the review
N = 25

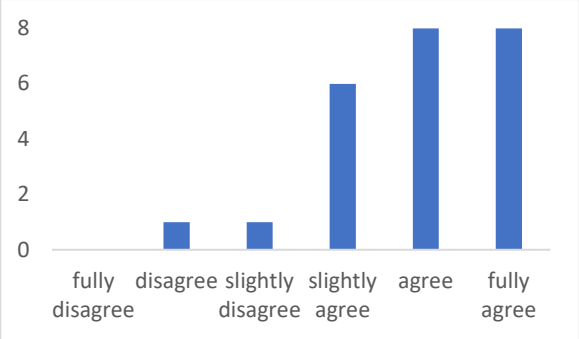


Figure 28. I found it easy to be the first to start a review
N = 25

Phases in the PC process: The PC lasted for quite a long time, the first round for eleven weeks, and the second round for eight weeks. To stimulate interaction between the reviewers, two "interactive periods" of two weeks were defined in each of the rounds. Most PC members found the interactive periods modestly useful (Fig. 29), and the role of the interactive periods was not clear to all reviewers, given the scores on the question "I felt I had to review during the interactive periods": Only half of the PC members agreed with this (Fig. 30). The analysis of the logfiles (Maps 1 and 2, 7.1) suggests that the review activities were generally concentrated at the beginning and the end of the Peer Circle periods, not always within the interaction periods. Examples are Zoology with a different pattern in both rounds, and Inorganic Chemistry and Materials Science in the second round.

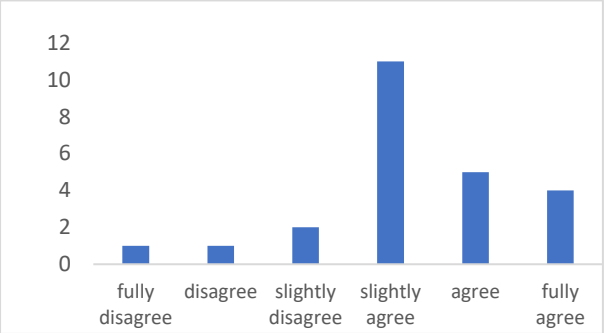


Figure 29. The interactive periods were useful
N = 24

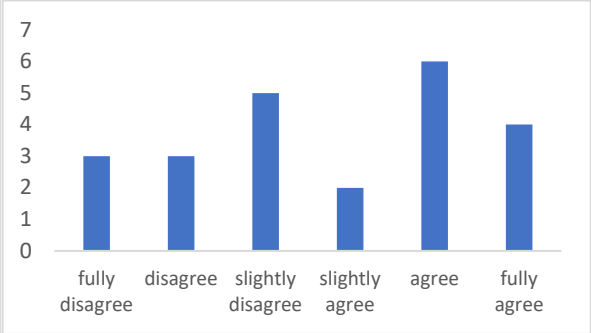


Figure 30. I had the impression that I should review only during the interactive period
N = 23

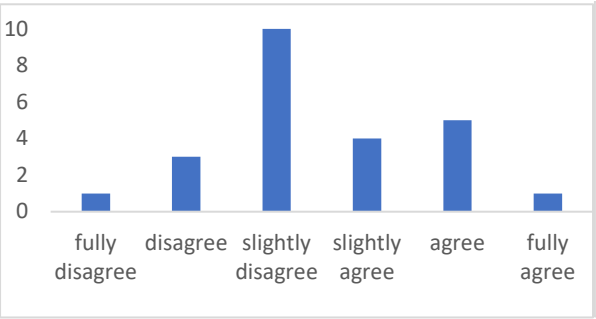


Figure 31. I plan to distribute my activities over a longer period of time
N = 24

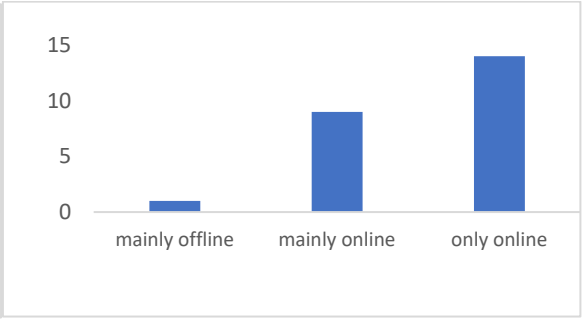


Figure 32. How did you prepare the review/comments
N = 24

The interviews linked these findings to the differences in work styles of the PC members. Some interviewees mentioned that the duration of the PC should be shorter and more structured, with several deadlines for the first reviews and for the commenting.

