Research Evaluation, 26(3), 2017, 181–189
doi: 10.1093/reseval/rvx021
Advance Access Publication Date: 30 May 2017
Article

Blinding applicants in a first-stage peer-review process of biomedical research grants: An observational study

Maite Solans-Domènech[1,2,3,*], Imma Guillamón[1,2], Aida Ribera[2,4], Ignacio Ferreira-González[4], Carme Carrion[3,5], Gaietà Permanyer-Miralda[1,4] and Joan M. V. Pons[1,2]

[1]Agency for Health Quality and Assessment of Catalonia (AQuAS), Roc Boronat 81-95, Barcelona, Catalonia 08020, Spain, [2]Epidemiology and Public Health Network (CIBER ESP), Roc Boronat 81-95, Barcelona, Catalonia 08020, Spain, [3]Faculty of Health Sciences, Universitat Oberta de Catalunya, Barcelona, Catalonia 08018, Spain, [4]Epidemiology Unit, Cardiology Service, Vall d'Hebron Hospital, Passeig de la Vall d'Hebron, 119-129, Barcelona, Catalonia 08035, Spain and [5]Laboratory of Translational Medicine and Decision Science (TransLab Research Group), Department of Medical Sciences, Faculty of Medicine, University of Girona, Girona, Catalonia 17071, Spain

*Corresponding author: E-mail: mtsolans@gencat.cat.

Abstract

To blind or not researcher's identity has often been a topic of debate in the context of peer-review process for scientific publication and research grant application. This article reports on how knowing the name and experience of researchers/institutions influences the qualification of a proposal. We present our experience of managing the peer-review process of different biomedical research grants. The peer-review process included three evaluation stages: first, blinded assessment; second, unblinded assessment by the same reviewer; and final, assessment of the better qualified proposals by an *ad hoc* committee. The change between the first (applicants blinded) and the second assessments (unblinded) for each evaluation and reviewer was evaluated. Factors associated with change were analysed, taking into account the characteristics of proposals, reviewers, and researchers. A qualitative content analysis of the reviewers' comments was also carried out to assess the reasons for change. The analysis of 5,002 evaluations indicated that in 18.5% of the evaluations (from 10.5 to 27.7% depending on the year of the edition), the reviewer changed the second assessment: either for better (11.9%) or worse (6.6%). Our findings also suggest that a change in the second assessment was highly correlated with a positive evaluation of the experience of the principal investigator or research team. With a change of 1 in 10 to 1 in 4 depending on the year of the edition, we believe that concealing the identity of researchers/institutions could help to focus exclusively on the proposal and reduce some of the common biases of the peer-review process in grant decisions.

Key words: grant peer review; financing; organized; double-blind method; qualitative research; quality control.

1. Introduction

Peer review is the most widespread system used to allocate research funds and to appraise scientific manuscripts for publication. Although imperfect, it is an accepted instrument for self-policing and ensuring quality in scientific research, having gained as much legitimacy in the scientific world as among the lay public (Gurwitz et al. 2014). Since the 80s, there has been a quest to improve the peer-review process in its robustness, fairness, and transparency.

As an object of scientific study, the peer-review process is only a few decades old. Most of the literature has focused on the peer-

review process regarding submissions to journals (Mayo et al. 2006). Blinding reviewers and/or authors of manuscripts has often been a topic of discussion (Wessely 1998; Regehr and Bordage 2006; Baggs et al. 2008; Mulligan et al. 2013). In principle, blinding can reduce bias against some researchers' characteristics (i.e. female or junior researchers) (Triggle and Triggle 2007; Budden et al. 2008) or proposal characteristics (i.e. untried, very innovative or disruptive, and interdisciplinary proposals) (Wessely 1998). Avoidance of any interpersonal conflict of interest is another argument supporting the blinding of identities (Baggs et al. 2008). Most of the evidence available related to the effect of blinding the author on the quality of reviews of manuscripts submitted to the editorial peer-review system demonstrates little or no effect (Justice et al. 1998; Van Rooye et al. 1998, 1999; Alam et al. 2011).

There has been less written on peer review of grant applications. One of the best-known studies of peer review of grant applications was that performed in the 70s to 80s at the request of the Committee on Science and Public Policy of the US National Academy of Sciences on the review process of the National Science Foundation (NSF). Experimentally, half of the reviewers received proposals that had been edited in an attempt to conceal the applicant's identity; the other half received copies identical to those but without concealing the applicants, as they had been submitted to the NSF. Results of the peer review in the NSF showed that an applicant's age and track record had little effect on the chances of getting a grant and that reviewers treated proposals from researchers at prestigious institutions no differently than proposals from workers at less prestigious institutions (Garfield 1987). Reviewers of blinded proposals were also asked whether the removal of title pages, list of references, budgets, or any other identifying information made the proposal more difficult to evaluate. The study found that it was difficult to conceal authorship because it made the proposal almost unreadable. Also, complete blinding often seems difficult to achieve (because of the many internal clues pointing towards authorship included in the articles) (Van et al. 1999). The Committee concluded that the blinding process of grant applications severely compromised the integrity of the proposals. Contrary to manuscript peer review, which is an ex-post research assessment, review of grant for funding (ex ante research assessment) was considered highly dependent on the principal investigator's (PI)/research team's ability to adequately implement the proposal. Another study that analysed the gap in success rates between different races and ethnicities showed a bias against black or Asian researchers, even after controlling for education, country of origin, training, employer characteristics, previous research awards, and publication record (Tabak and Collins 2011).

