Individual versus general structured feedback to improve agreement in grant peer review: A randomized controlled trial

Jan-Ole Hesselberg (janohes@student.sv.uio.no)
Foundation Dam (Norway) https://orcid.org/0000-0001-5183-7933
Knut Inge Fostervold
University of Oslo: Universitetet i Oslo
Pål Ulleberg
University of Oslo: Universitetet i Oslo
Ida Svege
OsloMet - storbyuniversitetet

Research

Keywords: peer review, inter-rater agreement, funding, reliability, feedback, training

Posted Date: May 6th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-461626/v1

Abstract

Background

Vast sums are distributed based on grant peer review, but studies show that interrater reliability is often low. In this study, we tested the effect of receiving a short individual feedback report compared to a short general feedback report on the agreement between reviewers.

Methods

A total of 42 reviewers at the Norwegian Foundation Dam were randomly assigned to receive either a general feedback report or an individual feedback report. The general feedback group received one report before the start of the reviews that contained general information about the previous call in which the reviewers participated. In the individual feedback group, the reviewers received two reports, one before the review period (based on the previous call) and one during the period (based on the current call). In the individual feedback group, the reviewers were presented with detailed information on their scoring compared with the review committee as a whole, both before and during the review period. The main outcomes were the proportion of agreement in the eligibility assessment and the average difference in scores between pairs of reviewers assessing the same application.

Results

A total of 2398 paired reviews were included in the analysis. There was no difference between the two groups in terms of the average difference. There was a significant difference between the two groups in the proportion of absolute agreement on whether the application was eligible for the funding programme, with the general feedback group demonstrating a higher rate of agreement. However, the overall levels of agreement remained critically low in both groups and for both outcomes.

Conclusions

Individual feedback did not improve agreement between reviewers. This finding is in line with related studies of journal peer review. The low levels of agreement remain a major concern in grant peer review, and research to identify contributing factors as well as the development and testing of interventions to increase agreement rates are still needed.

Trial registration

The study was preregistered at OSF.io/n4fq3.

Background

Worldwide, vast sums are distributed through grants that use peer review processes. Despite this, evidence of the ability of peer review to identify the “best” future projects – however defined – seems

sparse. The latest Cochrane review of peer review to improve the quality of grant applications concludes that “there is little empirical evidence on the effects of grant-giving peer review” [1], and in the latest review of grant peer review, Guthrie et al. [2] conclude that there is “fairly clear evidence that peer review is, at best, a weak predictor of future research performance”. Because funding decisions relying on peer review have a considerable impact on how public funds are spent, the trajectories of a field's knowledge base and the careers of researchers, the lack of predictive value is a concern.

One factor that is likely to contribute to poor predictive value is low levels of agreement between reviewers assessing the same application. Several studies have shown that grant reviews have very low levels of agreement. A study of nine different programmes at the National Science Foundation (NSF) found average rating intraclass correlations (ICCs) between 0.18 and 0.37 [3], and a study of 23,414 ratings in six different research areas at the Austrian Science Fund found average rating ICCs between 0.38 and 0.55 [4]. In their review of the literature on grant peer review, Marsh et al. [5] conclude that “interrater reliability estimates are not adequate, falling well below acceptable levels”.

Without reasonable levels of agreement, interrater reliability will remain low, and the funding decisions based on it will become inconsistent. One reason for the low agreement levels might be different scoring styles and different interpretations of the scoring scale among reviewers. Without a common understanding of how to interpret and use the scale, differences in scoring among reviewers easily occur in the evaluation process. Reviewer training delivered by funders is one way to increase the level of shared understanding, but studies show that such initiatives are lacking. A survey of 57 international public and private grant-giving organizations found that only 9 % of the reviewers had received any formal training and that “64 % of the reviewers said they would be interested in receiving training if funding organizations provided it” [6].

In funding programmes where a set of reviewers are engaged to review many applications in multiple calls for proposals over some time, simple feedback on how the reviewers use the scale compared to other reviewers might be a cost-efficient way to calibrate reviewers’ use of the scale. A lenient reviewer with a tendency to use only the top scores might employ feedback in a way that makes him or her use the lower scores more frequently and distribute them more in accordance with the score distribution of the other reviewers. Reduced variability in the distribution of scores between reviewers will increase the likelihood of absolute agreement (similar scores by reviewers). Since this feedback might also contribute to larger variability in the distribution of scores within reviewers and thereby change the individual reviewer’s ranking of the applications, it might also increase relative agreement.

Research on the effects of reviewer training on any outcome, let alone reviewer feedback on agreement in grant peer review, is sparse. The most up-to-date systematic review of the effect of training in journal peer review included only five studies [7]. Based on the five qualified studies, the review concluded that "training did not improve the quality of the peer review report". Most studies identified in this review and in other training trials [8, 9] test the effect of structured training or mentoring sessions, and the outcomes seldom include interrater reliability or agreement as the outcome measure. In addition, most of the

studies focus on journal peer review rather than grant peer review. To our knowledge, the study by Sattler et al. [8] is the only study to test the effect of training on interrater reliability in grant peer review. They found that interrater reliability was significantly higher in the training group. However, in this study, the participants did not review applications; instead, they "received criteria consistent with specific rating scale values and were asked to assign ratings that were consistent with those criteria".

