Research Evaluation, 2021, 1–12
doi: 10.1093/reseval/rvab011
Article

Does reviewing experience reduce disagreement in proposals evaluation? Insights from Marie Skłodowska-Curie and COST Actions

Marco Seeber [1,*], Jef Vlegels[2], Elwin Reimink[3], Ana Marušić[4] and David G. Pina[5]

[1]Department of Political Science and Management, University of Agder, Kristiansand, Norway, [2]Department of Sociology, Ghent University, Ghent, Belgium, [3]European Cooperation in Science and Technology (COST), Brussels, Belgium, [4]Department of Research in Biomedicine and Health, University of Split School of Medicine, Split, Croatia and [5]Research Executive Agency, European Commission, Brussels, Belgium

*Corresponding author. Email: marco.seeber@uia.no

Abstract

We have limited understanding of why reviewers tend to strongly disagree when scoring the same research proposal. Thus far, research that explored disagreement has focused on the characteristics of the proposal or the applicants, while ignoring the characteristics of the reviewers themselves. This article aims to address this gap by exploring which reviewer characteristics most affect disagreement among reviewers. We present hypotheses regarding the effect of a reviewer's level of experience in evaluating research proposals for a specific granting scheme, that is, scheme reviewing experience. We test our hypotheses by studying two of the most important research funding programmes in the European Union from 2014 to 2018, namely, 52,488 proposals evaluated under three funding schemes of the Horizon 2020 Marie Sklodowska-Curie Actions (MSCA), and 1,939 proposals evaluated under the European Cooperation in Science and Technology Actions. We find that reviewing experience on previous calls of a specific scheme significantly reduces disagreement, while experience of evaluating proposals in other schemes—namely, general reviewing experience, does not have any effect. Moreover, in MSCA—Individual Fellowships, we observe an inverted U relationship between the number of proposals a reviewer evaluates in a given call and disagreement, with a remarkable decrease in disagreement above 13 evaluated proposals. Our results indicate that reviewing experience in a specific scheme improves reliability, curbing unwarranted disagreement by fine-tuning reviewers' evaluation.

Key words: project evaluation; peer review; reliability; reviewing experience; reviewers disagreement; MSCA and COST Actions

1.Introduction

Grant funding represents a major channel of resources for research activity. However, the evaluation of the quality and potential impact of research proposals presents several challenges. Scholars have explored factors that affect evaluation scores and may hinder the identification of the best proposals. For example, Boudreau et al. (2012) found a negative bias towards novel research proposals in medicine, and several studies showed that interdisciplinary proposals have a lower chance of receiving funding (Laudel 2006; Sandström and Hällsten 2008; Bromham et al. 2016; DFG 2016). Scholars debate whether the gender of the applicant significantly affects research proposal evaluation and related funding opportunities (Mutz et al. 2012a; Van der Lee and Ellemers 2015; Volker and Steenbeek 2015). Reviewers in general give lower scores when

applicants belong to a different scientific domain, and higher scores when they have a conflict of interest with the proposal (Tamblyn et al. 2018). Furthermore, early funding is an asset for acquiring later funding, regardless of scientific achievements, suggesting a 'rich-club' or 'Matthew' effect (Bol et al. 2018).

A major issue in the evaluation of research proposals is low inter-reviewer reliability (IRR), namely, a high level of disagreement between reviewers in the score assigned to the same research proposal (Hodgson 1997; Jayasinghe et al. 2003; Mayo et al. 2006; Marsh et al. 2008; Mutz et al. 2012b; Pier et al. 2018). This problem also occurs in peer review of scientific articles (Cicchetti 1991; Bornmann et al. 2010; Bornmann and Daniel 2010) and it is known as the 'luck of the reviewer draw', meaning that the fate of a manuscript is largely predetermined by the selection of reviewers (Cole and Simon 1981).

While some level of disagreement between reviewers is arguably unavoidable and even desirable, unwarranted sources of disagreement can be harmful because they threaten the ability to identify the best proposals,[1] in addition to the legitimacy of grant evaluation as a procedure to allocate resources (Roumbanis 2019). Existing research suggests that specific traits of the proposal, the applicants, and the reviewer(s), may affect the level of disagreement in the evaluation of a proposal. Disagreement is stronger for proposals in the Social Sciences and Humanities than in other disciplines (Mallard et al. 2009; Lamont and Huutoniemi 2011; Mutz et al. 2012b; Pina et al. 2015). Disagreement between reviewers may depend on the fact that reviewers have different scientific expertise and backgrounds, affecting their assumptions about appropriate goals, methods, and theories. It may also be due to differences in their individual preferences, or differences in the biases affecting reviewers' evaluations: in general, for example, towards higher or lower scores or towards a specific proposal (Marsh et al. 2008). Moreover, several studies found that strictness or leniency in evaluation systematically vary as a function of the social categories to which reviewers belong (Lee et al. 2013), with variations observed by gender—that is, female reviewers being stricter than their male colleagues (Jayasinghe et al. 2003; Borsuk et al. 2009; Lane and Linden 2009; Wing et al. 2010), disciplinary affiliation (Lee and Schunn 2011), and nationality (Wood 1997; Marsh et al. 2008).

However, while such differences in peer evaluations may represent sources of contamination or error in assessments of a submission's true quality value (Marsh et al. 2008), research exploring what factors affect disagreement in project evaluation did not yet consider the individual characteristics of the reviewers. In fact, the key limitation of the existing research is that disagreement has been explored at the proposal level, hence focusing on the characteristics of the proposal or the applicants while neglecting the characteristics of the reviewers themselves.

This article aims to address this gap by studying disagreement at the reviewer level, exploring which characteristics of the reviewers affect reliability of evaluation; most importantly a reviewer's level of experience in evaluating research proposals of a specific granting scheme.

We do this by studying two of the most important research funding programmes in the European Union from 2014 to 2018, namely, 52,488 proposals evaluated under three funding schemes of the Horizon 2020 Marie Sklodowska-Curie Actions (MSCA), and 1,939 proposals evaluated under the European Cooperation in Science and Technology (COST) Actions.

In the next section, we present our hypotheses on how reviewing experience affects reliability. Section 3 presents the data and method, and the results are described in the Section 4. In the final section, we discuss the article's main findings, their theoretical and practical implications, and suggest ways to improve reliability of research proposal evaluation.

2.Theoretical framework

2.1 Reliability in grant evaluation

IRR has been mainly studied in the context of peer-reviewed journal manuscripts. In their seminal study, Peters and Ceci (1982) resubmitted 12 articles that had been previously published and 8 were rejected. Review articles of Cicchetti (1991) reported a very low IRR ranging from 18% (or 82% of disagreement in reviewers' recommendations) to 57%, whereas Weller's (2001) reported peer agreement from 14 to 98%, with an average of 49%. Disagreement is typically stronger in Social Sciences and Humanities journals, and more consensus is observed in certain hard sciences (specialized physics) than in others (general physics, medicine, and the behavioural sciences) (Cicchetti 1991), which may depend on the fact that disciplines have different degrees of consensus on research priorities, appropriate research methods, and theories (Hargens 1988, 1990).

Relatively little research has specifically focused on the IRR between the reviewers of research grant proposals. The evaluation of research proposals typically encompasses a first phase where reviewers independently assess and score the proposal, without any contact among them, followed by a second 'consensus' phase leading to the final score, typically with reviewers discussing and finally agreeing. IRR in proposal evaluation refers to the extent to which individual with independent reviews of the same proposal agree with each other. Studies that have looked at the grant application review process have reported low agreement in medical research (Mayo et al. 2006), and in the cross-disciplinary Australian Research Council (Marsh et al. 2008). Hodgson (1997) submitted the same 256 proposals simultaneously to two agencies and found that the level of agreement in the evaluations of the two agencies was only slightly above the agreement expected from two random sets of evaluations. Pier et al. (2018) studied 43 individual reviewers' ratings and written critiques of the same group of 25 National Institute of Health grant applications; they found no agreement among reviewers either in the qualitative and quantitative evaluations of the same proposal, or in how reviewers 'translated' a given number of strengths and weaknesses into a numeric rating. Mutz et al. investigated a sample of more than eight thousand grants submitted to the Austrian Science Fund and reported discipline-dependent and generally low levels of reviewer agreement (Mutz et al. 2012b).