Since then, more literature has been added and many different biases have been reported in the peer-review process: age, institution, 'cronyism', discipline, gender, etc. A non-systematic review of the existing studies on peer review for awarding grants was published at the end of last century by Wessely (1998) which examined issues of equity, efficiency, and failure to promote the best science. A lack of reliability in the rankings of reviewers was, among other things, one of the main weaknesses considered. More recently the Cochrane Collaboration performed another more comprehensive and systematic review, examining the effects of peer review in awarding grants, taking into account different ways of screening, assigning, or masking submissions; different ways of eliciting internal or external opinions; different ways of carrying out procedures (single person or group); and different types of feedback given and revisions done of applications (Demicheli and Di Pietrantonj 2007). The authors of the review concluded that 'there is little empirical evidence of the effects of peer review in awarding grants' (importance, relevance, usefulness, soundness of methods, soundness of ethics, completeness, and accuracy) because they were unable to find comparative studies assessing the actual effect of peer-review procedures on the quality of the research funded. There was an urgent need for research to fill this gap and, as Wessely mentioned, the absence of controlled trials in this area of scientific decision making was ironic (Wessely 1998).

A recent call for greater transparency in reviewing grant applications once again mentioned both the low agreement between reviewers in their qualifications and also the more recent case of the NSF where reviewers faced with blinded proposals selected a different set of projects for funding than those chosen by reviewers with unblinded versions of the same proposal (Bhattacharjee 2012; Gurwitz et al. 2014). In any case, a recent survey has shown that blinding applicants is extremely rare among public research financing agencies, with only 4% of the organizations doing so (ESF 2011).

The peer-review process described here, and managed for more than 10 years now, consists in a first blinded assessment followed by a second unblinded assessment. So, the goal of this study is to analyse whether concealing the identity of researchers and their institutions from peer reviewers in the first assessment stage of a research project changes the reviewer's assessment when the name of the PI/research team, their experience, and the institution they represent is revealed in a second stage. Specifically, our intention was to quantify the change, its direction, and to know whether there are any factors associated and the reasons that led to it.

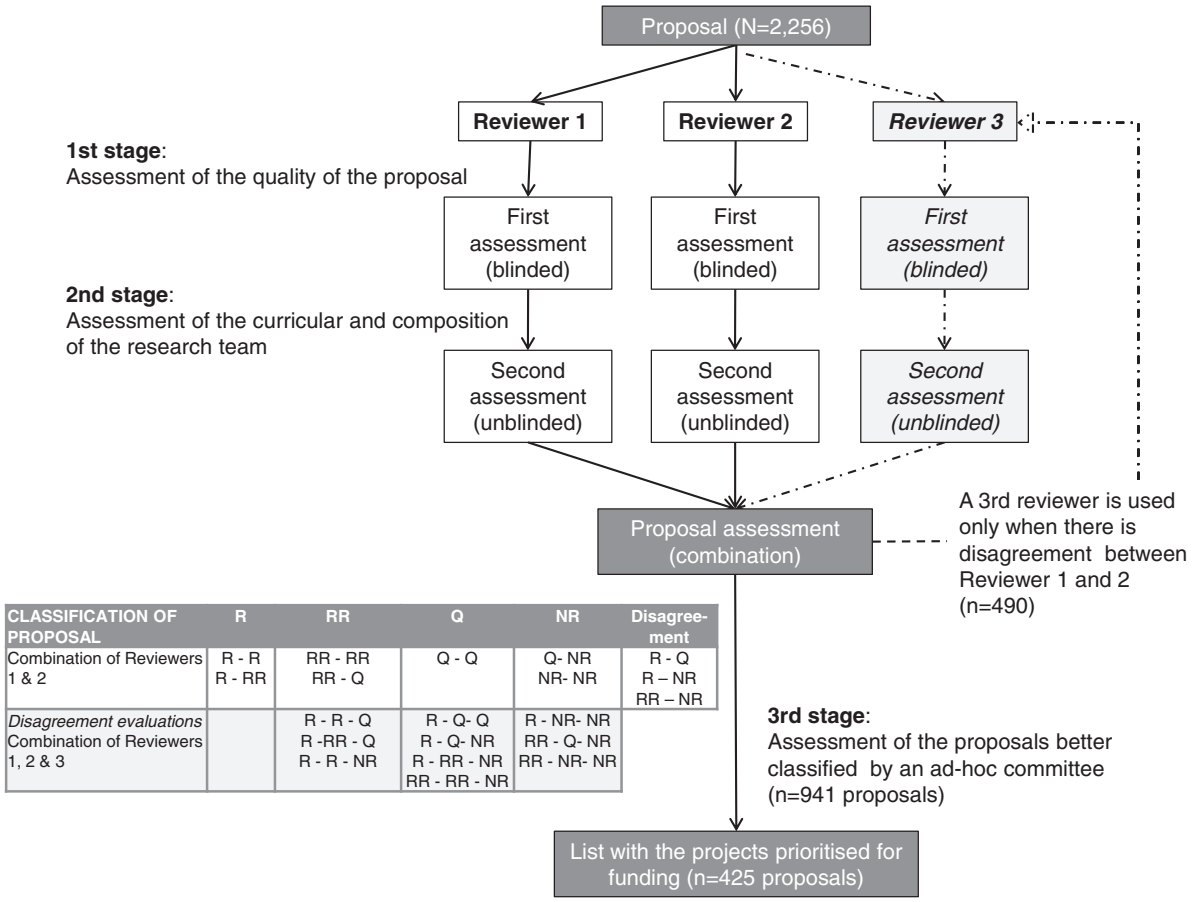
2. Methodology

A retrospective observational study of the peer-review process carried out to evaluate different biomedical research grant applications. The sample consisted of all research proposals (N = 2,256) presented in 14 annual calls from 2002 to 2015 (2001–2014 editions). Projects were evaluated by 1,475 international reviewers (about two to three reviews for each proposal, on average 2.2). Overall, 5,002 evaluations were conducted. First, we must describe the assessment process.

2.1 Peer-review process overview

Since 2001, the Agency for Health Quality and Assessment of Catalonia (AQuAS) has been in charge of managing the peer-review process of different annual research calls for biomedical research projects. Unlike other research calls, the main topic of interest in these calls changes each year. From the outset, proposals have had to be submitted in English, and the peer-review process, done in three stages, only has non-Spaniard reviewers who independently assess the scientific quality of projects (Fig. 1). Reviewers are mainly selected by searching through medical literature, or from scientific societies, editorial journals, or reviewer repositories.