Grading: The PC members were only asked to give one overall grade for the applications, and not one for every evaluation criterion. The interviewees addressed the issue that is felt important but sometimes also difficult: How to grade the total application in cases where the reviewer did not go into all aspects of the application? Grading of the individual evaluation criteria (like in the conventional procedure) would be more appropriate precisely for the Peer Circle.

The four-point scale was also addressed. Since, in practice, only a part of the scale is used, it was felt by several interviewees to be too narrow for a proper comparison of the applications. This is an important point specifically because comparative assessment is one of the main advantages of the PC.

Online versus offline working: Almost all PC members worked fully or mainly online (Fig. 32), which makes the logfiles of the online platform a useful dataset for analyzing PC members' activities (see 7.2). Some reviewers suggested that working offline should be possible, with automatic uploading to the platform. The background to this is that one does not always have internet access when one has time for reviewing (e.g., while traveling).

Inviting the PC members: Finding members for the PC seemed to be easier than finding conventional reviewers according to the interviews with AvH staff and the committee members. This was partly because the invited PC members were all from the AvH network, but more importantly, the number of PC members is much lower than the number of conventional reviewers that would have been needed for the same number of applicants.

In this regard, several interviewees noted that reviewers should remain a PC member for a period of two or three years so that they gain sufficient experience. This would also save time in finding reviewers. And to avoid large fluctuations in experience in the PC, all PC members should not be replaced at the same moment.

Finally, the platform is in German, as are most of the reviews and the comments. This hampers the participation of those who do not speak German.

Preparation of the material: The PC is an experiment, and therefore it is not (yet) embedded in the standard digital procedures of the AvH. This means that considerable work is involved in preparing the documents for the PC and the online platform, in another format than normally provided to the reviewers. The interviews with the case handlers made clear that in the current situation the administrative workload of the PC is considerable. However, they expected that this problem would resolve itself once a dedicated PC platform was integrated into the normal workflow. Until then, the PC needs additional effort at the level of administrative handling.

Coordinating the Peer Circle as a new task: The conventional way of reviewing does not require intervention and coordination, apart from stimulating the reviewers to submit the review in time and checking the quality after the review has been received. In the conventional review process there is no interaction with the reviewers, but the PC enables the AvH staff and the committee members to ask for clarification about things that remain unclear and to stimulate the PC members to have a look at something again. More generally,

the PC is a group activity with more interdependence and coordination needs. That coordination task of the AvH staff required logging into the platform quite often, monitoring activity or the lack thereof, and asking PC members to begin reviewing, or responding to specific issues that were yet covered. Answering questions asked by the PC members was also part of that coordination task. This leads to an additional workload which may decrease as reviewers get used to the new way of working, but it will remain an extra task for AvH staff. In other words, the PC implies some role and task changes for the involved AvH staff, and a modified workflow.

The interviews (with reviewers and with committee members) show that the coordinating tasks of AvH staff were highly valued and were also perceived as crucial for the functioning of the Peer Circle. This is also reflected in the survey: 21 out of 24 PC members agreed (strongly) that moderation by the AvH office was useful (Fig. 33), and for communication with the AvH office, 20 out of 25 PC members held that view (Fig. 34).

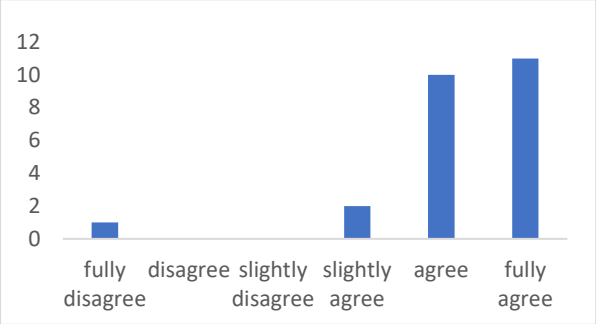


Figure 33. Moderation by the AvH was useful
N = 24

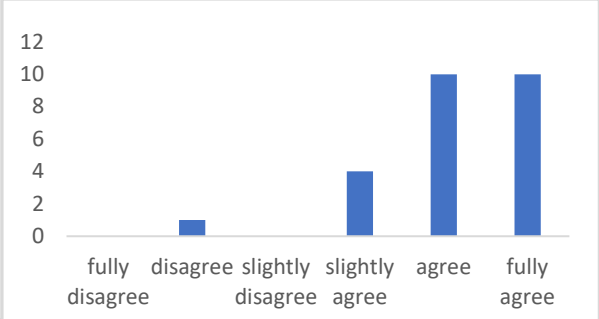


Figure 34. Communication with AvH was useful
N = 25

Summary report: The interviews with the subject group managers suggest that it became easier to write the summary report, as (i) the subject group manager is more actively involved in the review process, and (ii) when things were unclear it was possible to ask the reviewer(s) for further clarification.