Jayasinghe et al. (2003) found that some traits of the applicants (e.g. status of the university of affiliation) and of the reviewer (e.g. being professors or not) systematically affected evaluation scores of proposals to the Australian Research Council. In later studies (Jayasinghe et al. 2006; Marsh et al. 2008), they argued that a major factor contributing to low IRR in grant evaluation is that each reviewer only scores one or a few submissions. In response, they proposed and tested a 'reader system' in which a reader (i.e. reviewer) reads the first proposal, and then a second proposal. After comparing the two, the reviewer rates the first proposal before moving onto the third, and rating it in comparison with the second, and so on until the last proposal. They found that the system substantially increased IRR.

Disagreement in grant proposal evaluation is not necessarily problematic. Relying on several reviewers, instead of one reviewer, ensures that there is an integration of different views and opinions (Olbrecht and Bornmann 2010). Two or more experts have access to different resources and information than one expert, so their combined assessment is more likely to reflect an academic debate. Consequently, some level of disagreement may indicate that experts represent various views on what is good and valuable research (Langfeldt 2001), whereas in certain cases very high agreement can signal that the pool of reviewers is not diverse enough, leading to redundant information (Bailar 1991). At the same time, very high disagreement between reviewers is undesirable. If reviewer scores are hardly related, the legitimacy and validity of the peer review process are in danger, and the efficiency of the panel-phase in project proposal evaluations is at stake. Low IRR potentially undermines the ability to achieve agreement during the consensus phase, it threatens the procedure's objectivity, and impacts the reviewer's reflection of the funding agency's priorities (Tan et al. 2016; Derrick and Samuel 2017).

Therefore, disagreement may be due to different legitimate scientific perspectives on a proposal, but it is also important to curb unwarranted sources of low IRR. For example, if reviewers put very little effort into an evaluation, their scores will be somewhat random, and IRR will be low. Another important factor that may reduce IRR is the reviewers' lack of experience in evaluating a specific scheme. In a similar case, reviewers may have a vague idea, for example, of the evaluation criteria and the standard of quality in that specific evaluation context. The next paragraph explores why evaluation experience may improve IRR.

2.2 Reviewer's experience in evaluation and reliability

While striving for objectivity, science is characterized by different paradigms, perspectives, and complexities, which makes it difficult to agree on the value of a scientific claim (Kuhn 1962). Judging a scientific proposal is further complicated by the fact that it concerns research that will be developed and will produce results in the future, and reviewers must rely on the limited information available in a research proposal to predict its future success (Hemlin 2009). It is therefore not surprising that reviewers often disagree when scoring the same proposal.

To cope with the challenge of identifying the best proposals and to reduce the role of chance, agencies rely on expert reviewers. In fact, a prerequisite of a valid evaluation is in fact that the reviewer masters the content of the proposal, and that they are familiar with the research proposed, the theory, and the methods discussed. At the same time, proposals often entail a wide spectrum of expertise and a single reviewer may not be equipped to assess every aspect (Laudel 2006). Therefore, agencies typically rely on several reviewers, under the assumption that they will provide complementary insights, leading to an overall improvement in the ability to identify the best proposals.

The evaluation of a research proposal, however, is not the mere sum of reviewers' individual assessments based on their own scientific expertise, tastes, and heuristics. Through the experience gained from repeated evaluations of a specific funding scheme, the reviewer gains knowledge of three points of reference when judging and scoring a proposal, namely: (1) the objectives and evaluation criteria of a specific funding scheme; (2) the quality of proposals in that specific context; (3) the judgements and scores given by the other reviewers on similar past proposals.

First, while there are similarities across different funding schemes regarding the format of proposals and the evaluation criteria, each funding scheme also has specific objectives, and the funding agencies tailor their evaluation procedures accordingly. Through repeated evaluations, a reviewer becomes increasingly familiar with such idiosyncratic elements. In fact, higher IRR occurs when the evaluation priorities of the funding agency and those of the reviewers are aligned (Abdoul et al. 2012).

Secondly, consistency is the prime principle of human action, meaning that if two cognitive structures (e.g. evaluations) are logically inconsistent, arousal is increased due to cognitive dissonance, which activates processes aimed to increase consistency and reduce conflict in behaviour (Gawronski 2012). Therefore, through repeated evaluations, the reviewer can use the proposals already evaluated as a benchmark, in order to provide consistent scores across proposals. Consider, for example, a reviewer that evaluates a proposal as good and assigns a score of 8/10; then, she evaluates a second proposal also as good—yet slightly better than the first, so she can refine her judgement and provide a slightly higher score.

Finally, individual evaluations are typically followed by a discussion between reviewers to reach a final score. As shown by Steiner Davis et al. (2020), during this phase, reviewers typically observe other reviewers to determine what they find important in a proposal, and how they structure their assessment. There are also opportunities to interact with more experienced panellists and learn from this discourse.

In summary, through repeated evaluations, a reviewer embarks into socialization and learning process that is expected to gradually align the assessment of proposals to the specific evaluation context. As a result, evaluation scores will not only be affected by personal conceptions of value, but also by the objectives and criteria of the specific funding scheme, the quality of the proposals typically submitted to a funding scheme, and the way other reviewers in that context typically judge and score proposals. In turn, this process is expected to mould a reviewer's evaluation and scoring towards the correct way of reviewing proposals in each specific evaluation context. If we assume that the mean of individual scores approximates the 'correct' score in that context, then reviewers that have evaluated a greater number of a scheme's proposals will tend to provide scores closer to the mean of individual scores.

> Hypothesis 1: A reviewer's number of proposals evaluated in past calls of a research funding scheme decreases the disagreement with the mean of a proposal's the individual scores.

In the time between two calls, a reviewer's memory and experience may partially fade. Therefore, reliability may not only be affected by the number of proposals already evaluated in previous scheme calls, but also by the number of proposals evaluated in an ongoing round of evaluation.

However, increasing the number of proposals evaluated in a call arguably unleashes two counteracting effects. On the one hand, making repeated judgements or decisions depletes individuals' executive function and mental resources (Muraven and Baumeister 2000; Pocheptsova et al. 2009), and cognitive fatigue leads to decreased attention and poor information processing (van der Linden et al 2003 ; Boksem et al 2005), which is detrimental to individuals' judgements and decisions, even those of experts, like judges (Danziger et al. 2011) and physicians (Linder et al. 2014). Therefore, increasing the number of proposals evaluated in a call may gradually reduce the effort and time that a reviewer can

dedicate to each evaluation. This will arguably reduce accuracy and increase the gap between an appropriate score in that context, thus increasing divergence from the mean of the individual scores. On the other hand, comparison is a key leverage of learning (Alfieri et al. 2013; Patterson and Kurtz 2020), and by evaluating more proposals, a reviewer can make more comparisons between them and provide more accurate evaluations. This effect is arguably exponential, as judging one proposal enables zero comparisons, judging two proposals allows one comparison, judging three proposals allows three comparisons, judging four proposals allows six comparisons, judging five proposals allows 10 comparisons, and so forth.

In turn, we expect that:

> Hypothesis 2: A reviewer's number of proposals evaluated in the current call of a specific scheme has a linear positive effect (increase) and a negative quadratic effect (decrease) on the disagreement with the mean of a proposal's individual scores.

3. Data

3.1 MSCA and COST programmes

The Horizon 2020 MSCA is one of the European Union research flagship programmes promoting researchers' mobility. MSCA comprehends of four research funding schemes: doctoral training research networks (ITN); individual postdoctoral fellowships (IF); international and intersectoral staff exchanges between cooperating organizations, aimed to turn creative ideas into innovative products, services or processes (RISE); and co-funding of doctoral and postdoctoral programmes at the regional, national, and international level (COFUND).

The COST is the oldest pan-European framework for funding in science and technology. It was created in 1971 with the idea to connect national research systems through transnational networks, called 'COST Actions'. The proposals are developed by a group of researchers, and they can combine many disciplines.

There are some similarities and differences regarding the evaluation process of MSCA and COST proposals.

In MSCA, for all but the COFUND scheme, applicants select one of the eight scientific panels that best match their proposal's field of expertise. Moreover, to facilitate the matching of proposal and reviewer expertise, applicants and reviewers were asked to select a set of 'descriptors'[2] from a discipline classification system structured in panels, subpanels, and descriptors. In COST, the proposals are pooled in a common pot, and there are no panels. The applicants select up to five research areas of expertise that are relevant for the proposal (equivalent to the 'descriptors') from a discipline classification system structured into fields, subfields, and areas. Each reviewer identifies one area of core expertise and—optionally—other areas of high and medium expertise.