The objective of this study was to evaluate the effect on the agreement between reviewers of receiving a simple individual feedback report compared to a general feedback report. Our main hypothesis was that agreement, either in the assessment of eligibility or the scoring of proposals, would differ between reviewers who received individual feedback reports on their scoring and reviewers who received general and nonspecific feedback reports. Our secondary hypothesis was that the perceived usefulness of the feedback report would differ between reviewers who received an individual feedback report and those who received a general feedback report.

Methods

Trial design

We conducted a randomized controlled pretest-posttest design with two parallel groups. One group of reviewers received an individual feedback report, and the other group received a general feedback report. The study was preregistered at Open Science Framework (osf.io/n4fq3).

Participants

The participants included in this study comprise all reviewers who served in the funding programme Health ("Helse") at the Norwegian Foundation Dam for the year 2018. The program has two calls for proposals every year, and each call has two different review committees (Health spring and Health fall), with 24 reviewers in each.

Reviewers were excluded from participation in the study if they had not served on the corresponding review committee the previous year. The reason for this exclusion criterion was that the individual feedback report was based on the reviews that the reviewer had conducted the previous year. Reviewers who served on both committees were treated as members of the Health spring committee.

All participants signed an informed consent form (electronically) before enrolment, and no participants were paid to participate in the study. Baseline characteristics of the participants included age, gender and years of experience as a reviewer for the foundation.

Interventions

Two interventions were delivered: general feedback and individual feedback. Both were delivered electronically in the form of one-page reports that were designed and generated in an Excel workbook. The general feedback was designed as a control intervention. It consisted of one simple report containing

information only on the total number of applications rated by the review committee as a whole in the previous call in 2017 and a line chart of the distribution of the committee's review scores. The reviewers in the general feedback group received the report two days before the start of the review period (Figure 1). The individual feedback intervention consisted of two reports with detailed information on each recipient's reviews compared to the reviews by the committee as a whole. The first individual feedback report was delivered three days before the start of the review period and was based on the reviews from the previous call in 2017, and the second was delivered ten days before the end of the review period and was based on the reviews conducted to date in the ongoing call (Figure 1).

Both individual feedback reports comprised information on the reviewer's assessment of eligibility, scoring of the applications, and level of agreement with other reviewers of the same applications. This information was presented in tables, figures and text. The report included a table showing the proportion of applications the reviewer rated as ineligible and a table showing the reviewer's average review scores compared to the review committee as a whole. In addition, the report contained a graph showing the distribution of the reviewer's scores compared to the score distribution of the committee as a whole. The tables and charts were accompanied by standardized text segments describing the same information. Translated versions of the two reports can be viewed at dam.no/feedback.

Outcomes

Main outcomes: agreement

All applications were reviewed by three reviewers independently of each other. Each review was conducted in two steps in an electronic review form related to each application. First, the reviewers decided whether the applications fulfilled the eligibility criteria for the funding programme by answering "yes" or "no" to the question, "Is the proposal eligible for the programme?" Second, the reviewers were asked to score the quality of the application by giving the application a total score "on a scale from one to ten, where ten is the best".

The main outcome measures were as follows:

1. the average absolute difference in application score between pairs of reviewers within the same intervention group who had reviewed the same application divided by the number of pairs. If one application was reviewed by three reviewers from the same group, the formula would be (|Review1-Review2| + |Review1-Review3| + |Review2-Review3|) / 3. In the following, this is referred to as the "average absolute difference";
2. the agreement in the scoring of eligibility between pairs of reviewers within the same intervention group who had reviewed the same application, categorized as "agree" or "disagree". In the following, this is referred to as "eligibility agreement".

Secondary outcomes: perceived usefulness

The secondary outcome measure was the reviewers' perceived usefulness of the report. All reviewers received a web-based survey with two questions regarding the perceived usefulness of the interventions. They answered the following questions:

- "To what degree did you find the feedback you received useful?" on a five-point Likert scale (To a very small degree, To a small degree, To some degree, To a large degree or To a very large degree).
- "If you were offered this feedback next time, would you want it?" by choosing "yes", "no" or "I don't know".

Changes to outcomes

In the study, preregistration agreement regarding the application score was defined as the "difference in total review score of 0 or 1". Consequently, the planned main outcome measure was dichotomous ("agree" and "disagree"). However, the average absolute difference provided a continuous, more fine-grained measure of agreement, and it was therefore chosen as the main outcome in the study. Both outcomes were derived from the reviewer's quality scoring, and the predefined outcome measure was used in the sensitivity analyses.

Sample size

We used a convenience sample. All eligible reviewers were included as participants. Hence, no power analysis was conducted prior to inclusion.

Randomization

We used a block randomization procedure. There were two blocks, and each committee (fall and spring) constituted one block.

Blinding

This was a double-blinded trial. The reviewers were not informed about the differences between the two interventions provided. They were only told that they would receive one of two feedback reports and did not know what information the other group received in their report. The personnel who interacted directly with the reviewers were not aware of the assigned interventions. In addition, the author who conducted the initial statistical analysis (IS) was blinded and was provided with anonymized data sets.