8.4 Functionality and user friendliness

The Peer Circle was implemented in an existing online platform that is not integrated in the AvH back-office procedures, and a dedicated platform is needed when the PC continues. The current evaluation is therefore not targeted at the usability of the online platform. However, we did ask about it in the first survey and in the interviews, to see if usability issues affected the experiment. This did not appear to be the case, even though some users sometimes did have problems. Most PC members[16] found the platform fine and user-friendly: intuitive, easy to use, self-explanatory. The interviews suggest the same, and of the twelve interviewees in the first round, nine were positive about the online platform, and three had a more mixed view. This did not change after the second round. The committee members were more critical than the reviewers, more in the first than in the second interview. Many interviewees came up

[16] Eleven out of thirteen in the first survey.

with suggestions for improvement when designing a dedicated Peer Circle platform. Two of these issues are worth mentioning here:

Firstly, only about half to two thirds of the reviewers (in the first round) found it relatively easy to get an overview of the status of individual reviews and of when a proposal had been sufficiently reviewed. These opinions may change as reviewers become more experienced, but the fall survey only found a difference for "ease of keeping track". Many more respondents agree to some extent that this is easy (Fig. 35). But it remains unclear to quite a few PC members when an application had been adequately reviewed (Fig. 36).

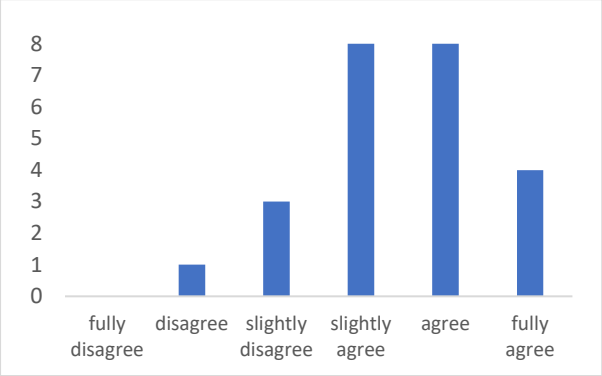


Figure 35. It was easy to keep track of the review process
N = 24

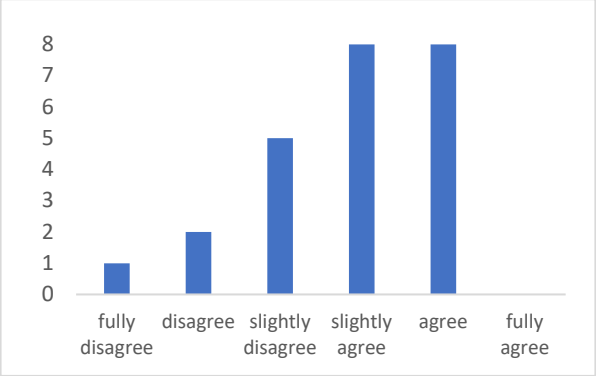


Figure 36. It was always clear when an application was sufficiently reviewed
N = 24

Another issue is the way the documents are organized in the online platform. In the first round, one third slightly disagreed that the documentation was better organized in the PC than in conventional forms of reviewing. After the second round this changed somewhat, and about 80% gave a positive score (Fig. 37). However, the average score remained the same (3.9 in both surveys) indicating that the critical PC members gave even lower scores in the second survey. On this item there were quite a few missing values, but the non-respondents and respondents did not differ in age or research field.

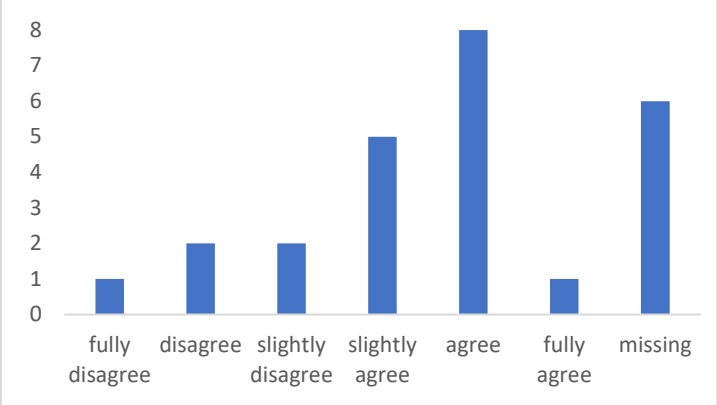


Figure 37. The documentation was better organized
N = 19

Part III – Conclusion and recommendations

9. Conclusion

The evaluation aims to answer the following questions about the Peer Circle:

1. Are the resulting PC reviews of at least the same quality as conventional reviews?
2. Does the interaction within the PC result in premature consensus, hindering critical consideration of the applications?
3. Do the reviewers and committee members perceive the PC as a better alternative?
4. Is the PC more efficient than the conventional approach, such as the amount of time it takes?
5. Does the PC committee select the best applicants?
6. Do the PC reviews influence the gender balance in the selection outcome?