MSCA and COST reviewers are suggested by algorithms, which look for matches between the expertise of the reviewer and the expertise (descriptors) marked in the proposal. The final selection is performed by MSCA[3] and COST staffs who vet the reviewers for conflicts of interest, and they assure that expertise for all (or most) areas are covered.

Three reviewers (in rare cases four) evaluate each proposal individually and independently, resulting in three individual evaluation reports, and later through a consensus discussion among the reviewers, a consensus report is produced with a final score. Proposals are evaluated based on three criteria: (1) scientific excellence; (2) expected impact; and (3) proposed implementation. Each criterion is divided into several sub-criteria and is rated on a scale of 0 (fail) to 5 (excellent).[4] In IF, evaluations consider the proposal and the CV of the applicant, whereas the other schemes only focus on the proposal's content. MSCA evaluations are single-blind (i.e. reviewers know the identity of the authors, but not vice versa), whereas COST evaluations are double-blind (i.e. both reviewers and applicants' identities are unknown to each other).

In MSCA, each reviewer receives all the proposals at once and is given a timeframe for completing their assessment and proposal scoring. Therefore, the reviewer has some flexibility in deciding whether to read all the proposals at once, or sequentially. In COST, reviewers receive, evaluate, and score each proposal individually, rather than reviewing in a batch.

The final score is then converted on a scale from 0 to 100 in MSCA and 0 to 65 in COST.[5] The final step occurs when groups of proposals are reviewed by ad hoc review panels—which in rare instances can suggest changes in the consensus report and score. In COST, the proposals with the highest scores are automatically selected from the pool of proposals at large,[6] whereas in MSCA, the highest scores within each panel are selected.

3.2 Sample

The sample of MSCA data includes 52,230 proposals (43,250 for IF, 7,460 for ITN, and 1,520 for RISE) covering the first 5 years of Horizon 2020 (2014–2018). These projects received 129,838 evaluation scores in IF, 23,979 in ITN, and 4,560 in RISE. A group of 6,165 unique reviewers completed 158,377 evaluations for all three MSCA schemes. COFUND is not included in our study. Overall, 9,669 proposals were funded, although the success rates vary from 10.2% (ITN) to 19.3% (IF) and 35.9% (RISE). The sample for COST evaluations includes data from five calls, between March 2015 and September 2017, for a total of 1,939 proposals, and 5,821 evaluations conducted by 3,244 reviewers. Overall, 166 proposals have been funded (8.5%).

3.3 Variables

3.3.1 Dependent variable

Past studies examined reviewers' disagreement at the proposal level, using measures such as the standard deviation of the scores and the intraclass correlation coefficients (Mutz et al. 2012b). We employ a measure at the reviewer level, namely, the extent to which a reviewer's evaluation of a proposal deviates from the mean of the individual scores of the proposal. This measure explores how both proposal and reviewer traits affect disagreement. Formally, for each evaluation Score_{ij} of a proposal i, performed by a reviewer j, we calculate disagreement D_{ij} as:

$$D_{ij} = \left|\text{Score}_{ij} - \frac{\sum_{j=1}^{n} \text{Score}_{ij}}{n}\right|$$

where n stands for the number of reviewers j in a specific project i. As we take the absolute value of the differences, the resulting variable represents the difference from the mean evaluation scores, independent of the direction.

Table 1 includes descriptive statistics on reviewers' disagreement in three MSCA schemes, and in COST. Within MSCA, disagreement is stronger on average in RISE than IF and ITN. The large

disagreement in the RISE may be due to the intersectoral nature of the funding scheme, including academic and non-academic beneficiaries, implying more diversity in methods and approaches, and more difficulty in reaching agreement in reviewer scores.

Disagreement in COST is stronger than in MSCA: 7.46 points on a 0–65 scale, compared with 7.14–8.40 points on a 0–100 scale. One possible reason is that reviews in COST are double-blind, while MSCA is not. However, while one might expect single-blind reviewers to agree more than double-blind reviewers, for instance, because they would tend to share a preference for proposals by famous authors, Tomkins et al. (2017) found in a study of journal peer review that the agreement of single-blind reviewers is not significantly superior to the agreement of double-blind reviewers.

3.3.2 Independent variables

The level of *experience* with evaluation in a funding scheme—*meaning the number of proposals evaluated*—is our main predicting factor.

In the MSCA dataset, we identify four variables of experience, by distinguishing between the evaluation experience within the specific funding scheme of the evaluated proposal (e.g. IF) and the experience in other MSCA funding schemes (e.g. RISE plus ITN), and between experience in the *current call* from the sum of experiences in *previous calls*.

As the COST dataset only includes one research funding scheme, we only distinguish between experiences in the current call versus the sum of past experiences in previous calls.

Descriptive properties of the different experience variables are included in Table 1. In COST, experts review, on average, less than two proposals per call and have evaluated just one proposal in previous calls. Reviewer experience is much higher in MSCA schemes. Current within-scheme experience is higher in IF (on average 15.46) than in ITN (8.65) and RISE (8.79).[7] Previous calls within-scheme experience is much higher in IF (on average 12.31) than in ITN (6.40) and RISE (4.11).

3.3.3 Control variables

3.3.3.1 Reviewer characteristics. We consider several variables representing reviewer's characteristics that may affect evaluation scores and hence, indirectly, the level of disagreement with other reviewers.

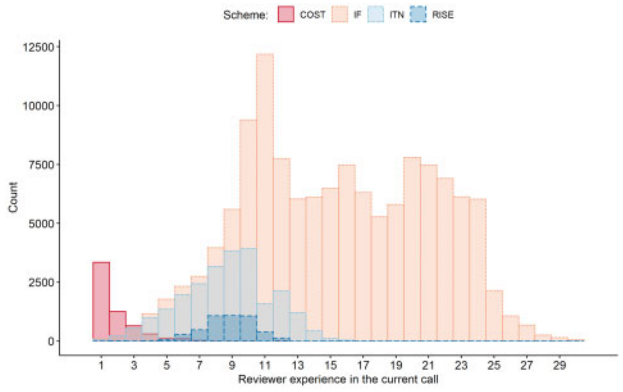

Figure 1. Distribution of reviewer experience (number of proposals evaluated) in the current call by funding scheme.

A reviewer's assessment can be affected by personal biases, such as the propensity to give high or low scores in general and/or to proposals or applicants with given traits. We consider two demographic traits, namely, gender and age that may be associated with some systematic or idiosyncratic biases. In the MSCA dataset, we also include a third indicator for *industrial experience* (yes or no), which is not available in the COST dataset.

In addition to the main effect of these individual characteristics, we also tested the effect of a proxy for similarity–difference of these traits with other proposal reviewers, and an interaction of this index with the main effect of the individual trait. This tests whether similarity in individual characteristics among the reviewers is associated with lower disagreement, and whether this similarity effect is different depending on the gender, industrial experience, and age of the reviewers.

Gender similarity is calculated for each proposal i and each reviewer j as the number of reviewers in proposal i with same gender as j, divided by the total number of reviewers for proposal i. Likewise, for industrial expertise similarity, we divided the number of reviewers in proposal i with the same industrial expertise as j by the total number of reviewers in proposal i. The result is a proportion, where 0 indicates no similarity at all, and 1 indicates perfect similarity. Age similarity for each proposal i and each reviewer j is given by the average age difference with other reviewers in proposal j.

Descriptive statistics of these variables are included in Table 1. The mean age of a reviewer is very similar in COST and the three different MSCA schemes, ranging between 47.60 and 49.40. The gender distribution shows a small overrepresentation of male reviewers in the MSCA schemes (between 53.31 and 55.87%), and a larger gender gap in COST, where male reviewers represent 63.55% of the total sample. Industrial expertise of reviewers is quite common among RISE reviewers (51.43%), less among ITN reviewers (47.03), and significantly less among IF reviewers (34.44%). These numbers probably reflect that within RISE- and to a lesser extent ITN, proposals are typically more industry-oriented than in IF.