The first assessment stage (blinded) begins when the details of all researchers linked to a research proposal and the institutions they represent are concealed. Researchers' names, manuscripts' titles, volume, and pages (but not journal identification or year of publication) are also suppressed from self-references in the text and in the bibliographic section. Therefore, the assessment only focuses on the content of the research proposal and performed by using a



CLASSIFICATION OF PROPOSAL	R	RR	Q	NR	Disagree-ment
Combination of Reviewers 1 & 2	R - R R - RR	RR - RR RR - Q	Q - Q	Q- NR NR- NR	R - Q R – NR RR – NR
Disagreement evaluations Combination of Reviewers 1, 2 & 3		R - R - Q R -RR - Q R - R - NR	R - Q- Q R - Q- NR R - RR - NR RR - RR - NR	R - NR- NR RR - Q- NR RR - NR- NR	

Figure 1. Workflow of the review process.
Note: R = recommended for funding; RR = recommended with reservations; Q = Questionable; NR = not recommended.

structured questionnaire with response categories on a Likert scale. At the end of the questionnaire, the reviewers are asked to state whether, in qualitative terms, the project should be recommended (R), recommended with reservations (RR), questionable (Q), or not recommended (NR) for funding. In the second stage (unblinded), reviewers once again use a structured questionnaire to evaluate the experience and track record of the PI, the research team, and the suitability of the institution where the research is planned. At this point, the reviewers have to give their second and definitive qualitative assessment of the project in the same terms (R, RR, Q, or NR). When there is a two-level disagreement in this second stage between two reviewers, the proposal is sent to a third reviewer for another independent evaluation performed in the same way. The time elapsed between the first and second assessments is approximately 1–2 weeks. The second stage begins when the AQuAS receives the first completed questionnaire. The two or more categories obtained from the assessments are combined for each proposal into one that allows the proposal to be classified in self-excluding categories (R, RR, Q, and NR) (see Fig. 1).

The third and final stage of the process is performed by an *ad hoc* committee composed of some of the international reviewers who meet over a 2-day period. These reviewers are in charge of assessing the best qualified proposals in the second stage (usually those classified as R, and sometimes also the RR's). The committee drafts a list of the projects prioritized for funding, taking into account the amount of funds available. This list is presented to the respective Scientific Advisory Committee, and the final decision is made by the board of trustees. The process ends when the proposals awarded grants are made public, and a report describing the entire peer-review process and the reviewers participating in it is published.

2.2 Quantitative analysis

The primary variable of analysis was the change made between the first (researchers/institutions blinded) and the second assessments (unblinded) for each evaluation and for each reviewer. The variable was categorized on an ordinal scale of 0 = no change, 1 = improvement (change for a better category), and 2 = worsening (change for a worse one). The relationship between the primary variable and the adequation of the PI/research team/institution was analysed with a chi-square test. The PIs/research teams/institutions were rated in the second-stage questionnaire by the reviewers according to their appropriateness to carry out the proposal and placed in one of four categories (strongly agree, agree, disagree, or strongly disagree). This variable was dichotomized for this study. Agreement between the first (blinded) and the second assessments (unblinded) was calculated using the weighted kappa statistic (k).

An adjusted multinomial logistic regression model was used to identify those factors associated with change with 'no change' as the reference category. Relative risk ratios (RRRs) were calculated

for improvement and for worsening. Several covariates were included in the data analyses as predictors, both at the level of proposals and PIs, and reviewers. At the level of proposals, 'year of the edition', 'research area of the proposal' identified by the applicant, 'requested grant sum', 'PI's gender', 'PI's age', and 'adequate experience of the research team' were taken into account. The maximum sum of the grant requested allowed for variations depending on the type of proposal; individual with a single group/institution coordinated between two groups or coordinated among three or more participating groups from different institutions. At the level of reviewers, the 'reviewer's gender', 'world region of the reviewer', and 'reviewer h-index' were included. The gender (and age) of the applicant was obtained from the forms completed on submission of the research project or in the reviewer acceptance form. If gender was not specifically identified, we conducted a manual search (using websites or corresponding addresses). In particular, Asian names were difficult to assign to a specific gender, so in those cases the gender variable was left blank. The h-index was determined for each reviewer using *Web of Science* and taking into account the year of the evaluation. The value of h is equal to the number of papers of the reviewer (N) that have N or more citations. The statistical significance was set at $P \leq 0.05$. Statistics were calculated using SPSS18.

2.3 Qualitative analysis

To analyse the reasons that indirectly influenced a change in the second assessment (unblinded), a qualitative content analysis was carried out. As the evaluation questionnaire does not have a specific field for documenting the reasons for a change, we analysed the open field for additional comments included in the second-stage evaluation form. The only cases examined were those in which a change was made and a comment written by the reviewer. Of the 922 evaluations that modified the assessment, half of them (n = 461) filled in the open field for additional comments. The content analysis was conducted in an inductive two-dimensional way: reasons for change and nature of the reason. First, one author read all the comments to get an overall understanding of the reviewers' comments and to extract the initial categories and subcategories. Then, the comments were classified into these categories and subcategories by all the authors. Secondly, we determined whether the justifications were of a positive, negative, or neutral nature. A justification was defined as positive when the reviewer described excellent aspects of the project; a negative justification meant that the characteristics evaluated did not seem appropriate or were absent; and a neutral character was considered when it was impossible to classify the justification as either positive or negative. The results were triangulated by two coders, who reached a consensus and discussed those cases classified as 'doubtful' within the multidisciplinary team.

3. Results

Table 1 displays a description of the different editions analysed. During the period 2001–2014, more than 101 million € have been distributed among 376 projects, with a success rate of 18.6% for the 2,256 proposals presented. Overall 1,475 international reviewers participated in this assessment process (some participated in more than one edition). A third reviewer to solve discordances was needed in 490 cases (9.8%).

