Attitudes of Referees in a Multidisciplinary Journal: An Empirical Analysis

Niccolò Casnici
Department of Experimental and Clinical Sciences, University of Brescia, 25125 Brescia, Italy. E-mail: niccasnici@gmail.com

Francisco Grimaldo
Departament d'Informàtica, University of Valencia, Avinguda de la Universitat, 46100 Burjassot-València, Spain. E-mail: francisco.grimaldo@uv.es

Nigel Gilbert
Centre for Research in Social Simulation, Department of Sociology, Faculty of Arts and Human Sciences, University of Surrey, Guilford GU2 7XH, United Kingdom. E-mail: N.Gilbert@surrey.ac.uk

Flaminio Squazzoni
Department of Economics and Management, University of Brescia, 25122 Brescia, Italy. E-mail: flaminio.squazzoni@unibs.it

This paper looks at 10 years of reviews in a multidisciplinary journal, *The Journal of Artificial Societies and Social Simulation (JASSS)*, which is the flagship journal of social simulation. We measured referee behavior and referees' agreement. We found that the disciplinary background and the academic status of the referee have an influence on the report time, the type of recommendation and the acceptance of the reviewing task. Referees from the humanities tend to be more generous in their recommendations than other referees, especially economists and environmental scientists. Second, we found that senior researchers are harsher in their judgments than junior researchers, and the latter accept requests to review more often and are faster in reporting. Finally, we found that articles that had been refereed and recommended for publication by a multidisciplinary set of referees were subsequently more likely to receive citations than those that had been reviewed by referees from the same discipline. Our results show that common standards of evaluation can be established even in multidisciplinary communities.

Received April 13, 2015; revised September 21, 2015; accepted September 21, 2015

 • Published online 00 Month 2016 in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/asi.23665

Introduction

Reviewing for journals is a kind of moral duty in the scientific community, being instrumental to the Mertonian ethical norms that regulate science as an organized institution (Huutoniemi 2015; Merton 1942). We know that the quality of our publications depends at least partially on comments and suggestions given by competent and cooperative referees (e.g., Mulligan, Hall, & Raphael, 2013). On the other hand, we know that science is a public good that can be maintained only if we are unbiased in judgment and collaborate in distributing efficiently and more or less equally the reviewing effort (e.g., Hochberg, Chase, Gotelli, Hastings, & Naeem, 2009).

Given that review standards are not formalized and our decisions are typically confidential, it is likely that the way we accomplish this duty may depend on our background and experience, as well as on our commitment to the journal that asked our opinion. Given the lack of training on reviewing, the opacity of the process, and the weak incentives for referees, the way we review and the time we take to accomplish this important task might depend on attitudes and norms that can reflect the attitudes of the other members of our scientific community (e.g., Azar, 2008; Squazzoni & Gandelli, 2013).

This means that looking at referee behavior could potentially help to reveal scientist misbehavior or situations where referees could benefit from their gatekeeping role at the

expense of editors and/or authors (e.g., Bornmann, Weymuth, & Daniel, 2010; Lamont, 2009; García, Rodriguez-Sánchez & Fdez-Valdivia, 2015); it could also inform us about the nature of the social norms of reviewing and so help to counterbalance possible bias (e.g., Bornmann & Daniel, 2009a; Demarest, Zhang, & Sugimoto, 2014; Lee, Sugimoto, Zhang, & Cronin, 2013; Mahoney, 1977; Sugimoto & Cronin, 2013), especially in multidisciplinary journals where heterogeneous norms coexist and could conflict (e.g., Huutoniemi, Thompson Klein, Bruun, & Hukkinen, 2010). Observing the behavior of referees is even more important for multidisciplinary journals, in which problems of incompatible standards of judgment can arise.