3.3.3.2 Proposal characteristics. Discipline profile. A reviewer's scientific background impacts their understanding of what are considered to be the appropriate research objectives, contributions, theories, and methods of a proposal, meaning the extent to which reviewers share mutual understanding will affect the level of disagreement. As the Social Sciences are more fragmented than the Natural Sciences when it comes to understandings of appropriate goals, theories, and methods (Whitley 2000), such fragmentation may contribute to higher levels of disagreement for the evaluation of proposals in the Social Sciences (Mallard et al. 2009; Lamont and Huutoniemi 2011; Mutz et al. 2012b; Pina et al. 2015).

We construct the variables to measure the disciplinary profile of the proposals. For the MSCA scheme, we consider the panel to which the proposal is submitted.[8] For COST, each proposal classifies a maximum of five specialization areas from six broad disciplinary fields; the variable measures the *disciplinary focus* of the proposal as the share of areas in each of six disciplinary fields. Table 1 shows the distribution of the proposals among the different panels of MSCA, and the average disciplinary focus for each of the six fields in COST. These figures show that within MSCA, the proportion of proposals in each panel differs substantially. For example, the most popular panel within IF is Life Sciences, while the most popular in ITN and RISE is Information Sciences and

Table 1. Descriptive statistics

Variable	MSCA			COST
	IF	ITN	RISE	
Evaluation level	$N = 129,838$	23,979	4,560	5,821
Reviewers' disagreement		**Scale: 0–100**		**0–65**
Mean (standard deviation)	7.41 (6.31)	7.13 (5.99)	8.70 (7.01)	7.46 (5.83)
Median (Q1–Q3)	5.80 (2.67–10.40)	5.67 (2.67–10.00)	7.00 (3.33–12.47)	6.00 (3.00–10.67)
Min–Max	0.00–61.00	0.00–56.20	0.00–49.93	0.00–36.00
Scheme experience: PREVIOUS CALLS				
Mean (SD)	12.31 (17.16)	6.40 (9.19)	4.11 (6.98)	0.99 (2.51)
Median (Q1–Q3)	0 (0–21)	0 (0–10)	0 (0–9)	0.00 (0–1)
Min–Max	0–87	0–46	0–37	0–27
Scheme experience: CURRENT CALL				
Mean (standard deviation)	15.46 (5.61)	8.65 (2.69)	8.79 (1.49)	1.88 (1.37)
Median (Q1–Q3)	15 (11–20)	9 (7–10)	9 (8–10)	1 (1–3)
Min–Max	1–30	1–16	2–12	1–12
Other schemes experience: CURRENT CALL				
Mean (SD)	1.21 (3.27)	5.44 (7.78)	8.10 (9.67)	–
Median (Q1–Q3)	0 (0–0)	0 (0–11)	0 (0–15)	–
Min–Max	0–24	0–36	0–39	–
Other schemes experience: PREVIOUS CALLS				
Mean (SD)	0.92 (4.23)	6.28 (14.25)	8.24 (16.29)	–
Median (Q1–Q3)	0 (0–0)	0 (0–0)	0 (0–12)	–
Min—Max	0–59	0–89	0–110	–
Gender overlap				
Mean (SD)	0.53 (0.36)	0.54 (0.35)	0.53 (0.36)	0.57 (0.40)
Median (Q1–Q3)	0.50 (0.50–1)	0.50 (0.33–1)	0.50 (0.5–1)	0.50 (0.50–1)
Missing (%)	629 (0.48)	487 (2.03)	576 (12.63)	880 (15.12)
Industrial expertise overlap				
Mean (SD)	0.59 (0.38)	0.53 (0.35)	0.54 (0.37)	–
Median (Q1–Q3)	0.50 (0.5–1)	0.50 (0.33–1)	0.50 (0.50–1)	–
Missing (%)	629 (0.48)	487 (2.03)	576 (12.63)	–
Proposal level	$N = 43,250$	7,460	1,520	1,939
Panel				
Chemistry (%)	5,289 (12.23)	866 (11.61)	142 (9.34)	–
Economics (%)	1,059 (2.45)	88 (1.18)	72 (4.74)	–
Information Sciences and Engineering (%)	5,199 (12.02)	2,279 (30.55)	462 (30.39)	–
Environment and Geosciences (%)	5,598 (12.94)	903 (12.10)	204 (13.42)	–
Life Sciences (%)	1,1401 (26.36)	2,015 (27.01)	231 (15.20)	–
Mathematics (%)	981 (2.27)	102 (1.37)	47 (3.09)	–
Physics (%)	4,671 (10.80)	545 (7.31)	151 (9.93)	–
Social Sciences and Humanities (%)	9,052 (20.93)	662 (8.87)	211 (13.88)	–
Field focus				
Natural Sciences (mean–SD)	–	–	–	0.25 (0.36)
Engineering and Technology (mean–SD)	–	–	–	0.22 (0.34)
Medical and Health Sciences (mean–SD)	–	–	–	0.19 (0.34)
Agricultural Sciences (mean–SD)	–	–	–	0.06 (0.19)
Social Sciences (mean–SD)	–	–	–	0.24 (0.37)
Humanities (mean–SD)	–	–	–	0.05 (0.18)
Interdisciplinarity (HI)				
Mean (SD)	−0.86 (0.22)	−0.81 (0.24)	−0.80 (0.25)	−0.80 (0.25)
Median (Q1–Q3)	−1 (−1 to −0.63)	−1 (−1 to 0.56)	−1 (−1 to 0.56)	−1 (−1 to −0.56)
Missing (%)	8,841 (20.40)	1,711 (22.90)	324 (21.30)	41 (2.11)
Reviewer level	$N = 5,193$	1,851	361	3,244
Age				
Mean (SD)	47.60 (9.00)	49.10 (9.22)	49.20 (8.88)	49.40 (9.00)
Median (Q1–Q3)	46 (41–54)	48 (42–54)	48 (42–55)	48 (42–55)
Min–Max	26–83	27–83	29–74	22–80
Missing (%)	30 (0.58)	67 (3.62)	46 (12.70)	344 (10.60)
Gender				
Female (%)	2,306 (44.66)	833 (46.69)	139 (44.13)	1,057 (36.45)
Male (%)	2,857 (55.34)	951 (53.31)	176 (55.87)	1,843 (63.55)
Missing (%)	30 (0.58)	67 (3.62)	46 (12.74)	344 (10.60)

(continued)

Table 1. Continued

Variable	MSCA			COST
	IF	ITN	RISE	
Industrial expertise				
No (%)	3,385 (65.56)	945 (52.97)	153 (48.57)	–
Yes (%)	1,778 (34.44)	839 (47.03)	162 (51.43)	–
Missing (%)	30 (0.58)	67 (3.62)	46 (12.74)	–

Engineering. In COST, Natural Sciences and Social Sciences are the two main fields of proposal focus.

Interdisciplinarity of the proposal. Proposals with a higher degree of interdisciplinarity may be conducive to stronger disagreement for three main reasons. First, when a proposal spans a broad or unusual set of expertise, reviewers are more likely to possess different scientific backgrounds, and hence to disagree on its value (Laudel 2006). Secondly, when a reviewer lacks (some of) the scientific expertise needed to make a valid assessment, the score will largely be due to chance. This may happen more frequently for interdisciplinary proposals, where reviewers may lack some of the expertise needed to make a valid assessment (Porter and Rossini 1985; Bruun et al. 2005; Porter et al. 2012). A third reason is that proposals with a high degree of interdisciplinarity are intrinsically more difficult to explain and justify (Lee 2006; Mansilla et al. 2006; Uzzi et al. 2013), which increase randomness and likely disagreement.

Interdisciplinarity is a complex and multidimensional concept (Leydesdorff and Rafols 2011; Wagner et al. 2011). We use a measurement that conceptualizes interdisciplinarity as integration, namely research which combines and builds on *diverse* disciplines[9] (Rafols et al. 2012). Diversity entails three main features (Stirling 2007): (1) variety, that is, the number of disciplines; (2) balance, that is, their relative proportion; and (3) disparity, that is, the extent to which they differ, their cognitive distance.[10] Given that the subpanels in MSCA or the areas in COST of a proposal are not weighted,[11] we compute the Herfindahl index (HI) at panel level for MSCA and at the field level for COST. The measure of balance is given by the number of subpanels within a panel, or the number of areas within a specific field, respectively.

$$\mathrm{HI}_{\mathrm{interdisciplinarity}} = \sum_{i,j}(p_i p_j)*(-1)$$

Where i and j are the two distinct panels/fields and p_i is the proportion of research subpanels/areas assigned to each value of i. We multiplied the HI with -1 to facilitate a more logic interpretation, namely, a low value indicates low interdisciplinarity and vice versa.