3.1 Quantitative analysis

We analysed 5,002 evaluations. In most of them (81.5%) and after the PI/research teams/institution were unblinded, reviewers did not change their second assessment of the proposal, while in 18.5% (n = 922) cases, there was a change in the assessment: it was for better in 11.9% (n = 594) cases and for worse in 6.6% (n = 328). Depending on the year of the edition year, the percentage of change ranged from 10.5 to 27.7% (Table 1).

The association between an adequate experience of the PI/research team and the change in the second assessment was statistically significant ($P < 0.05$). The change was for the better when there was substantial agreement on the 'adequate experience' of the PI/research team, while it was for the worse when there was considerable disagreement.

Both the first (blinded) and the second (unblinded) assessments included similar rates of the four possible categories (R, RR, Q, and NR). A weighted Kappa statistic indicates a very good agreement ($k = 0.75$) between the first and second assessments.

In the adjusted multinomial model, a positive evaluation of the experience of the PI/research team/institution showed the strongest association with a positive change (RRR = 2.63; $P < 0.005$), and it was less likely to have a negative change (RRR = 0.25; $P < 0.005$). Earlier editions, from 2001 to 2007 and 2009 edition) were also found to be a factor associated with positive change, and there were only negative changes in the years 2002 and 2004. Multivariate analyses also showed that, compared to no change, a positive change was less likely to be present with reviewers coming from North America (RRR = 0.62; P = 0.001), in comparison to European reviewers. In contrast, when a budget of more than €300,000 was requested, there was less likelihood (RRR = 0.49; P = 0.01) of a negative change. Female PI was also a statistical significant factor associated with a negative change (RRR = 1.42; P = 0.001) with no change as reference category (Table 2). Other factors such as research area of the proposal, PI's age, reviewer's gender, or reviewer h-index had no association with any positive or negative change.

3.2 Qualitative analysis

There was an added comment in 50% (461) of the evaluations which changed the qualification of the project in the second assessment stage. Of these comments, 5% (n = 23) were inconsistent with the change, i.e. the reviewer improved the second assessment, although comments were unfavourable (n = 18), or he/she worsened the second assessment without a clear rationale for change (n = 5). That means that the justification could be 'positive', when it presented the strengths of the research team or the proposal, or 'negative', when the assessment implied its weaknesses.

The justification ranged in length from 2 ('Excellent team' [EvaluationID 90]) to 452 words. The themes emerging from the reviewers' comments fell into three main groups related to: (1) an evaluation, (2) a suggestion, or (3) a comment about the lack of information either related to the PI/research team or to the proposal. The comments that describe an evaluation (first group) referred to, first, the skills and experience of the research team and/or the PI that demonstrate whether they have an adequate background for carrying out the proposal, mostly measured by the publications of group members. Also, it was measured regarding multi-institutional collaboration, which was not only related to the need for more participating centres (i.e. to collect enough samples) but also to the aspect of how the multiple centres should be coordinated. Finally, it also

Table 1. Descriptive characteristics of the different editions

Year	2001	2002	2003	2004	2005	2006	2007	20008	2009	2010	2011	2012	2013	2014	Total
Number of proposals presented	107	80	105	260	200	81	159	151	244	100	205	317	147	100	**2,256**
Research area of proposal: number (%)															
Basic	49 (46)	38 (48)	32 (30)	132 (51)	139 (70)	26 (32)	66 (42)	47 (31)	128 (52)	40 (40)	119 (58)	162 (51)	81 (55)	27 (27)	**1,086 (48)**
Clinical	33 (31)	16 (20)	47 (45)	69 (27)	42 (21)	39 (48)	49 (31)	60 (40)	53 (22)	35 (35)	48 (23)	62 (20)	29 (20)	28 (28)	**610 (27)**
Epidemiology/health services	15 (14)	2 (3)	10 (10)	22 (8)	3 (2)	8 (10)	14 (9)	20 (13)	5 (2)	4 (4)	6 (3)	22 (7)	3 (2)	6 (6)	**140 (6)**
Combination	10 (9)	24 (30)	16 (15)	37 (14)	16 (8)	8 (10)	30 (19)	24 (16)	58 (24)	21 (21)	32 (16)	71 (22)	34 (23)	39 (39)	**420 (19)**
Type of proposal: number (%)															
Single institution	88 (82)	55 (69)	85 (81)	214 (82)	144 (72)	71 (88)	106 (67)	99 (66)	140 (57)	62 (62)	121 (59)	185 (58)	71 (48)	63 (63)	**1,504 (67)**
Coordinated	19 (18)	25 (31)	20 (19)	46 (48)	56 (28)	10 (12)	53 (33)	52 (34)	104 (43)	38 (38)	84 (41)	132 (42)	76 (52)	37 (37)	**752 (33)**
Number of international reviewers	52	34	52	169	201	92	145	131	134	111	105	257	96	59	**1,638[a]**
Number of proposals funded	21	20	26	30	35	28	25	26	20	30	29	42	44	43	**376**
Amount of money collected (million €)	3.9	3.8	3.3	7.7	6.9	7.1	7.0	6.1	6.3	8.7	8.0	11.3	11.2	10.3	**101.6**
Average funding per proposal (thousand €)	186.1	188.8	128.6	256.4	196.8	254.1	281.0	236.4	317.5	291.1	277.1	269.9	255.0	239.6	**243.2**
Proposals success rate[b] (%)	19.6	25.0	24.8	11.5	17.5	34.6	15.7	17.2	8.2	30.0	14.1	13.2	29.9	43	**18.6**
Percentage of evaluations that change the qualitative category (%)	26.2	24.4	24.7	23.7	27.7	17.3	21.9	15.3	20.9	11.3	10.5	12.7	13.5	11.1	**18.5**

Note: [a]Some of the reviewers participated in more than one edition.
[b]Proposal success rate was calculated dividing the number of proposals funded by the number of proposals presented.

included the composition and multidisciplinary nature of the team, which was described in relation to the different backgrounds, expertise, and specialization of its members. On the other hand, the assessment of a proposal was related to its strengths and weaknesses, such as relevance, quality, or importance, characteristics that were in fact part of the first-stage assessment. The second group (a suggestion) includes suggestions regarding the research team, which were mainly related to the necessity of additional expertise and suggestions regarding the proposal, including a wide array of aspects such as budget, hypothesis, planning. Finally, the third group, regarding a lack of information, was considered when reviewer justification described that the proposal did not possess enough information to evaluate one specific aspect (either of the research team or the proposal). An overview of the results (and quotes) is shown in Table 3.