Statistical methods

Main outcome: agreement

To examine whether the intervention had an effect on the difference in scores between reviewers, a linear mixed-effects regression model (LMM) was used. The use of an LMM was preferred since three reviewers were nested within the same application, yielding at most three difference scores for each application given that all reviewers found the application eligible. Thus, the reviewers' difference in scores was

defined as level 1 in the model, and applications were defined as level 2. The difference in score was the dependent variable, and random intercepts for each reviewer were included in the linear mixed-effects model. The covariance matrix of within-subject measurements was variance components. The time point (baseline vs. follow-up) and group (individual feedback report vs. general feedback report) were included as dummy-coded fixed effects. We also included the interaction between group and time point as a fixed effect to estimate whether the change in the difference score from baseline to follow-up was dissimilar in the two groups. A significant interaction effect would indicate a significant effect of the individual feedback report. To examine the effect of the intervention on eligibility agreement, we used a linear mixed binary logistic regression model. In this case, the rate of eligibility agreement was the dependent variable.

In addition, we calculated the intraclass correlation for the average absolute difference and Cohen's kappa for the eligibility agreement for the two groups at baseline and follow-up. Since the raters were not the same for all subjects, we used the one-way random effects model, ICC [1, 2], for the average absolute difference.

The LMM analyses were a deviation from the preregistered analysis plan. Initially, we planned to calculate the proportion of absolute agreement (for both eligibility and application score) and compare the groups using Fisher's exact test. The LMM analysis provided a more fine-grained measure of agreement and the possibility to control for the interaction between group and time point. The preplanned analyses were also conducted and included as a sensitivity analysis. All analyses were performed according to the intention-to-treat principle.

Secondary outcome: perceived usefulness

We used the Mann-Whitney U test to compare the reviewers' evaluation of the usefulness of the feedback report between the two groups.

Results

Participant enrolment and characteristics

The 43 reviewers assigned to review applications in the funding programme Health at the Norwegian Foundation Dam for the year 2018 were assessed for eligibility (Figure 2). One was excluded due to not being part of the review committee the previous year. The remaining 42 reviewers were included in the study and randomized to either the general feedback group (n=23) or the individual feedback group (n=19). Five reviewers served in both calls and were thus excluded from the Health fall call.

One participant in the general feedback group could not perform his reviews due to acute illness. Hence, he did not receive the allocated intervention. Follow-up data and data on compliance were retrieved for the remaining 41 participants, and none of them were excluded from the analyses. All participants in the general feedback group and 95 % of the participants in the individual feedback group confirmed that they had received and read the feedback report.

Participant characteristics were similar in the two intervention groups (Table 1).

Table 1. Baseline characteristics of the study participants.

Baseline measures	General feedback group (n=22)	Individual feedback group (n=19)	Total (n=41)
Age, years	58 ± 11,4	49 ± 9,4	54 ± 11,4
Women, count (%)	14 (64 %)	9 (47 %)	23 (56 %)
Years of experience as reviewer for the foundation	2,6 ± 2,97	2,4 ± 3,17	2,5 ± 3,08
Reviews assessing the application as eligible, count (%)	651 (91 %)	484 (89 %)	1135 (90 %)
Application score (1-10)	6,3 ± 1,93	5,7 ± 2,14	6,1 ± 2,05
Average absolute difference	2,0 ± 1,54	2,2 ± 1,59	2,1 ± 1,56

Values are mean ± SD unless otherwise stated

Numbers analysed

A total of 2398 paired reviews were analysed (Table 2). In 2038 of these cases, the two reviewers agreed that the application was eligible, and the average absolute difference could be calculated.

Table 2. Numbers analysed at baseline and follow-up.

	General feedback group (n=22)			Individual feedback group (n=19)		
	Baseline	Follow-up	Total	Baseline	Follow-up	Total
Total number of paired reviews	715	642	1357	545	496	1041
Number of applications included in analyses	511	450	961	409	376	785
Number of paired reviews included in analyses of average absolute difference	601	594	1195	434	409	843
Number of paired reviews included in eligibility agreement analyses	715	642	1357	545	496	1041

Outcomes and estimation

Main outcome: agreement

The results from the LMM analyses (Table 3) showed that the time×group interaction effect was not significant, indicating an equal decrease in the difference score over time for the individual feedback and the general feedback groups. There was an overall significant decrease in the difference score from baseline to follow-up (b = - 0.32, p =.004). At baseline, the reviewers within the general feedback group had a lower average absolute difference compared to the individual feedback group (b = - 0.24, p = .020).

Table 3. Linear mixed regression model analysis estimating the change in differences score over time by group.

	b (*se*)	*t*-value	*p*	95 % CI
Intercept	2.20 (0.08)	28.593	.000	[2.05, 2.35]
Time[a]	-0.32 (0.11)	-2.924	.004	[-0.54, -0.11]
Group[b]	-0.24 (0.10)	-2.329	.020	[-0.44, -0.04]
Time×Group	0.17 (0.15)	1.206	.228	[-0.11, 0.46]

A total of 2038 average absolute differences nested within 1500 applications were included in the analysis. At baseline, 434 difference scores were from reviewers in the intervention group, and 601 difference scores were from reviewers in the control group. At follow-up, 409 difference scores were from reviewers in the intervention group and 594 from reviewers in the control group. [a]Baseline=0, follow-up=1, [b]Individual feedback group=0, General feedback group=1.