3.4 Method

To explore how reviewing experience affects reviewers' disagreement, we construct a multilevel cross-classified linear regression model. This model accounts for the fact that individual project evaluation scores and derivative disagreement scores are nested simultaneously in reviewers and in proposals. It controls for the fact that the relationship between *experience* and *disagreement* may be partly due to an unobserved variable affecting both. As reviewers can assess more than one project, the nesting is not hierarchical but cross-classified. A chi-square test on the difference in -2 loglikelihood of the three-level null model with a one-level null model points to a significant difference for all four models, and hence confirms the need for a multilevel model.

Missing values were deleted listwise in all models. Continuous variables used in interaction effects (i.e. gender overlap and industrial expertise overlap) are mean centred to facilitate a better understanding of the effect.

4. Results

Tables 2 and 3 present the results of cross-classified multilevel regressions, respectively, for MSCA and COST.

Concerning reviewing experience, results for IF, ITN, and RISE show that the number of proposals evaluated in previous calls of one specific scheme predicts significantly lower disagreement, supporting Hypothesis 1. The effect is not significant in COST, arguably because the level of experience accumulated is often too small to produce a visible effect (mean value of less than one).

In the IF scheme, in line with Hypothesis 2, we observe that a reviewer's number of proposals evaluated in the current call has a linear positive effect (increase) and a negative quadratic effect (decrease) on the disagreement with the mean of individual proposal scores. This is in line with the argument that increasing the number of proposals has a double effect. On the one hand, it reduces evaluation accuracy by reducing the time and effort to evaluate each proposal; on the other hand, it improves accuracy by exponentially increasing the potential to compare proposals.

These combined effects lead to an inverted U relationship with reviewers' disagreement (Figure 2). The predicted values, shown in Figure 2, suggest that reviewers assessing only one proposal can expend significant effort and time on the one proposal, leading to accurate evaluations and hence moderate disagreement from the mean individual evaluations. Up to approximately seven proposals evaluated, the (linear) effect of reduced accuracy due to less time and effort for each evaluation is stronger than the effect of increased comparisons, leading to greater disagreement. The turning point of the curve is situated at eight proposals: from there onwards, the marginal effect of an additional proposal on the comparison effect becomes stronger than the marginal effect of less effort and time. Above 13 proposals, the overall comparison effect overcomes the effort–time effect, leading to considerably lower levels of disagreement. Concretely, Figure 2 shows that by increasing the number of proposals evaluated in a given call can decrease disagreement substantially. For example, from 18 to 28 proposals evaluated, disagreement decreases by around 15%.

In ITN and RISE, there are similar effects, but not significant, while in COST they have different and non-significant results. This may be due to the fact that the total number of proposals evaluated in IF (43,014) is much larger than ITN (7,331), RISE (1,333) and COST (1,688). Moreover, the effect of experience in reducing disagreement becomes strong enough only when the number of proposals evaluated by each reviewer in a given call becomes large enough; in IF, for example, above 13 proposals. However, while in

Table 2. Cross-classified multilevel regressions for MSCA

Variable	IF				ITN				RISE			
	Estimate	Std. error	P-value	Std. coef.	Estimate	Std. error	P-value	Std. Coef.	Estimate	Std. error	P-value	Std. coef.
(Intercept)	6.479	0.213	**0.000**	0.000	6.506	0.453	**0.000**	0.000	2.857	3.116	0.359	0.000
Reviewers' experience												
Scheme experience: PREVIOUS CALLS	−0.017	0.001	**0.000**	−0.047	−0.041	0.005	**0.000**	−0.063	−0.047	0.019	**0.015**	−0.044
Scheme experience: CURRENT CALL	0.059	0.019	**0.002**	0.053	0.111	0.078	0.153	0.050	0.672	0.695	0.334	0.139
Squared-Scheme experience: CURRENT CALL	−0.004	0.001	**0.000**	−0.105	−0.009	0.005	0.057	−0.069	−0.024	0.040	0.540	−0.088
Other schemes experience: CURRENT CALL	−0.014	0.007	**0.047**	−0.007	−0.009	0.006	0.170	−0.011	−0.017	0.015	0.253	−0.023
Other schemes experience: PREVIOUS CALLS	−0.004	0.005	0.457	−0.003	−0.002	0.004	0.588	−0.005	−0.006	0.009	0.517	−0.014
Proposal panel (ref. cat is Chemistry)												
Economics	1.821	0.185	**0.000**	0.045	2.093	0.526	**0.000**	0.038	2.721	0.838	**0.001**	0.084
Information Sciences and Engineering	1.059	0.116	**0.000**	0.055	0.890	0.213	**0.000**	0.069	2.044	0.588	**0.001**	0.133
Environment and Geosciences	0.665	0.114	**0.000**	0.035	0.248	0.252	0.323	0.014	1.508	0.661	**0.022**	0.073
Life Sciences	0.418	0.101	**0.000**	0.029	0.740	0.215	**0.001**	0.055	1.949	0.630	**0.002**	0.099
Mathematics	0.751	0.201	**0.000**	0.018	1.290	0.524	**0.014**	0.025	2.325	1.024	**0.023**	0.054
Physics	−0.293	0.123	**0.018**	−0.014	−0.012	0.296	0.967	−0.001	1.122	0.725	0.121	0.048
Social Sciences and Humanities	2.190	0.106	**0.000**	0.141	1.693	0.274	**0.000**	0.081	2.191	0.654	**0.001**	0.107
Reviewer characteristics												
Age	0.012	0.003	**0.000**	0.017	0.001	0.006	0.875	0.001	0.018	0.016	0.253	0.022
Gender: Male	−0.292	0.056	**0.000**	−0.023	−0.089	0.114	0.435	−0.007	−0.695	0.288	**0.016**	−0.049
Gender overlap (centred)	0.193	0.082	**0.019**	0.011	0.248	0.190	0.190	0.015	0.626	0.520	0.228	0.032
Male gender * gender overlap	−0.427	0.115	**0.000**	−0.018	−0.380	0.265	0.152	−0.017	−0.921	0.729	0.206	−0.035
Industrial expertise: Yes	0.305	0.064	**0.000**	0.023	0.103	0.115	0.371	0.009	0.116	0.296	0.696	0.008
Industrial expertise overlap (centred)	−0.270	0.074	**0.000**	−0.016	0.001	0.187	0.994	0.000	−0.606	0.518	0.242	−0.032
Industrial expertise * industrial expertise overlap	0.509	0.123	**0.000**	0.019	0.137	0.276	0.621	0.006	0.820	0.746	0.272	0.031
Random effects												
σ^2		24.322				23.295				32.651		
τ00 reviewer level		1.747				2.126				1.889		
τ00 proposal level		12.543				9.854				14.417		
Marginal R^2/conditional R^2		0.023/0.385				0.012/0.348				0.019/0.346		
No. of evaluations		129,209				23,492				3,984		
No. of proposals		43,014				7,331				1,333		
No. of reviewer		5,192				1,847				361		

Significance of bold is $P < 0.05$.

IF the average number of proposals in the current call per each reviewer is 15.46, it is much lower in ITN (8.65), RISE (8.79), and especially in COST (1.88). In turn, the combined effect of total number of proposals evaluated and the number of evaluations per reviewer in each call, contributed to explaining why the effect is remarkable and significant in MSCA IF, and not significant in ITN, RISE, and COST. For COST, the difference in the signs of the coefficients can be explained by the fact that reviewers receive, evaluate, and score each proposal individually, rather than reviewing in a batch, as well as substantially lower number of proposals evaluated in the current call per each reviewer (1.88). In fact, conducting each evaluation one by one reduces the room for learning and adjusting scoring through comparison, and with such a low number of evaluated proposals per reviewer, reviewer's fatigue is very unlikely to occur, and personal limits of time and effort are unlikely to be reached.