1. Quality of the reviews

Compared to conventional reviews, the PC review has a different format, which was intended: it is shorter, more informal and less a constructed story leading to a verdict. This is the consequence of stimulating reviews of only those parts of the application a reviewer is familiar with, and of the interactive online format where reviewers can comment on one another. We also found that there is less emphasis on the technicalities of the proposed project, which according to most interviewees is an advantage, as the emphasis in the review should be on the person's overall achievements and not on the project. This avoids too much focus on technical details, which detracts from the overall quality of an application. Only a small number of interviewees found this a disadvantage of the PC.[17]

Most participants in the experiment find the reviews and comments overall to be at least of the same quality as conventional reviews. The same holds for comprehensiveness: All aspects of the applications are covered in the PC review.

Is there a risk that for some of the applications none of the PC members has the required expertise? With the role of first reviewer, and with the possibility of adding further reviewers to the Peer Circle for specific (parts of) applications, this possible mismatch is solved according to most interviewees.

The main advantage mentioned is that in the Peer Circle a larger group of reviewers formulate their views and assessments of the applications. This has many advantages, such as making the review less dependent on the choice of one or two reviewers, more objectivity, better coverage of the various evaluation criteria, a better self-correction mechanism to avoid wrong assessments, and better final evaluations. This contributes to the quality of the review process. In contrast, the conventional review often includes only two perspectives: one of the external reviewer and one of the committee member.

Most of the committee members find the PC outcomes useful in formulating their advice to the committee, and the PC reviews do not seem to result in more questions and longer discussions in the committee – which is a positive outcome.

[17] Quite a few interviewees mentioned that the weight and content of the various criteria needs more clarification. As this is not a specific Peer Circle issue, we do not address that issue in this report.

Among the PC members there are also relatively inexperienced scientists. Most interviewees considered this heterogeneity as useful, and only a single committee member found that only experienced reviewers can come to meaningful contributions.

To summarize, the first question can be answered positively: the PC reviews have the required quality.

2. Premature consensus

The possibility of premature convergence of opinions did not materialize or hardly materialized in the experiment. Consensus emerged, but not too quickly and not without the possibility for articulating contradicting opinions – which happened regularly. The group interaction and the possibility to see other PC members' reviews and comments was seen as positive, helpful, and stimulating for reflexivity.

3. Acceptance

The participants in the Peer Circle experiment are not randomly selected but come from the AvH network. So, their opinion about the Peer Circle cannot be considered representative – which is a problem anyway because of the small size of the sample. On the other hand, they come from different research fields, creating sufficient disciplinary variety.
Within these constraints, the Peer Circle was appreciated overall and seen as a good way forward. Only three out of 28 reviewers preferred the conventional review approach. These three were from the Modern History PC, which consisted of ten reviewers in total.

There were several incentives for participating in the Peer Circle, such as curiosity about this new model for reviewing grant applications, and the perceived need to serve the community by participating in reviewing. Additionally, the PC members found it very useful to see how others review, which helps to improve review skills. Furthermore, by seeing a set of applications (instead of only one, as in the conventional review procedure), one learns what make a good and poor proposal – something that helps to improve one's own grant applications.

4. Efficiency

Workload and time used: The reviewers differ on whether the Peer Circle saves time. Writing time goes down as the texts are shorter, more informal, and one does not need to build an extensive argument leading to an overall conclusion. And when one agrees with comments/ assessments of others, one does not need to formulate these again. However, there is more reading than for a single review, which may not (always) be compensated by the reduced writing time.

However, the large majority feels that *per application* the PC procedure takes less time, and that the time the Peer Circle takes is reasonable. The logfiles show that the time spent online differs strongly between the reviewers.

Independently of the time used by the individual Peer Circle members, the PC concept should also be evaluated from the perspective of the scientific community. And for the community,

the PC saves time as the number of reviewers involved is much lower than in the conventional approach.