The ICC (one-way random, average measures) at baseline for the general feedback group and individual feedback group was 0.276 and 0.323, respectively. At follow-up, the values were 0.303 for the general feedback group and 0.401 for the individual feedback group (Figure 3).

The mean application score at baseline in the general feedback group was 6.3 (95 % CI from 6.19 to 6.49), and the mean score in the individual feedback group was 5.7 (95 % CI from 5.52 to 5.90) (Table 4). At follow-up, the mean score in the general feedback group was 6.6 (95 % CI from 6.46 to 6.76), and the mean score in the individual feedback group was 6.2 (95 % CI from 6.04 to 6.41).

Table 4. Application eligibility, application score and average absolute difference.

	General feedback group (n=22)		Individual feedback group (n=19)	
	Baseline	Follow-up	Baseline	Follow-up
Reviews assessing the application as eligible, count (%)	651 (91 %)	613 (96 %)	484 (89 %)	442 (89 %)
Eligibility agreement, count (%)	612 (86 %)	599 (93 %)	453 (83 %)	417 (84 %)
Application score (1-10)	6,3 ± 1,93	6,6 ± 1,85	5,7 ± 2,14	6,2 ± 1,97
Average absolute difference	2,0 ± 1,54	1,8 ± 1,47	2,2 ± 1,59	1,9 ± 1,48

Values are mean ± SD unless otherwise stated

The results from an LMM analysis (Table 5) found no main effect of either time or group but a significant time×group interaction effect (b = 0.77, p = .006, OR = 2.17). The interaction effect indicated an increase in the proportion of eligibility agreement over time for the general feedback group only.

Table 5. Linear mixed binary logistic regression model analysis of eligibility agreement over time by group.

	b (*se*)	*t*-value	*p*	*Odds Ratio* [95 % CI]
Intercept	1.65 (0.13)	12.6	< .001	5.19 [4.02, 6.70]
Time[a]	0.08 (0.19)	0.43	.667	1.09 [0.75, 1.58]
Group[b]	0.19 (0.18)	1.05	.293	1.21 [0.85, 1.71]
Time×Group	0.77 (0.28)	2.74	.006	2.17 [1.25, 3.79]

A total of 2398 paired eligibility assessments nested within 1746 applications were included in the analysis. At baseline, 545 paired eligibility assessments were from reviewers in the intervention group and 715 paired eligibility assessments were from reviewers in the control group. At follow-up, 496 paired eligibility assessments were from reviewers in the intervention group and 642 from reviewers in the control group. [a]Baseline=0, follow-up=1, [b]Individual feedback group=0, General feedback group=1.

Cohen's kappa at baseline for the general feedback group and the individual feedback group was 0.097 and 0.197, respectively. At baseline, the values were 0.154 for the general feedback group and 0.082 for the general feedback group (Figure 4). The rate of eligibility agreement was high in both groups at baseline and follow-up, ranging from 83 % to 93 % (Table 4).

Secondary outcome: perceived usefulness

Table 6 displays the perceived usefulness of the interventions. The results show that 95 % (n=18) in the individual feedback group and 68 % (n=15) in the general feedback group responded to the question “To what degree did you find the feedback you received useful?” after finishing the reviews. An independent sample t-test of the mean scores showed that there was no significant difference between the groups (*p*=0.442).

Table 6. Perceived usefulness of the interventions.

	General feedback group (n=15)	Individual feedback group (n=18)
“To what degree did you find the feedback you received useful?”, mean score (scale 1-5)*	3,5 ± 0,74	3,7 ± 0,75
“If you were offered this feedback next time, would you want it?”, the proportion of “Yes”	80 % (n=12)	94 % (n=17)

* The question was answered on a five-point Likert scale (To a very small degree, To a small degree, To some degree, To a large degree or To a very large degree) and coded from 1 (To a small degree) to 5 (To a very large degree).

Sensitivity analyses

The sensitivity analysis showed no difference between the two groups in the proportion of absolute agreement on application score, defined as a difference in score of 0 or 1 being similar (p=1.000), supporting the findings of the main analysis of average absolute differences.

There was a significant difference between the two groups in the proportion of absolute agreement on whether the application was eligible for the funding programme, with the general feedback group demonstrating a higher rate of eligibility agreement (p<.01) in the 2018 review.

Discussion

The results of this study revealed that agreement between reviewers who received an individual feedback report did not improve compared to reviewers who received a general feedback report. In addition, the increase in agreement on application eligibility was significantly higher in the general feedback group.

To our knowledge, the present study is the first to evaluate the effect of training interventions in grant peer review in a real-world setting. One previous controlled study investigated the effect of reviewer training in grant peer review and found that interrater reliability was significantly improved in the training group [8]. The training applied in the previous study was an 11-minute training video focusing on the general importance of the review and “how to assign evaluation scores”. Hence, both the training intervention and the review task differed from our study.