Concerning our control variables, we observe strong disciplinary differences in all schemes. In MSCA, disagreement is much stronger in Social Sciences and Humanities compared with the reference category of Chemistry, and this holds for all three MSCA schemes. In COST, disagreement is much higher in the reference category of Humanities than in all other disciplines.

Some characteristics of the reviewers predict the level of disagreement.

The main effect of male gender is significantly negative in IF and COST, meaning it predicts less disagreement. In addition, in IF, the effect of gender overlap is positive, while the interaction effect is significantly negative for male gender. This means that for female reviewers (Figure 2), having same-gender reviewers on a proposal will lead to more disagreement, while for male reviewers, same-gender reviewers on the same proposal will lead to less disagreement. These surprising effects are possibly explained by the fact that female reviewers tend to use a wider range of evaluation scores: the standard deviation of individual scores in IF is 14.8 for male reviewers and 15.5 for female reviewers (see Supplementary Appendix).[12]

In IF, reviewers generally display higher disagreement when they have industrial expertise, while the industrial expertise overlap leads to less disagreement. At the same time, the interaction between

Table 3. Cross-classified multilevel regressions for COST

Variable		COST			
		Estimate	Std. error	P-value	Std. coef.
(Intercept)		7.967	0.787	**0.000**	0.000
Reviewers' experience					
	Scheme experience: PREVIOUS CALLS	−0.016	0.059	0.791	−0.004
	Scheme experience: CURRENT CALL	−0.223	0.284	0.431	−0.044
	SQUARED: Scheme experience: CURRENT CALL	0.068	0.050	0.171	0.078
Proposal field focus (ref. cat is Humanities)					
	Natural Sciences	−2.603	0.641	**0.000**	−0.160
	Engineering and Technology	−1.416	0.646	**0.028**	−0.083
	Medical and Health Sciences	−1.323	0.641	**0.039**	−0.079
	Agricultural Sciences	−1.535	0.795	0.053	−0.051
	Social Sciences	−1.140	0.653	0.081	−0.070
Reviewer characteristics					
	Age	0.031	0.010	**0.002**	0.047
	Gender: Male	−0.566	0.214	**0.008**	−0.047
	Gender overlap (centred)	0.252	0.400	0.528	0.017
	Male gender * gender overlap	−0.125	0.515	0.809	−0.006
Random effects					
	σ^2		22.224		
	τ00 reviewer level		2.120		
	τ00 proposal level		9.600		
Marginal R^2/conditional R^2			0.017/0.357		
No. of evaluations			4,428		
No. of proposals			1,688		
No. of reviewers			2,538		

Significance of bold is $P < 0.05$.

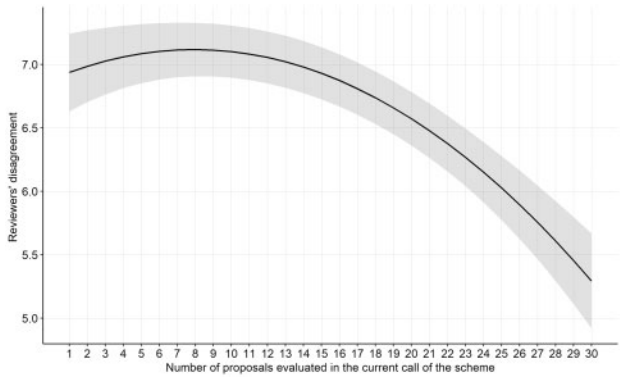

Figure 2. Predicted values of reviewers' disagreement by scheme experience in current call for MSCA IF.

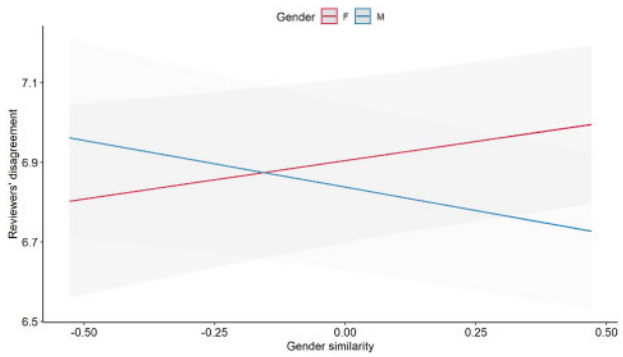

Figure 3. Predicted values of reviewers' disagreement by gender similarity and gender for MSCA IF.

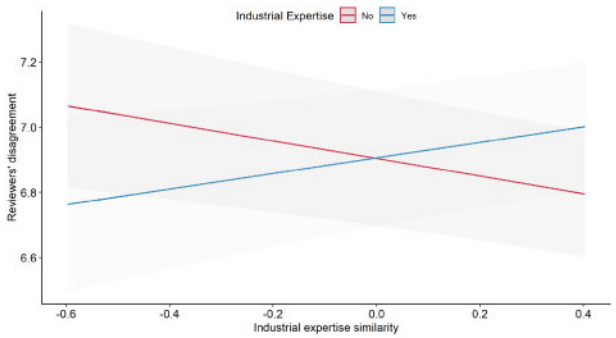

Figure 4. Predicted values of disagreement by industrial expertise similarity and industrial expertise for MSCA IF.

having industrial expertise and the degree to which industrial expertise overlaps is positive and significant. This means that when reviewers with industrial expertise are paired with co-reviewers who have industrial expertise there is high disagreement, while a reviewer with no industrial expertise coupled with co-reviewers without industrial expertise relates to lower disagreement (Figure 3). In other words, there is more variation in the way reviewers with industrial experience assess a proposal than the reviewers without such experience (see Supplementary Appendix). This result may be due to the research focus of MSCA schemes, whereas in a hypothetical research funding scheme focused on innovation reviewers with industrial expertise may actually display less disagreement.

Older reviewers display stronger disagreement. Looking for an explanation, we tested two different hypotheses. First, that older reviewers give more extreme individual scores; this was not the case (see Supplementary Appendix). Secondly, that the effect of

experience is more pronounced for younger than older reviewers (viz, younger learning faster). This could not be confirmed as the interaction of age and accumulated experience is not significant.[13]

Finally, one of the schemes showed a significant effect of proposal interdisciplinarity on disagreement. As this variable presents around one in four missing cases (Table 1), hence reducing the sample size significantly, we report the results without this variable.

5. Conclusions

We explored proposals' and reviewers' characteristics affecting reliability in research proposal evaluations. We elaborated hypotheses on how disagreement with the other reviewers is affected by the research scheme reviewing experience, namely the number of proposals from a grant scheme evaluated: (1) in previous calls and (2) in the current call. In order to gain more robust insights, we tested our hypotheses by studying two of the most important research funding programmes in the European Union from 2014 to 2018, namely, 52,488 proposals evaluated under three funding schemes of the Horizon 2020 MSCA, and 1,939 proposals evaluated under the COST Actions.

The empirical analysis shows that past scheme experience reduces disagreement in all MSCA schemes. The level of experience accumulated in COST is arguably too small (mean of 0.99) to produce a visible effect. In the IF scheme, we observed an inverted U relationship with experience in the current call: disagreement increases from one to seven proposals evaluated, and then decreases substantially. Older reviewers, those with industrial expertise, and women, display a higher level of disagreement. For the latter two categories, a possible explanation is that they are keener to provide very low and very high scores. For industrial expertise, also the specific focus of MSCA schemes on fundamental and basic research, rather than innovation, may account for the greater disagreement.

Disagreement is higher in COST than in MSCA. Within MSCA, disagreement is stronger in RISE compared with IF and ITN, possibly because their cross-sectoral focus implies a content more difficult to evaluate and putting together reviewers with a more diverse background. As in previous studies, we found that disagreement is higher for proposals in the Social Sciences and Humanities, while this is the first study to explore whether the value of interdisciplinary proposals is more contested, and all tests conducted with several different measures show that this is not the case.

Some choices and limitations should be discussed. Reliability may be regarded to be of minor importance when compared with validity. However, disagreement among reviewers typically leads to lower consensus scores, in turn harming validity as well (Pina et al. 2015). Moreover, while some level of disagreement is unavoidable and even desirable, the common very high level of disagreement found in proposal evaluation threatens the legitimacy of the peer review process and its capability to identify the best submissions (Tan et al. 2016; Derrick and Samuel 2017; Roumbanis 2019). Our results show that scheme reviewing experience decreases disagreement; at the same time, exactly how many proposals are advisable to evaluate may vary across funding schemes, as well as reviewers, for example, in relation to the complexity of the proposal, existing incentives, etc. Future research may shed light on similar additional factors. While findings are consistent with the initial hypotheses, observational and in-depth research are arguably needed to further deepen the underlying mechanisms and explanations, for example, about the reason why some reviewers are keener to give extreme scores, or the reasons for the greater disagreement displayed by older reviewers.