Opinions of committee members vary. Most of them report that their tasks take less time thanks to the PC input, for example because they do not have to read all the applications. For those applications where there is convincing unanimity in the PC, the committee member can adopt this conclusion and turn their attention to the other applications. Other committee members see this differently and report that they need to read all the applications as they did under the conventional approach. A few of them adopted a reviewer role for all or several applications.

Inviting reviewers requires much less time by AvH staff for the Peer Circle than for traditional peer review. On the other hand, the preparation of the material took much longer, as it was not integrated in the standard workflow. This should be resolved in future versions of the platform.

The PC needs coordination and moderation, and that is a new task for AvH staff. This task is crucial for the efficiency and effectiveness of the review process, and it was appreciated and effective. If the PC is to be implemented on a larger scale, coordinating Peer Circles may become the principal task of more AvH staff members.

To conclude, the PC saves time for the research community overall, and the PC members find their time investment reasonable. For AvH staff it saves time when it comes to finding reviewers, but more time is needed for the new task of moderating and coordinating the Peer Circle. Among the committee members the opinions diverge.

Other aspects of efficiency: Although the online platform is not completely suited for the tasks of the Peer Circle, most participants reported that the platform worked reasonably well and was not difficult to learn. The documents uploaded to the system were appreciated, although the navigation functions could be improved. The PC members found the material provided in the platform better organized than in the conventional peer review procedure.

The Peer Circle was new for all the members. It would be good to have PC members functioning for a few years, in order to accumulate experience.

Some PC members mentioned that the current schedule of the Peer Circles is not optimal and not efficient and hinders interaction. More structured planning of the Peer Circle processes may increase efficiency and save more time. An alternative could be to schedule it, e.g., in three shorter phases with their own deadlines: initial review phase, a first commenting phase, and a second commenting phase. Several PC members found it difficult to integrate the review tasks into their other activities, and suggested scheduling the Peer Circle longer in advance.

The translation of PC results into a proposal for the committee could become more efficient if committee members relied more on Peer Circle results, which some but not all committee members already do. It would help if the Peer Circles created a ranking, which could be coordinated by committee members. The roles of the committee members seem to need reconsideration, if the PC is to be implemented at a larger scale. This also holds for the AvH staff, as mentioned above.

Finally, one problem of the conventional review procedure is that reviews regularly come in late, which then hinders the decision-making process. Late reviews result in postponing the consideration of applications four months to the next round. The PC does not suffer from this problem, as all the work is done before the deadline – including because of the coordinating activities of AvH staff.

5. Are the best selected?

According to several committee members, the outcome of the PC selection process would not have been different if the conventional approach had been deployed. But were the best applicants selected? We addressed that issue, using bibliometric indicators for both Inorganic Chemistry (PC) and Solid State Chemistry (control group), as in these fields journal publications and top journals play a core role in scientific communication. The analysis shows for both fields that the overall bibliometric performance scores of granted applicants are no higher than those of rejected applicants, which suggests that publications, citations, and journal impact are less important in the decision-making process than expected from review reports and committee meeting observations. The approach to address this question is useful, but more quality aspects need to be included to come to a final answer, as well as a larger dataset.

6. Gender balance

It is often argued that the way in which review and selection processes are organized affects the outcome. Would the PC influence the gender balance in the outcomes of the selection process? A comparison of the success rates between the conventional method and the Peer Circle (in the eight fields included in this study) shows that the success rates for men and women are the same in both groups. At the level of the individual research fields, the outcomes differ considerably between fields and between years. Whether this represents bias or merit-based gender differences cannot yet be answered and would require more data and further analysis. For now, one may conclude that the Peer Circle is neutral in terms chances for women, but this issue needs further investigation.

10. Recommendations

The recommendations below emerge from this study:

- Continue with the Peer Circle for those funding schemes in which the emphasis is more on the applicant than on the project.
- Change the scheduling of the PC, so that there are fixed periods for the different tasks with their own deadlines. For example, a phase for reviews by first reviewers; a questioning and commenting phase; a last phase for discussion and possibly producing the ranking for the committee.
- In relation to this, rethink the tasks of the committee members.
- Implement a similar scoring scheme in the PC as is used in the conventional procedure.
- Plan the Peer Circle long in advance, so the specific activities can be better integrated into the schedule of PC members.
- Develop a protocol for what should be done in which phase of the PC. This may also lead to a more equal (and for some a higher) level of activity.
- As coordination of the PC is crucial, develop a protocol for this. In the experiment, this was done differently in the different Peer Circles.
- The new platform should be in English. Currently most of the communication within the Peer Circle is in German, which creates a barrier to participation for non-German-speaking AvH laureates and others.
- Future experiments may be organized on a larger scale as that makes more extended evaluation possible.