This paper aims to understand these problems empirically by looking at 10 years of reviews in a multidisciplinary, online journal, *The Journal of Artificial Societies and Social Simulation (JASSS)* (see: http://jasss.soc.surrey.ac.uk/JASSS.html). Established in 1998 and indexed in the Institute for Scientific Information (ISI; now Thomson Reuters), Scopus, and other major journal databases, *JASSS* is the flagship journal of social simulation, that is, the study of social processes through computer simulation. Given that it typically publishes contributions by social scientists, economists, computer scientists, and other specialists often in collaboration, who apply computer simulation to analyze a wide range of complex social processes, from opinion dynamics to market behavior, *JASSS* is ideal for looking at situations where referees from different disciplinary backgrounds remotely collaborate to judge multidisciplinary submissions (e.g., Meyer, Lorscheid, & Troitzsch 2009; Squazzoni & Casnici, 2013).

The journal received 1,272 submissions and published 606 articles and 236 book reviews from 1st January 1998 to 24th February 2015. The rejection rate for submitted articles increased as a proportion of the total article submissions from 50% in 2006 to 75% in 2015. It applies a double-blind model of peer review, has an average decision time of ~60 days from author submission to the editorial decision, and an average report time by referees of 30 days, all rounds included, for example, first, second, or more rounds of reviews. It is truly international, as about 20% of first submission authors come from the US, 13% from the UK, 10% from Germany, 9% from China, and the rest worldwide, from Japan to Australia (source: internal journal statistics).

In order to look at peer review empirically, we measured referee behavior, that is, time to report, recommendation, and length of the report, and looked at the implications of their disciplinary background, academic status, and position in the journal's editorial board. We also measured the degree of agreement about the recommendations.

We found that the disciplinary background and the academic status of the referee have an influence on the report time and the recommendation. By measuring the degree of consensus among referees and the number of citations of published articles, we found that combining referees from different disciplines was beneficial in selecting highly successful articles.

This suggests that common standards of evaluation can be developed even outside single disciplines. This is important especially in fields, such as social simulation, that have a strong, multidisciplinary nature and so require the integration of competent disciplinary judgments.

Methods

Data included 10 years of peer review in *JASSS* from 2001 to 2011. Data were extracted on 12 May 2012 from the system used by the journal to manage submissions and reviewing, epress (http://www.epress.ac.uk), with the agreement of the journal editor at the time. They included 915 submissions as well as information on 1,819 authors and referees, that is, 581 who were both authors and referees, 921 authors who had never refereed for the journal, and 317 referees who had not authored a submission. For the sake of comparability, we restricted our analysis to the first round of reviews. We excluded from the data set all submissions before 2001, that is, the first 5 years of the journal, in which some submissions were invited and rejections were rare, and all those submitted after 2011 because some of these were still under review as of 12 May 2012.

For each submission, we had: referee names and recommendations, that is, acceptance, minor or major revisions, or rejection; time taken by each referee to report; and editorial decision. For each personal record, we had: disciplinary background; status (e.g., full/associate professor, private researcher, or junior researcher); whether a member of the journal's editorial board; affiliation; and a rating of referee quality assigned by the editor. With regard to referees' disciplinary background, a classification was performed by hand from the biographical information included in the database, augmented as required using a Google search.

Most submissions were reviewed by two or three peers (in 45% and 44% of the cases, respectively) and only a few by four referees or just one (2% and 9%, respectively). Referees only rarely recommended accepting submissions at the first round (9%), mostly asking authors to revise and resubmit (40%), while rejections and minor revisions were recommended nearly equally often (25% and 26%, respectively).

In order to measure the quality of the referee reports, we considered the length of the text and the number of days spent by referees for reporting. The length was calculated by summing the number of words of the report's text, excluding all review guidelines included in the journal's review format and possible pleasantries used by the referees when reporting via e-mail (in such cases, the text was copied and pasted manually from the e-mails by the journal editor in the epress management tool). We excluded ~25% of referee reports, as they were directly evaluated by journal editors or due to missing recommendations or text. We assumed that the length of the referee text was a proxy for the quality of information provided by the referee. The length of reports is critical in two cases: (a) when referees recommend revisions, as authors expect to find suggestions to improve the quality of their submission, and (b) when referees recommended

DOI: 10.1002/asi

TABLE 1. Report length by recommendation.