Furthermore, journal peer review has several similarities to the review of grant applications, and research regarding training for journal review may also be relevant for funders. In a systematic review, Bruce et al. [7] found that training interventions to improve peer review in biomedical journals had a limited effect on the quality of the review report as assessed by journal editors.

The lack of effect of individual feedback compared to general feedback on scoring agreement may be related to the content and the simplicity of the individual feedback reports. The part of the individual feedback report addressing the reviewers' scoring history provided the reviewers with information on their average score and their score distribution compared to the committee as a whole. It did not provide specific information on how the reviewers should re-score the applications, interpret the content of the applications or interpret and weigh the different criteria. The results suggest that merely adjusting the scores to align better with the average distribution of scores might increase agreement somewhat, but likely not by much. Supplementation of the individual feedback report with more comprehensive guidance on interpretation and possible actions could have been beneficial. Nevertheless, an automatically generated feedback report is probably not suitable for providing such specific guidance.

In the part of the individual feedback report addressing eligibility, reviewers who had rated many applications as non-eligible were advised to reconsider the number of non-eligible applications. This guidance did not increase the proportion of reviews assessing the application as eligible or the proportion of agreement on eligibility in the individual feedback group. However, in the general feedback group, for which no information on eligibility was provided, there was a significant increase in the proportion of agreement on eligibility compared to the general feedback group. We have no plausible explanation for this difference and suspect this is an artefact. It should be noted that the absolute level of agreement regarding eligibility was high in both groups at both baseline and follow-up.

Limitations

One potential limitation of the study is that the score distribution provided in the general feedback report might have affected the reviewers in this group in a similar way as the reviewers who received the individual feedback report. A group that was offered no feedback would have provided a comparison with the usual process but would remove any possibility for blinding the participants (as was the intention with providing the general feedback). Combined with the limited number of reviewers, this means that we cannot rule out any potential beneficial effects of the individual feedback report in a study with sufficient power and a more heterogeneous sample.

Conclusions

Several factors may influence and decrease agreement between raters. Theoretically, training interventions have the potential to make reviewers focus on and value the same aspects of an application and to interpret and use the review scoring instrument more uniformly. However, to accomplish this, the training intervention should also focus on how to interpret the review criteria and

address other factors that are associated with low agreement levels in the specific setting in which they are intended to be used.

In light of this and to ensure compliance and uniform interpretation of the report, a separate training session with the reviewers might be necessary. The feedback report combined with a training session would, however, be a more complex intervention and would have to be tested in a separate study.

Despite the lack of effect of the individual feedback report compared to the general feedback report on scoring agreement, the individual feedback report might still be considered useful. The overall agreement increased significantly from baseline to follow-up. There might be causes other than the reports for this increase (e.g., more reviewer experience). However, given this increase and the fact that the intervention can be provided using a simple spreadsheet template and that the reviewers perceived the feedback report as useful, it might be reasonable to provide the report.

Even with the increase in agreement from baseline to follow-up, agreement levels were still critically low. Previous studies and reports have shown that this is a major concern across different programmes, funders and journals [2, 4, 5]. Research to identify factors contributing to this phenomenon as well as the development and testing of interventions to increase agreement rates are needed. Furthermore, an attempt should be made to justify costly peer review processes by evaluating their validity. A sufficient number of reviewers is crucial to ensure acceptable overall levels of reliability of the application selection process as a whole.

Abbreviations

ASE
Asymptotic standard error
CI
Confidence interval
ICC
Intra-class correlation
LMM
Linear mixed-effects regression model

Declarations

Ethics approval and consent to participate

All participants provided written informed consent before inclusion. According to Norwegian law, ethics approval is not required for studies of this kind.

Consent for publication

Not applicable

Availability of data and materials

Data and materials were uploaded to and freely available at OSF.io/6rdvc under the CC BY-NC licence (creativecommons.org/licenses/by-nc/4.0).

Competing interests

JOH is the chief programme officer and IS is the head of development at the funder of this study, the Norwegian Foundation Dam (Stiftelsen Dam).

Funding

The Norwegian Foundation Dam was the sole funder of this study. The authors JOH and IS are employees at the funder.

Authors' contributions

JOH and IS designed the intervention and the study. JOH delivered the intervention, collected responses from the participants and prepared the datasets for blinded analyses. ICS and PU performed the analyses. All authors were involved in drafting and critically revising the manuscript for important intellectual content. All authors have given final approval of the version to be published, have participated sufficiently in the work to take public responsibility for appropriate portions of the content, and agree to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

Acknowledgements

The authors wish to thank the reviewers who participated in the study and Secretary General Hans Christian Lillehagen at the Foundation Dam for considerable support. In addition, the employees at the Foundation Dam, particularly Jan Hoel Gulbrandsen, Jan Melby and Taran Søberg, provided valuable assistance in delivering the intervention.

References

1. Demicheli V, Di Pietrantonj C. Peer review for improving the quality of grant applications. Cochrane Database Syst Rev. 2007:MR000003. doi:10.1002/14651858.MR000003.pub2.
2. Guthrie S, Ghiga I, Wooding S. What do we know about grant peer review in the health sciences? F1000Res. 2017;6:1335.
3. Cicchetti DV. The reliability of peer review for manuscript and grant submissions: a cross-disciplinary investigation. Behav Brain Sci. 1991;14:119-35.