4. DISCUSSION

The results of this analysis indicate that in 18.5% of the evaluations, from 10 to 28% depending on the edition, the assessment of a research proposal was changed after the identity and experience of the researchers and their institutions were revealed. Therefore, most of the reviewers maintain their initial assessment of the blinded proposal. As expected, our findings also imply that the change in the second assessment was highly correlated with the evaluation of the appropriateness of the PI/research team's experience. This is not surprising, as the track record of the researchers and their institutions was the only new information received in the second stage. These results are also supported by the qualitative analysis.

The fact that almost 19% of the assessments were changed after the details of the researchers were made known opens a further discussion about the time and cost of blinding applicants. Is this percentage enough to continue insisting on and ensuring a researcher's anonymity for the first (blinded) assessment stage of proposals? It is not very common among agencies financing research to request that proposals be blinded as a recent survey has shown (ESF 2011), and it is even less common to ensure its effectiveness. However, there was no question in the survey regarding the blinding of applicants in the first-stage review process.

In the qualitative analysis of a reviewer's comments, the reasons for changing an assessment were mainly characterized by a positive or a negative evaluation of the PI/research team. The skills and experience of the research team and/or the PI, multi-institutional collaboration, or the composition and multidisciplinary nature of the team were factors that the reviewers commented on in the second phase of evaluation as a reason to change their assessment. Therefore, this played an important part in judging the feasibility of a proposal and supports the studies of the NSF which considered that blinding applicants compromised the integrity of a proposal. However, in almost half of the evaluations that changed in the second assessment, no comment was included by the reviewer and as such, in these cases, it was impossible to know what the justification was for the change. Moreover, in 5% of the cases, the comments made seemed to be incongruous with a change in the opposing direction of the assessment. Our results also seem to reaffirm the importance of evaluating a proposals' characteristics (quality, originality, methodology, innovation, etc.), as other studies have shown (Abdoul et al. 2012), because the assessments remained unchanged in the majority of cases.

Table 2. Factors predicting change between the first assessment (applicants blinded) and the second assessment (unblinded)

Factors	Number of evaluations (% total)	Positive change rate (% category)	Negative change rate (% category)	Adjusted model RRR (95% CI)	
				Positive change	Negative change
Edition year					
2001	214 (4.28)	41 (19.16)	15 (7.01)	**3.83 (1.91–7.68)**	1.45 (0.58–3.58)
2002	160 (3.20)	21 (13.13)	18 (11.25)	**2.55 (1.20–5.40)**	**2.47 (1.04–5.86)**
2003	231 (4.62)	33 (14.29)	24 (10.39)	**3.03 (1.53–6.03)**	2.16 (0.96–4.88)
2004	586 (11.73)	78 (13.31)	61 (10.41)	**2.35 (1.28–4.32)**	**2.42 (1.18–4.99)**
2005	451 (9.03)	96 (21.29)	29 (6.43)	**4.02 (2.19–7.37)**	1.20 (0.55–2.61)
2006	185 (3.70)	22 (11.89)	10 (5.41)	**2.33 (1.13–4.80)**	0.88 (0.34–2.25)
2007	356 (7.13)	56 (15.73)	22 (6.18)	**2.78 (1.46–5.32)**	1.48 8 (0.64–3.41)
2008	339 (6.79)	29 (8.55)	23 (6.78)	1.03 (0.47–2.24)	1.07 (0.42–2.68)
2009	540 (10.81)	74 (13.70)	39 (7.22)	**2.36 (1.29–4.34)**	1.94 (0.93–4.07)
2010	222 (4.44)	14 (6.31)	11 (4.95)	0.98 (0.45–2.17)	0.73 (0.28–1.88)
2011	457 (9.15)	29 (6.35)	19 (4.16)	0.94 (0.48–1.85)	0.80 (0.35–1.79)
2012	703 (14.07)	61 (8.68)	28 (3.98)	1.44 (0.78–2.67)	0.79 (0.37–1.70)
2013	326 (6.53)	25 (7.67)	19 (5.83)	1.20 (0.60–2.39)	1.19 (0.53–2.72)
2014	225 (4.50)	15 (6.67)	10 (4.44)	Ref.	Ref.
Research area of the proposal					
B	2,387 (47.79)	309 (12.95)	159 (6.66)	0.99 (0.62–1.57)	1.09 (0.61–1.98)
C	1,354 (27.11)	143 (10.56)	95 (7.02)	1.03 (0.80–1.32)	1.21 (0.84–1.73)
E	306 (6.13)	36 (11.76)	25 (8.17)	0.88 (0.65–1.18)	1.01 (0.68–1.51)
Combinations	948 (18.98)	106 (11.18)	49 (5.17)	Ref.	Ref.
Requested grant sum					
<€100,000	471 (9.44)	45 (9.55)	46 (9.77)	Ref.	Ref.
€100,000–€199,999	2,503 (50.15)	294 (11.75)	189 (7.55)	1.22 (0.84–1.76)	0.98 (0.66–1.45)
€200,000–€299,999	929 (18.61)	104 (11.19)	53 (5.71)	1.28 (0.84–1.95)	0.73 (0.45–1.19)
≥€300,000	1,088 (21.80)	150 (13.79)	40 (3.68)	1.49 (0.99–2.25)	**0.49 (0.29–0.82)**
PI's gender					
Female	1,661 (33.25)	190 (11.44)	126 (7.59)	0.99 (0.81–1.21)	**1.42 (1.10–1.83)**
Male	3,334 (66.75)	404 (12.12)	202 (6.06)	Ref.	Ref.
PI's age					
≤40	799 (16.76)	103 (12.89)	58 (7.26)	1.04 (0.81–1.33)	1.10 (0.80–1.51)
>40	3,969 (83.24)	459 (11.56)	253 (6.37)	Ref.	Ref.
Adequate experience of the research team					
Strongly agree/agree	4,310 (87.57)	562 (13.04)	213 (4.94)	**2.63 (1.73–4.00)**	**0.25 (0.19–0.32)**
Disagree/strongly disagree	612 (12.43)	28 (4.58)	111 (18.14)	Ref.	Ref.
Reviewer's gender					
Female	933 (18.84)	109 (11.68)	59 (6.32)	1.02 (0.80–1.29)	0.88 (0.64–1.22)
Male	4,018 (81.16)	484 (12.05)	268 (6.67)	Ref.	Ref.
World region of the reviewer					
Europe	2,838 (56.82)	322 (11.35)	208 (7.33)	Ref.	Ref.
North America	1,697 (33.97)	228 (13.44)	101 (5.95)	**0.62 (0.43–0.89)**	0.65 (0.37–1.15)
Other	460 (9.21)	44 (9.57)	19 (4.13)	0.84 (0.69–1.03)	1.27 (0.96–1.69)
Reviewer h-index					
≤15	2,279 (46.95)	292 (12.81)	164 (7.20)	1.08 (0.88–1.32)	1.09 (0.83–1.43)
>15	2,575 (53.05)	287 (11.15)	154 (5.59)	Ref.	Ref.