Recommendation	Report length (words)	Std. Err.	[95% Conf.	Interval]
Accepted	301.0	29.2	243.7	358.4
Minor revisions	528.9	18.7	492.1	565.6
Major revisions	696.7	17.6	662.0	731.4
Rejected	534.8	19.4	496.6	572.9

rejection, as in these cases referees are expected to provide justifications for their opinions to editors and authors.

As an external validation of this measure of quality, we used the internal ratings of referees provided by the journal editors. Note that this was possible only when reports were submitted via e-mail and not directly through the epress management platform. The internal ratings were significantly correlated with the length of the reports (Kruskal–Wallis p-value $< .05$). We also included the time that the referee took to respond because providing responses quickly to authors is essential to avoid delaying publication and so is commonly perceived as an indicator of quality by authors (e.g., Hartonen & Alava, 2013).

Results

Table 1 shows report length by recommendation. Not only was length not normally distributed among the referee reports (Shapiro–Wilk W test $p < .05$); it was also dependent on the type of recommendation (Kruskal–Wallis $p < .05$). Reports where referees recommended accepting the submission or asked for minor revisions were shorter than average, and reports where referees asked for major revisions were longer (Dunn's post-hoc test[1]). This confirms that the length of the report was viewed by the referees as a means of providing detail both to editors and authors to justify their opinion and/or to share comments and suggestions to improve the quality of the authors' work. Although time for reporting was also not normally distributed (Shapiro–Wilk W test $p < .05$), a Kruskal–Wallis test showed no differences between the recommendations in terms of days spent by referees for reporting (Kruskal–Wallis $p = .23$).

Table 2 compares the 301 recommendations of referees who were members of the journal's editorial or management boards with the 1,387 recommendations from external referees. External referees rejected a higher proportion of the submissions they were sent than board members. Members of the editorial board were faster in reporting and tended to write shorter reports when recommending acceptance (11 days against 23 days, Wilcoxon $p < .05$). When they recommended rejection, they were faster than external referees (18 days against 28 days, Wilcoxon $p < .05$). When asking for major revisions, they wrote longer reports than external referees and did so more quickly (24 days against 28 days, Wilcoxon $p < .05$). A learning or commitment effect could account for these differences if members of the board were capable of detecting the quality of an author's submission more quickly than external referees or were more inclined to provide reviews of high quality to maintain the prestige of the journal.

TABLE 2. Recommendation by editorial board members versus external referees.

Recommendation	Editorial board	Management board	External referees
Accepted	10%	13%	9%
	(22)	(9)	(121)
Minor revisions	28%	34%	25%
	(63)	(25)	(352)
Major revisions	39%	34%	40%
	(90)	(25)	(559)
Rejected	23%	19%	26%
	(53)	(14)	(355)
Total	100%	100%	(100%
	(228)	(73)	(1387)

Note. The absolute values are reported in parentheses. The editorial board consisted of appointed experts and the management board included the journal, forum, and book review editors.

In order to check for possible bias due to reciprocity strategies (e.g., Squazzoni, Bravo, & Takács, 2013; Squazzoni & Gandelli, 2013), we looked at situations where referees had previously submitted to, or had published an article in the journal before being asked to review a submission. A previous negative experience as an author could have brought referees to reciprocate by rejecting a subsequent submission or a positive experience could have led referees to be more selective in order to defend the prestige of the journal in which they had already published.

Contrary to these hypotheses, we found no trace of strategic behavior by referees. We found that the recommendation was associated neither with having previously submitted (chi-square $p < .05$ but Kruskal gamma $= -0.09$) nor to being previously published (chi-square $p = .64$). The only differences were that (a) unpublished authors tended to send shorter reports in case of recommending acceptance and rejection (Wilcoxon $p < .05$) and (b) published authors tended to send longer reports when recommending rejection (Wilcoxon $p < .05$).