4. Mutz R, Bornmann L, Daniel HD. Heterogeneity of inter-rater reliabilities of grant peer reviews and its determinants: a general estimating equations approach. PLoS One. 2012;7:e48509.
5. Marsh HW, Jayasinghe UW, Bond NW. Improving the peer-review process for grant applications: reliability, validity, bias, and generalizability. Am Psychol. 2008;63:160-8.
6. Schroter S, Groves T, Højgaard L. Surveys of current status in biomedical science grant review: funding organisations' and grant reviewers' perspectives. BMC Med. 2010;8:62.
7. Bruce R, Chauvin A, Trinquart L, Ravaud P, Boutron I. Impact of interventions to improve the quality of peer review of biomedical journals: a systematic review and meta-analysis. BMC Med. 2016;14:85.
8. Sattler DN, McKnight PE, Naney L, Mathis R. Grant peer review: improving inter-rater reliability with training. PLoS One. 2015;10:e0130450.
9. Wong VSS, Strowd RE, 3rd, Aragón-García R, Moon YP, Ford B, Haut SR, et al. Mentored peer review of standardized manuscripts as a teaching tool for residents: a pilot randomized controlled multi-center study. Res Integr Peer Rev. 2017;2:6.

Figures

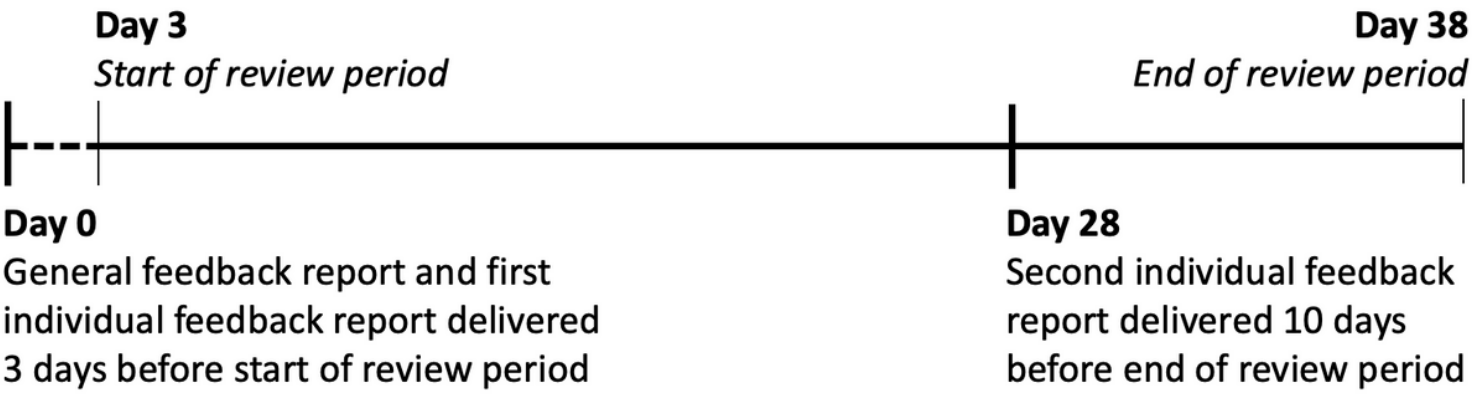

Figure 1

Timeline of intervention delivery in the review periods of 2018.