Note: Bold values indicate statistically significant findings.

Nagelkerke $R^2 = 0.102$.

CI = confidence interval; B = basic research; C = clinical research; E = epidemiological research; ref. = reference variable for RRR calculation.

In the design of our review process, significant efforts are made to ensure applicants are blinded in the first stage. All proposals are checked one by one to remove researcher's details, even though the research call instructions state that no information in the proposal may reveal a researcher's identity or affiliation. A random number of proposals are checked again with positive results in the majority of cases analysed.

We cannot ignore some points that favour concealing the identity of applicants and institutions. The main sources of bias in the peer-review process are related to the applicant, the project, or the potential relationship between the reviewer and the applicant (Lee 2000). In this last case, a strict policy of conflict of interest determines that the reviewer, either as collaborator or competitor, is not appropriate to evaluate the proposal, and this helps to prevent

Table 3. Description of the reasons for changing the qualitative assessment made by reviewers

Reasons to change	Classification	Type of change	Examples
Assessment of the team	Skills and experience	Positive	The experience of the researchers changed my mind regarding the unclear feasibility of the study. The study will be quite feasible. Even though some concerns regarding study design and data analysis remain, the study has potential. [ID 17]
		Negative	This project would be done by three people with rather varying degrees of expertise. Having reviewed the C's, it is doubtful to the reviewer whether this group may fulfil the ambitious goals set forth in the research plan. [ID 624]
	Multicentrism and collaboration	Positive	The different centres involved in the work for this proposal have been collaborating in the past with good results. This can guarantee the success of proposal, although the ambitions are very high. [ID 223]
		Negative	Given the multicentre structure, it is not sufficiently clear how this complicated effort will be coordinated. [ID 562]
	Composition and multidisciplinary	Positive	The multidisciplinary team is indeed composed of researchers having different backgrounds, expertise, and specialization. Most important, however, is the difference in scientific merits. [ID 667]
		Negative	My initial enthusiasm for this proposal is somewhat diminished by the lack of an expert in cell biology on the research team. The team is very strong in protein biochemistry, but not biological studies that are important for this project. [ID 1170]
Assessment of the proposal	Project	Positive	The applicant refers to important preliminary data that were not available in the first part of the evaluation; I have therefore reconsidered my evaluation in a positive way. [ID 2751]
	Project	Negative	Even though in the past this group has published studies on melanoma patients, in this project no clinical studies are scheduled. This makes the project not particularly interesting. [ID 939]
Suggestions about the team	Composition and multidisciplinary	Positive	The design and statistics of the protocol are flawed and suggest that a statistician would be a worthwhile member of the team. [ID 1592]
Suggestions about the proposal	Project	Positive	I advise the team of the project to focus on one or two tumours seen mostly in their centre, e.g. neuroblastoma (as most publications are on neuroblastoma and few are on sarcomas) and to focus only on one or two translational research issues (e.g. thyrosine hydroxylase . . .) and not to try to do many molecular techniques in the same project. [ID 1014]
Lack of information of the team	Skills and experience	Negative	The researchers did not follow instructions and provide information on their team. Instead they just provided a list of publications. They did describe a thesis which I presume that members of the team were mentors for these projects, but this was not clear. They also provided a list of funding, but it was not clear who received which funding. It is impossible to evaluate the researchers' expertise and experience fully with what they provided. [ID 2633]
Lack of information of the proposal	Project	Positive	I think this is a very interesting and important topic, but it is unclear how the researchers are going to disseminate the environmental information to patients, but of greater importance, there is no description on how any impact such information might have on asthma-related outcomes. Please see my initial review for further methodological issues. [ID 523]

ID = evaluation identification.

cronyism and nepotism. The bias related to a project usually comes from its innovative character or the degree to which it departs from mainstream science, the area where the majority of reviewers come from. However, the most common bias is related to the characteristics of researchers: gender, age, minority group and, specially, the well-known Matthew effect, related to the prestige, reputation, or recognition of researchers or institutions from which they come (Merton 1968). All these points assist us in favouring blinding.