The findings have implications on the theoretical understanding of evaluation processes and reviewers' behaviour. We argued that reviewers implicitly use different evaluation approaches, which represent an unwarranted source of disagreement. At the same time, our findings show that, through repeated evaluations, reviewers embark in a learning process that appears to mould their evaluation approach in an appropriate way of scoring proposals in that specific context. This process leads to a substantial improvement in reliability.

The results provide support for studies, which have argued that a major factor explaining low IRR in grant evaluation are reviewers who only score one or a few submissions (Jayasinghe et al. 2006; Marsh et al. 2008). A practical implication for the organization of the evaluation processes and to exploit the benefits of experience is to guarantee that reviewers judge enough proposals of a funding scheme and to establish long-term relationships with reviewers. At the same time, if agencies aim to increase the number of reviews per evaluator, they could monitor that the diversity of their scientific and socio-demographic background is preserved. It is also important to keep the proposal requirement as simple as possible, and to carefully ponder the trade-off between the amount of information in a proposal and the reviewers' processing capacity. Moreover, a key component of training evaluators should include looking at how past proposals have been evaluated in that specific context, and when possible, agencies could try to assemble a pool of reviewers with diverse traits like gender, age, and experience.

Supplementary data

Supplementary data are available at *Research Evaluation Journal* online.

Acknowledgements

We are grateful to Michael Wise for carefully proofreading the article. We thank COST Association and the Research Executive Agency for making available the data for this analysis.

Disclaimer

All views expressed in this article are strictly those of the authors and may in no circumstances be regarded as an official position of the Research Executive Agency, COST, or the European Commission.

Conflict of interest statement. None declared.

Notes

1. For example, disagreement between reviewers systematically lowers the final score of an application (Pina et al. 2015; Tamblyn et al. 2018).
2. A minimum one in 2014–6, and a minimum of three in 2017–8—with RISE already requesting a minimum of three descriptors in 2016.
3. Sometimes supported by so-called vice-chairs, who serve as external experts that assist MSCA staff in choosing the final reviewers.
4. Excellence, impact, and implementation weights on the final score are respectively 50, 30, and 20% in MSCA, and 38, 31, and 31% in COST.
5. As the data were gathered, a minor reform has changed the COST grading system; the final score now goes from 0 to 50.

6. In case of ties, the proposals for COST Actions are selected by the COST Scientific Committee on policy-relevant criteria.
7. IF experts tend to receive more proposals because these are normally simpler and shorter than ITN and RISE proposals, which involve several partners.
8. MSCA evaluation is organized into eight scientific panels: Chemistry (CHE), Physics (PHY), Social Sciences and Humanities (SOC), Economics (ECO), Mathematics (MAT), Life Sciences (LIF), Environment and Geosciences (ENV), and Information Sciences and Engineering (ENG).
9. We also tested other conceptualizations of interdisciplinarity: Leydesdorff (2018, 2019) measurement of diversity, and Bromham's (2016) measurement based on phylogenetic species evenness. However, none of these alternative operationalizations gave additional insights or different results.
10. Along diversity, Rafols et al. (2012) also consider coherence as a dimension of interdisciplinarity. However, the data available do not allow this dimension to be computed.
11. Computing the index at area level would mean that the balance will not affect the measure, as it will always equal 1: $(1/n)(1/n)*n^2$.
12. The multivariate model clearly controls for the slight under representation of female reviewers, as confidence intervals (and thus standard errors) will be wider (higher) for categories with less observations.
13. The main effect of age similarity and the interaction effect of age with age similarity are not significant and are therefore not reported in Tables 2 and 3.

References

Abdoul, H., Perrey, C., Amiel, P., Tubach, F., Gottot, S., Durand-Zaleski, I., and Alberti, C. (2012) 'Peer Review of Grant Applications: Criteria Used and Qualitative Study of Reviewer Practices', *PLoS One*, 7: e46054.

Alfieri, L., Nokes-Malach, T. J., and Schunn, C. D. (2013) 'Learning through Case Comparisons: A Meta-Analytic Review', *Educational Psychologist*, 48: 87–113.

Bailar, J. C. (1991) 'Reliability, Fairness, Objectivity and Other Inappropriate Goals in Peer Review', *Behavioral and Brain Sciences*, 14: 137–8.

Boksem, M. A., Meijman, T. F., and Lorist, M. M. (2005) 'Effects of Mental Fatigue on Attention: An ERP Study', *Cognitive Brain Research*, 25: 107–16.

Bol, T., de Vaan, M., and van de Rijt, A. (2018) 'The Matthew Effect in Science Funding', *Proceedings of the National Academy of Sciences*, 115: 4887–90.

Bornmann, L., and Daniel, H.-D. (2010) 'Reliability of Reviewers' Ratings When Using Public Peer Review: A Case Study', *Learned Publishing*, 23: 124–31.

Bornmann, L., Mutz, R., and Daniel, H.-D. (2010) 'A Reliability-Generalization Study of Journal Peer Reviews: A Multilevel Meta-Analysis of Inter-Rater Reliability and Its Determinants', *PloS One*, 5: e14331.

Borsuk, R. M. et al. (2009) 'To Name or Not to Name: The Effect of Changing Author Gender on Peer Review', *Bioscience*, 59: 985–9.

Boudreau, K. et al. (2012) 'The Novelty Paradox & Bias for Normal Science: Evidence from Randomized Medical Grant Proposal Evaluations', *Harvard Business School Working Paper Series# 13-053*, Harvard Business School.

Bromham, L., Dinnage, R., and Hua, X. (2016) 'Interdisciplinary Research Has Consistently Lower Funding Success', *Nature*, 534: 684–7.

Bruun, H. et al. (2005). *Promoting Interdisciplinary Research: The Case of the Academy of Finland*. Helsinki: Academy of Finland.

Cicchetti, D. V. (1991) 'The Reliability of Peer Review for Manuscript and Grant Submissions: A Cross-Disciplinary Investigation', *Behavioral and Brain Sciences*, 14: 119–35.

Cole, S., and Simon, G. A. (1981) 'Chance and Consensus in Peer Review', *Science*, 214: 881–6.

Danziger, S., Levav, J., and Avnaim-Pesso, L. (2011) 'Extraneous Factors in Judicial Decisions', *Proceedings of the National Academy of Sciences*, 108: 6889–92.

Derrick, G., and Samuel, G. (2017) 'The Future of Societal Impact Assessment Using Peer Review: Pre-Evaluation Training, Consensus Building and Inter-Reviewer Reliability', *Palgrave Communications*, 3: 1–10.

DFG. (2016). *Crossing Borders - Interdisciplinary Reviews and Their Effects*. Deutsche Forschungsgemeinschaft. <https://www.dfg.de/download/pdf/dfg_im_profil/geschaeftsstelle/publikationen/interdisciplinary_review_processes.pdf> 22 Sep 2020.

Gawronski, B. (2012) 'Back to the Future of Dissonance Theory: Cognitive Consistency as a Core Motive', *Social Cognition*, 30: 652–68.

Hargens, L. L. (1988) 'Scholarly Consensus and Journal Rejection Rates', *American Sociological Review*, 53: 139–51.

Hargens, L. L. (1990) 'Variation in Journal Peer Review Systems: Possible Causes and Consequences', *JAMA*, 263: 1348–52.

Hemlin, S. (2009) 'Peer Review Agreement or Peer Review Disagreement: Which is Better', *Journal of Psychology of Science and Technology*, 2: 5–12.

Hodgson, C. (1997) 'How Reliable is Peer Review? An Examination of Operating Grant Proposals Simultaneously Submitted to Two Similar Peer Review Systems', *Journal of Clinical Epidemiology*, 50: 1189–95.

Jayasinghe, U. W., Marsh, H. W., and Bond, N. (2003) 'A Multilevel Cross-Classified Modelling Approach to Peer Review of Grant Proposals: The Effects of Assessor and Researcher Attributes on Assessor Ratings', *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 166: 279–300.