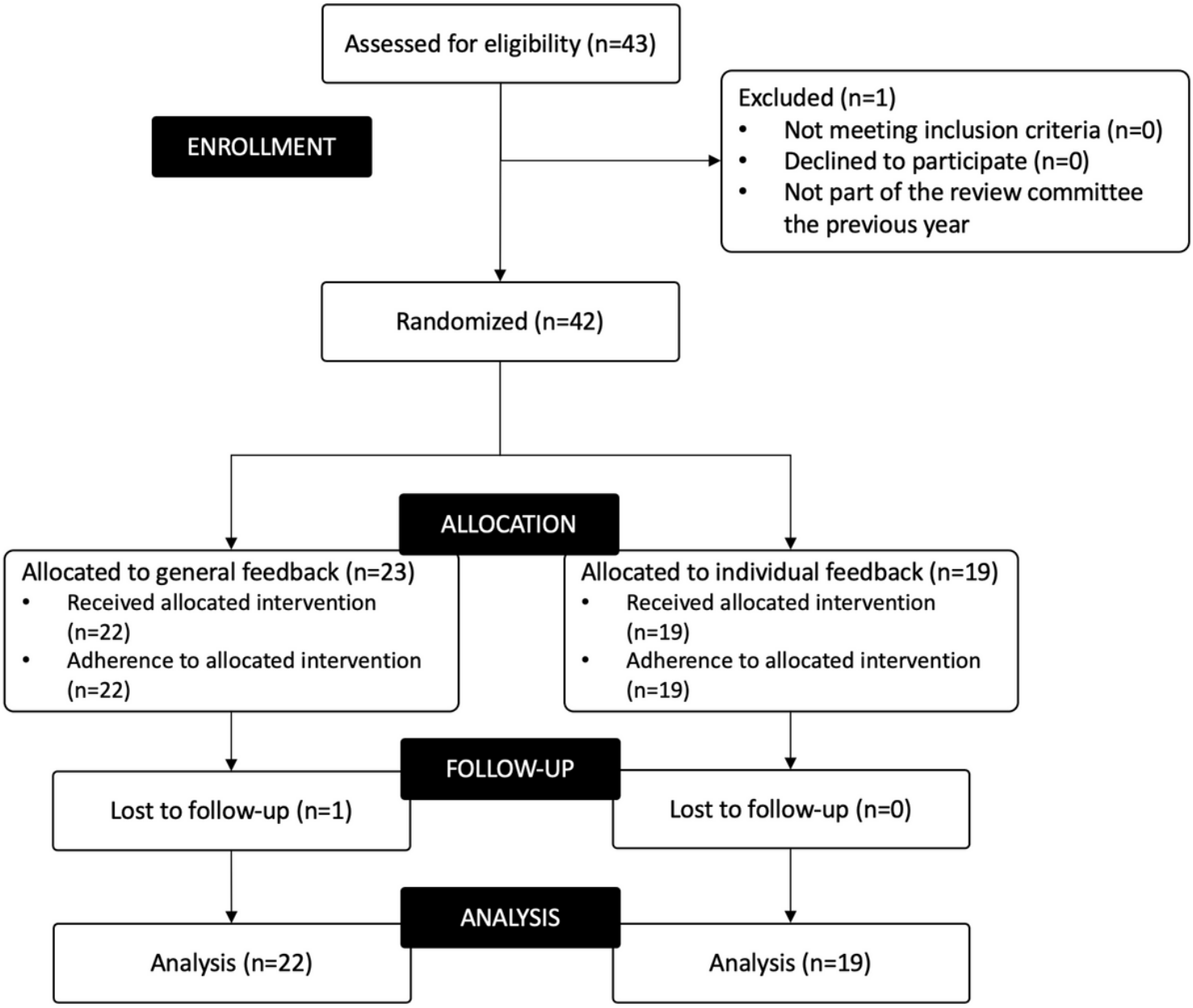

Figure 2

Study flowchart. Enrolment, randomization and follow-up of study participants.