4.1 Comparison with other studies

The only published study that assesses different ways of masking submissions, included in the Cochrane review, is a retrospective comparison between blind and open peer review of research proposals submitted to the Korea Science and Engineering Foundation (KOSEF). The process involved five reviewers for each proposal: three were sighted and two were blinded. A total of 1,978 proposals were sent to 917 reviewers; there were 562 answers, 331 from sighted reviewers and 231 from blinded ones. The study demonstrated that applicant's characteristics were the major factors leading to significant different evaluation scores between blinded and unblinded proposals. Results were considered proof of an obvious bias in the open evaluation of proposals towards researchers from top departments, senior researchers, and academically recognized researchers (Abdoul et al. 2012). In other words, it was reputation that made the difference. Similar results were found in the recent case of the NSF as mentioned above (Bhattacharjee 2012; Gurwitz et al. 2014), although no quantification was presented. Researcher and institutional prestige and even geographic location remain important sources of bias in research grant evaluation (Murray et al.

2016; Wahls 2016). The Matthew effect, or cumulative advantage, affects patterns of scientific collaboration, the growth of biological networks, the propagation of citations, scientific progress and impact, career longevity, as well as many other aspects of human culture (Perc 2014). In our case, it is worth mentioning that when the peer-review process is examined yearly, after the *ad hoc* committee (third and last stage) finishes its tasks, reviewers repeatedly praise the opportunity to assess a blinded proposal in the first stage.

In our study, the frequency of change in the second assessment and its direction (upward or downward) seemed to be affected by the year of the edition, the sum of the grant requested, the world region of the reviewer, and the gender of the PI. A probable explanation for why there were more changes in earlier editions than more recent ones might be that in earlier editions, improvement modifications in the review process were more common. For instance, modifications introduced included that only PIs should present a complete curriculum vitae with a short statement on the research team's experience since 2005, or the homogenization of the presentation of coordinated projects since 2007. Therefore, in more recent editions, the peer-review process has been more standardized, and only minor changes have been applied. We are unable to find another explanation for these differences. However, we must not forget that, as shown in different studies, there is always some inevitable degree of chance associated with the funding of grant applications and that a high number of reviewers is required to gain sufficient consistency to make decisions concerning proposals (Mayo et al. 2006).

Regarding differences depending on a reviewer's world region, it is worth mentioning that an evaluation of the Australian Research Council found that North American reviewers gave statistically significant higher ratings than those from other countries, such as those in Europe (Marsh et al. 2008). Also, the differences in the sum of the grant requested can be explained on the basis that lower amounts are related to single-group/institution proposals, to coordinated proposals having two groups from different institutions, and finally to proposals having three or more groups. This means that, in general, if lower numbers of PI/research teams are included in a proposal, a lower grant budget must be requested, and therefore, it will be more difficult to compensate for the expertise, specialization, and inclusion of different disciplines.

Finally, female PIs were associated with a negative change compared to no change and with respect to men. A meta-analysis of 66 different peer-review studies of grant applications showed a small gender effect in favour of men overall that was marginally statistically significant because of the large sample sizes, although a majority of individual studies showed no significant gender effect (Bornmann et al. 2007). Gender bias remains a disputed issue in the peer-review process with studies showing disparate results (Marsh et al. 2008; Van der Lee and Ellemers 2015). One recent review found only one study on interventions to mitigate gender bias in the peer review of grants (Tricco et al. 2017). This study found no difference in the proportion of women who were successful in receiving grant funding. Other authors speak of the 'Matilda effect', which highlights the historical tendency for women's work to be systematically omitted in the histories of scientific achievement (Rossiter 1993). However, we might assume that more recent editions are less gender biased and that some progress has been made in recent years (8.4% of the negative change in earlier editions versus 6.8% of the negative change in more recent editions, when the PI was female). An additional consideration is that fewer female PIs applied for these grants, though, in the end, the percentage of projects awarded grants to female PIs (33.1%) is almost exactly the same as their application ratio (33.7%). In any case, the design of our study precludes testing the role of gender experimentally, and association, as is well known, does not mean causality.

Although the evidence may not be conclusive, this study also shows that a change made in the second assessment is not affected by other characteristics that, according to the literature, might entail some bias such as a PI's age (Lee 2000) or the research area of a proposal, with some articles suggesting a bias against clinical research compared to molecular research (Marshall 1994). This fact is probably closely related to the reviewer's area of research, and thus, an accurate selection and matching of reviewers according to expertise (topic, and research area) has been attempted in the evaluation process.

4.2 Limitations of the study

The current study has certain limitations. First, the data come from a very specific peer-review process, where concealing applicants and institutions apply only to the first stage, so it might not be possible to extrapolate results to other systems. Secondly, the non-experimental nature of the study must be taken into account. We have no comparison group because data were extracted from the established peer-review process introduced in 2001 which, apart from the time constraints for the reviewers, is compensated for. We did not check if international reviewers were able to identify applicants in the blinded assessment phase, and it cannot be excluded in very specific research topics. However, with the help of electronic means, every effort is made to ensure anonymity when only the project is sent to reviewers (blinded assessment). The fact that the research peer-review process is an *ex ante* assessment of projects should be taken into account, and we do not know if this three-stage process with blinded applicants in the first stage improves the ex-post evaluation with a greater impact of the research, measured, for instance, with publications record and citations. Finally, since the questionnaire does not usually include a specific and compulsory direct question asking what the reasons for a change are, we only analysed an open field describing additional comments, so results were indirectly obtained.

4.3 Conclusions

In conclusion, blinding researchers/institutions in the first stage when assessing research grant proposals affects the second assessment in an average of 19% of the evaluations, from 10.5 to 27.7% depending on the year of edition. Attending to these rates, from 1 in 10 to 1 in 4, we believe that peer-review procedures that facilitate the focus only on the contents of a research proposal at first constitute a promising way to reduce some of the common biases described in the literature regarding researchers' characteristics and research grant decisions. Its implementation would reinforce transparency and accountability, so much in need nowadays for charities and public agencies financing research.