Jayasinghe, U. W., Marsh, H. W., and Bond, N. (2006) 'A New Reader Trial Approach to Peer Review in Funding Research Grants: An Australian Experiment', *Scientometrics*, 69: 591–606.

Kuhn, T. S. (1962) *The Structure of Scientific Revolutions*, original edn. Chicago: University of Chicago Press.

Lamont, M., and Huutoniemi, K. (2011) '*Opening the Black Box of Evaluation: How Quality is Recognized by Peer Review Panels*'. Bern (Switzerland): SWAG Bulletin.

Lane, J. A., and Linden, D. J. (2009) 'Is There Gender Bias in the Peer Review Process at Journal of Neurophysiology'? *Journal of Neurophysiology*, 101: 2195–6.

Langfeldt, L. (2001) 'The Decision-Making Constraints and Processes of Grant Peer Review, and Their Effects on the Review Outcome', *Social Studies of Science*, 31: 820–41.

Laudel, G. (2006) 'Conclave in the Tower of Babel: How Peers Review Interdisciplinary Research Proposals', *Research Evaluation*, 15: 57–68.

Lee, C. (2006) 'Perspective: Peer Review of Interdisciplinary Scientific Papers', *Nature*, 5034.

Lee, C. J., and Schunn, C. D. (2011) 'Social Biases and Solutions for Procedural Objectivity', *Hypatia: A Journal of Feminist Philosophy*, 26: 352–73.

Lee, C. J. et al. (2013) 'Bias in Peer Review', *Journal of the American Society for Information Science and Technology*, 64: 2–17.

Leydesdorff, L. (2018) 'Diversity and Interdisciplinarity: How Can One Distinguish and Recombine Disparity, Variety, and Balance?', *Scientometrics*, 116: 2113–21.

Leydesdorff, L., and Rafols, I. (2011) 'Indicators of the Interdisciplinarity of Journals: Diversity, Centrality, and Citations', *Journal of Informetrics*, 5: 87–100.

Leydesdorff, L., Wagner, C. S., and Bornmann, L. (2019) 'Interdisciplinarity as Diversity in Citation Patterns among Journals: Rao-Stirling Diversity, Relative Variety, and the Gini Coefficient', *Journal of Informetrics*, 13: 255–69.

Linder, J. A. et al. (2014) 'Time of Day and the Decision to Prescribe Antibiotics', *JAMA Internal Medicine*, 174: 2029–31.

Muraven, M., and Baumeister, R. F. (2000) 'Self-Regulation and Depletion of Limited Resources: Does Self-Control Resemble a Muscle?', *Psychological Bulletin*, 126: 247.

Mallard, G., Lamont, M., and Guetzkow, J. (2009) 'Fairness as Appropriateness: Negotiating Epistemological Differences in Peer Review', *Science, Technology, & Human Values*, 34: 573–606.

Mansilla, V. B., Feller, I., and Gardner, H. (2006) 'Quality Assessment in Interdisciplinary Research and Education', *Research Evaluation*, 15: 69–74.

Marsh, H. W., Jayasinghe, U. W., and Bond, N. W. (2008) 'Improving the Peer-Review Process for Grant Applications: Reliability, Validity, Bias, and Generalizability', *American Psychologist*, 63: 160.

Mayo, N. E. et al. (2006) 'Peering at Peer Review Revealed High Degree of Chance Associated with Funding of Grant Applications', *Journal of Clinical Epidemiology*, 59: 842–8.

Mutz, R., Bornmann, L., and Daniel, H.-D. (2012a) 'Does Gender Matter in Grant Peer Review? An Empirical Investigation Using the Example of the Austrian Science Fund', *Zeitschrift Für Psychologie*, 220: 121.

Mutz, R., Bornmann, L., and Daniel, H.-D. (2012b) 'Heterogeneity of Inter-Rater Reliabilities of Grant Peer Reviews and Its Determinants: A General Estimating Equations Approach', *PLoS One*, 7: e48509.

Olbrecht, M., and Bornmann, L. (2010) 'Panel Peer Review of Grant Applications: What Do we Know from Research in Social Psychology on Judgment and Decision-Making in Groups?', *Research Evaluation*, 19: 293–304.

Patterson, J. D., and Kurtz, K. J. (2020) 'Comparison-Based Learning of Relational Categories (You'll Never Guess)', *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 46: 851–71.

Peters, D. P., and Ceci, S. J. (1982) 'Peer-Review Practices of Psychological Journals: The Fate of Published Articles, Submitted Again', *The Behavioral and Brain Sciences*, 5: 187–95.

Pier, E. L. et al. (2018) 'Low Agreement among Reviewers Evaluating the Same NIH Grant Applications', *Proceedings of the National Academy of Sciences*, 115: 2952–7.

Pina, D. G., Hren, D., and Marušić, A. (2015) 'Peer Review Evaluation Process of Marie Curie Actions under EU's Seventh Framework Programme for Research', *PLoS One*, 10: e0130753.

Pocheptsova, A. et al. (2009) 'Deciding without Resources: Resource Depletion and Choice in Context', *Journal of Marketing Research*, 46: 344–55.

Porter, A. L., Garner, J., and Crowl, T. (2012) 'Research Coordination Networks: Evidence of the Relationship between Funded Interdisciplinary Networking and Scholarly Impact', *Bioscience*, 62: 282–8.

Porter, A. L., and Rossini, F. A. (1985) 'Peer Review of Interdisciplinary Research Proposals', *Science, Technology, & Human Values*, 10: 33–8.

Rafols, I. et al. (2012) 'How Journal Rankings Can Suppress Interdisciplinary Research: A Comparison between Innovation Studies and Business & Management', *Research Policy*, 41: 1262–82.

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

Sandström, U., and Hällsten, M. (2008) 'Persistent Nepotism in Peer-Review', *Scientometrics*, 74: 175–89.

Steiner Davis, M. L. et al. (2020) 'What Makes an Effective Grants Peer Reviewer? An Exploratory Study of the Necessary Skills', *PLoS One*, 15: e0232327.

Stirling, A. (2007) 'A General Framework for Analysing Diversity in Science, Technology and Society', *Journal of the Royal Society Interface*, 4: 707–19.

Tamblyn, R. et al. (2018) 'Assessment of Potential Bias in Research Grant Peer Review in Canada', *CMAJ*, 190: E489–99.

Tan, E. et al. (2016) 'Validating Grant-Making Processes: Construct Validity of the 2013 Senior Corps RSVP Grant Review', *VOLUNTAS: International Journal of Voluntary and Nonprofit Organizations*, 27: 1403–24.

Tomkins, A., Zhang, M., and Heavlin, W. D. (2017) 'Reviewer Bias in Single-versus Double-Blind Peer Review', *Proceedings of the National Academy of Sciences*, 114: 12708–13.

Uzzi, B. et al. (2013) 'Atypical Combinations and Scientific Impact', *Science*, 342: 468–72.

Van der Lee, R., and Ellemers, N. (2015) 'Gender Contributes to Personal Research Funding Success in The Netherlands', *Proceedings of the National Academy of Sciences*, 112: 12349–53.

Van der Linden, D., Frese, M., and Meijman, T. F. (2003) 'Mental Fatigue and the Control of Cognitive Processes: Effects on Perseveration and Planning', *Acta Psychologica*, 113: 45–65.

Volker, B., and Steenbeek, W. (2015) 'No Evidence That Gender Contributes to Personal Research Funding Success in The Netherlands: A Reaction to Van Der Lee and Ellemers', *Proceedings of the National Academy of Sciences*, 112: E7036–7.

Wagner, C. S. et al. (2011) 'Approaches to Understanding and Measuring Interdisciplinary Scientific Research (IDR): a Review of the Literature', *Journal of Informetrics*, 5: 14–26.

Weller, A. C. (2001). *Editorial Peer Review: Its Strengths and Weaknesses*. Medford, NJ: Information Today, Inc.

Whitley, R. (2000). *The Intellectual and Social Organization of the Sciences*. Oxford: Oxford University Press on Demand.

Wing, D. A. et al. (2010) 'Differences in Editorial Board Reviewer Behavior Based on Gender', *Journal of Women's Health*, 19: 1919–23.

Wood, F. Q. (1997). *The Peer Review Process*. Canberra, Australia: National Board of Employment, Education and Training.