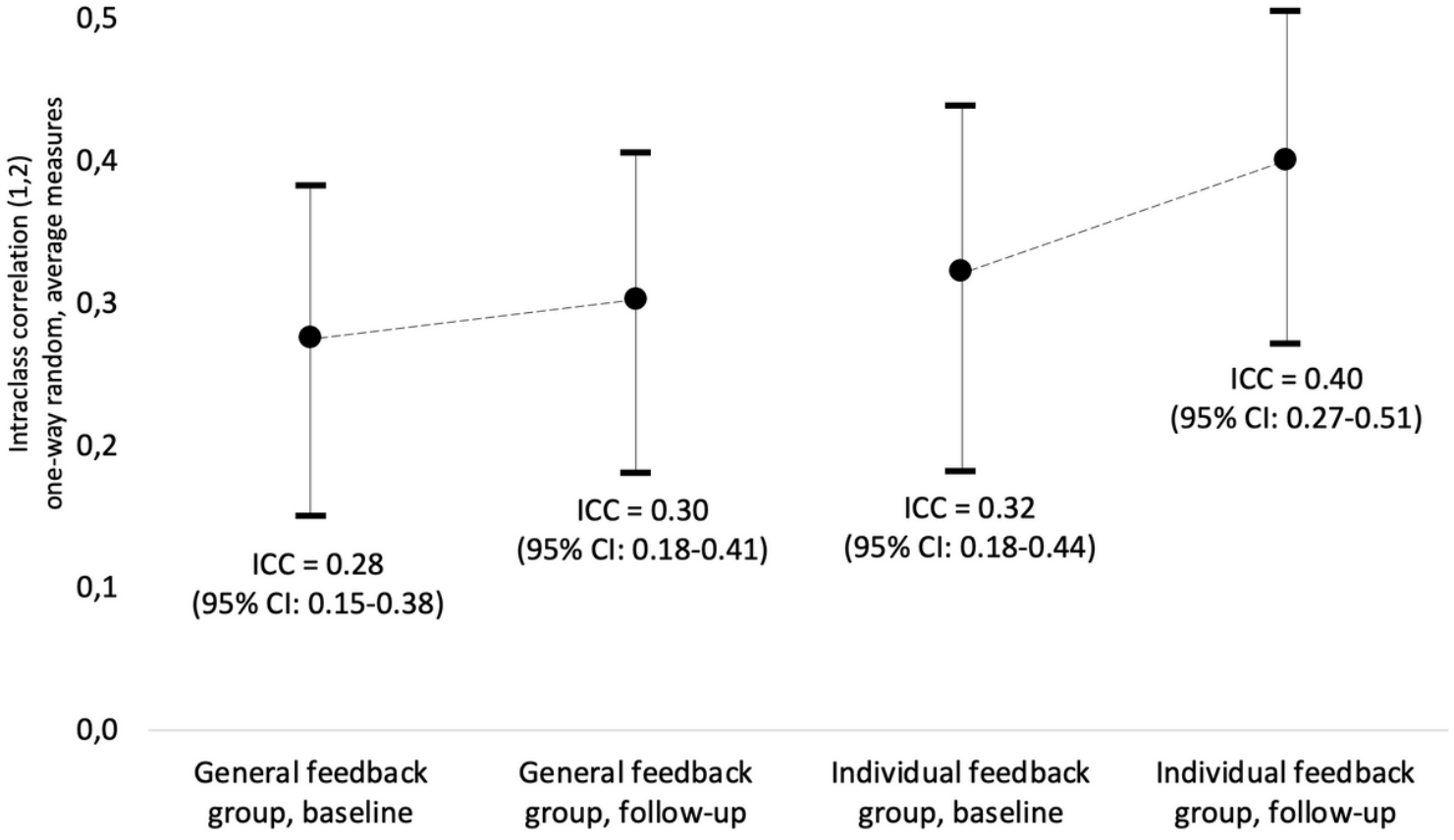

Figure 3

Intraclass correlation [1,2] in the general feedback and individual feedback groups at baseline and follow-up.

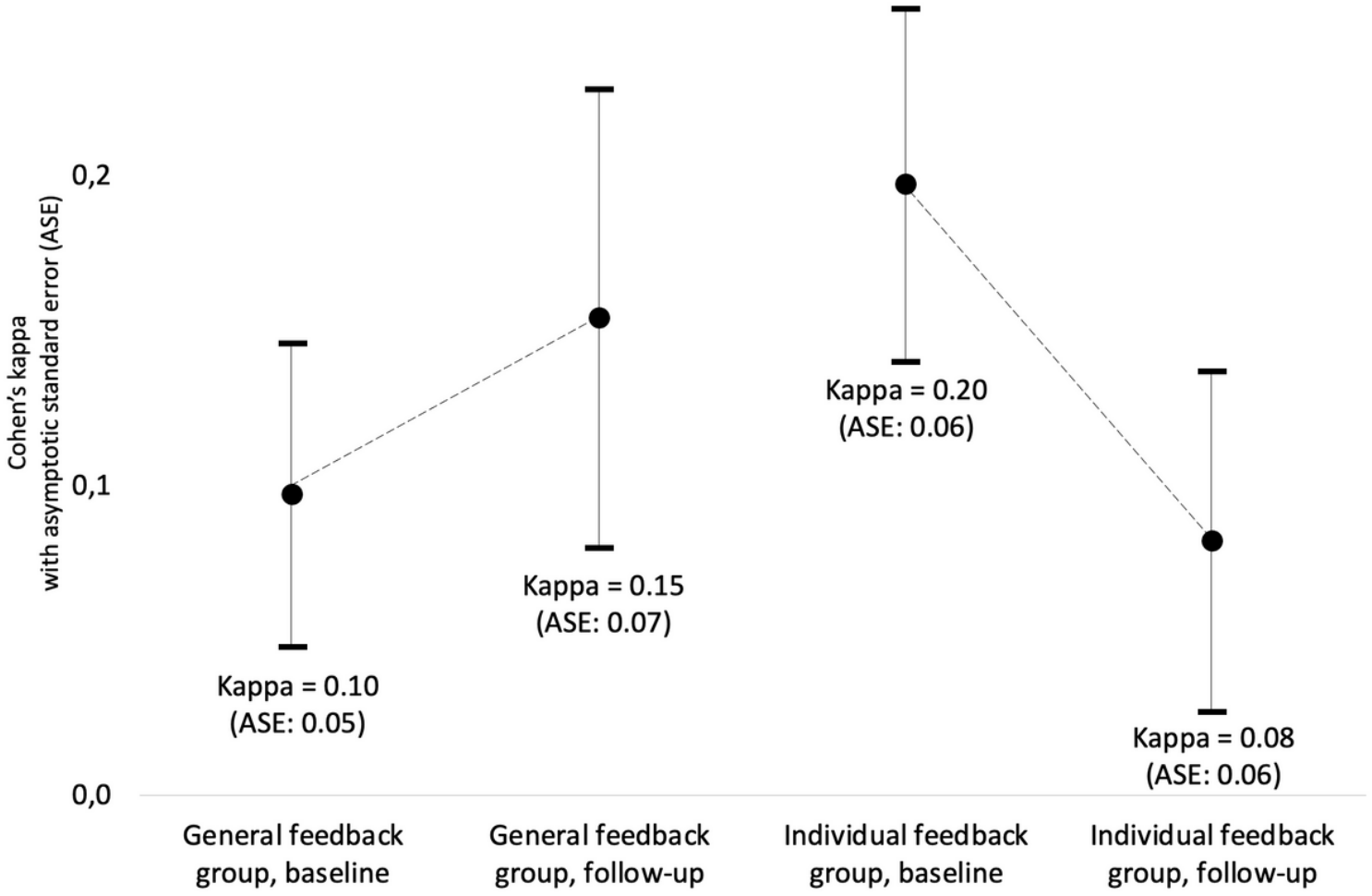

Figure 4

Cohen's kappa in the general feedback and individual feedback groups at baseline and follow-up.