5. What is already known on this topic

To blind an applicant's identity or not has often been a topic of debate in the context of a peer-review process for research funds allocation.

Most of the evidence available on the effect of blinding on the quality of reviews comes from publications submitted to the editorial peer-review system.

6. What this study adds

To our knowledge, there are very few studies about the effect of blinding applicants in a sequential peer-review process of grant allocation for biomedical research projects.

Blinding applicants in the first stage helps to focus on a proposal and reduces biases related to a researcher's characteristics.

Acknowledgements

The authors would like to thank to Marta Aymerich, Maria Dolors Navarro, and Emília Sanchez for their involvement with the AQuAS peer-review process throughout its years in existence. The authors are also grateful to former technicians of AQuAS who contributed to its management.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Transparency

The lead authors (the manuscript's guarantors) affirm that the manuscript is an honest, accurate, and transparent account of the study being reported; that no important aspects of the study have been omitted; and that any discrepancies from the study as planned have been explained.

References

Abdoul, H. et al. (2012) 'Peer Review of Grant Applications: Criteria Used and Qualitative Study of Reviewer Practices', *PLoS One*, 7/9: e46054.

Alam, M. et al. (2011) 'Blinded vs. Unblinded Peer Review of Manuscripts Submitted to a Dermatology Journal: A Randomized Multi-Rater Study', *British Journal of Dermatology*, 165/3: 563–7.

Baggs, J. G. et al. (2008) 'Blinding in Peer Review: The Preferences of Reviewers for Nursing Journals', *Journal of Advanced Nursing* 64/2: 131–8.

Bhattacharjee, Y. (2012) 'Science Funding. NSF's ′Big Pitch′ Tests Anonymized Grant Reviews', *Science*, 336/6084: 969–70.

Bornmann, L. R., Mutz, R., and Daniel, D. (2007) 'Gender Differences in Grant Peer Review: A Meta-analysis', *Journal of Informetrics*, 1/3: 226–38.

Budden, A. E. et al. (2008) 'Double-blind Review Favours Increased Representation of Female Authors', *Trends in Ecology and Evolution*, 23/1: 4–6.

Demicheli, V., and Di Pietrantonj, C. (2007) 'Peer Review for Improving the Quality of Grant Applications', *The Cochrane Database of Systematic Reviews*, 18/2: MR000003.

ESF (2011) *ESF Survey Analysis Report on Peer Review Practices* [Internet]. France: European Science Foundation <http://www.esf.org/fileadmin/Public_documents/. . ./pr_guide_survey.pdf'> accessed May 2015.

Garfield, E. (1987) 'Refereeing and Peer Review: Part 4. Research on the Peer Review of Grant Proposals and Suggestions for Improvement', *Essays of an Information Scientist*, 10: 27–33.

Gurwitz, D., Milanesi, E., and Koenig, T. (2014) 'Grant Application Review: The Case of Transparency', *PLoS Biology*, 12/12: e1002010.

Justice, A. C. et al. (1998) 'Does Masking Author Identity Improve Peer Review Quality? A Randomized Controlled Trial. PEER Investigators', *JAMA*, 280/3: 240–2.

Lee, M. (2000) 'The Bias of Sighted Reviewers in Research Proposal Evaluation: A Comparative Analysis of Blind and Open Review in Korea', *Scientometrics*, 48/1: 99–116.

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*, 63/3: 160–8.

Marshall, E. (1994) 'Does NIH Shortchange Clinicians?', *Science*, 265/5168: 20–1.

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*, 59/8: 842–8.

Merton, R. K. (1968) 'The Matthew Effect in Science: The Reward and Communication Systems of Science are Considered', *Science*', 159/3810: 56–63.

Mulligan, A., Hall, L., and Raphael, E. (2013) 'Peer Review in a Changing World: An International Study Measuring the Attitudes of Researchers', *Journal of the American Society for Information Science*, 64: 132–61.

Murray, D. L. et al. (2016) 'Bias in Research Grant Evaluation has Dire Consequences for Small Universities', *PLoS One*, 11/6: e0155876.

Perc, M. (2014) 'The Matthew Effect in Empirical Data', *Journal of the Royal Society, Interface*, 11/98: 20140378.

Regehr, G., and Bordage, G. (2006) 'To Blind or not to Blind? What Authors and Reviewers Prefer', *Medical Education*, 40/9: 832–9.

Rossiter, M. W. (1993) 'The Matthew Matilda Effect in Science', *Social Studies of Science*, 23/2: 325–41.

Tabak, L. A., and Collins, F. S. (2011) ' Sociology. Weaving a Richer Tapestry in Biomedical Science', *Science*, 333/6045: 940–1.

Tricco, A. C. et al. (2017) 'Strategies to Prevent or Reduce Gender Bias in Peer Review of Research Grants: A Rapid Scoping Review', *PLoS One*, 6, 12/1: e0169718.

Triggle, C. R., and Triggle, D. J. (2007) 'What is the Future of Peer Review? Why is there Fraud in Science? Is Plagiarism out of Control? Why do Scientists do Bad Things? Is it all a Case of: "all that is Necessary for the Triumph of Evil is that Good Men do Nothing"?', *Vascular Health and Risk Management*, 3: 39–53.

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 of the USA* , 112/40: 12349–53.

Van Rooyen, S. et al. (1998) 'Effect of Blinding and Unmasking on the Quality of Peer Review: A Randomized Trial', *JAMA*, 280/3: 234–7.

Van Rooyen, S. et al. (1999) 'Effect of Blinding and Unmasking on the Quality of Peer Review', *Journal of General Internal Medicine*, 14/10: 622–4.

Wahls, W. P. (2016) 'Biases in Grant Proposal Success Rates, Funding Rates and Award Sizes Affect the Geographical Distribution of Funding for Biomedical Research', *PeerJ*, 4: e1917.

Wessely, S. (1998) 'Peer Review of Grant Applications: What do we Know?', *Lancet*, 352/9124: 301–5.