Then we looked at possible disciplinary-specific attitudes of referees. We found a consistent correlation between the referees' disciplinary background and their recommendations (Cramer's $V = 0.12$, $p = < .05$). Referees from the humanities tended to give more favorable evaluations than referees having other disciplinary backgrounds: 58% recommended acceptance or minor revisions compared with 23% for economists, 24% for geographers, and 28% for environmental scientists. Referees from economics were more inclined to reject submissions and, together with geographers, were in general more demanding than other referees, recommending a higher percentage of rejections and major revisions (in 76% and 77% of the cases, respectively) (Figure 1).

[1]Dunn's test (Dunn 1964) is a post-hoc pairwise multiple comparisons procedure appropriate to follow the rejection of a Kruskal–Wallis test.

DOI: 10.1002/asi

FIG. 1. Recommendations related to the disciplinary backgrounds of the referees. The area of the disks indicates the percentage of reviews for each recommendation. The total column indicates the number of reviews by reviewers from each disciplinary background.

TABLE 3. Revision time and disciplinary background.

Background	Revision time	Std. Err.	[95% Conf.	Interval]
Humanities	23.05	2.21	18.72	27.38
Social sciences	27.07	1.49	24.15	29.99
Behavioral sciences	27.57	1.93	23.78	31.36
Physics	31.26	2.99	25.40	37.12
Environmental sciences	31.85	2.66	26.63	37.07
Computer sciences/ Engineering	24.75	0.92	22.95	26.55
Math	15.50	2.32	10.95	20.05
Geography	31.39	3.26	25.01	37.78
Medicine	19.38	3.95	11.63	27.13
Economics	29.31	1.68	26.01	32.60
Management	26.24	1.93	22.45	30.02

This could be due to the development of heterogeneous standards of judgment across different disciplines (e.g., Lamont, 2009). For instance, economists have developed a common understanding of what a formal model of the economy should look like, and so could judge submissions against this benchmark, while scholars from the humanities generally do not share a common view of the foundations of their disciplines (e.g., Richiardi, Leombruni, Saam, & Sonnessa, 2006). Given the special situation of *JASSS*, which often publishes articles on economic models by noneconomists (e.g., computer scientists), this might explain the severity of judgment by economist referees.

We also found disciplinary-specific variations in revision time and report length. The average time for reporting varied significantly among referees of different disciplinary backgrounds (Kruskal–Wallis $p < .05$). While referees from math and medicine submitted their reports more promptly (respectively, 16 and 19 days on average), physicists, environmental scientists, and economists took more time to complete their reports, spending 31 days on average (Table 3; there is more detail about these differences for each pair of disciplines in Table A1 in the Appendix). This is in line with recent findings on norms of time delay in economics (e.g., Azar, 2008; Ellison, 2002). The same is true for the report length, which varied significantly between referees of different disciplinary backgrounds (Kruskal–Wallis $p < .05$; more detail about the differences for each pair of disciplines can be found in Table A2 in the Appendix).

Referees having a background in management, environmental sciences, or social sciences tended to provide longer reports than mathematicians, computer scientists, and geographers (Table 4). This could reflect different norms in the former set of disciplines, where submissions are typically longer, and so, most likely, are the length of typical reviews.

Differences were also found in the relationship between the academic status of the referees and the recommendations they made (the two variables were significantly associated;

DOI: 10.1002/asi

TABLE 4. Report length and disciplinary background.

Background	Revision length	Std. Err.	[95% Conf.	Interval]
Humanities	551.2	45.9	461.1	641.4
Social sciences	648.0	29.8	589.4	706.6
Behavioral sciences	576.6	38.7	500.6	652.5
Physics	533.8	40.5	454.3	613.2
Environmental sciences	662.3	45.2	573.6	751.0
Computer sciences/ Engineering	545.6	17.7	510.8	580.5
Math	497.0	72.8	354.0	639.9
Geography	514.7	44.2	427.9	601.5
Medicine	544.0	67.8	410.9	677.2
Economics	541.6	26.7	489.1	594.1
Management	668.1	42.7	584.2	751.9

TABLE 5. Status and type of recommendation.

Referee status	Accepted	Minor revision	Major revision	Rejected
Full/associate professors	62	193	278	194
	8.5%	26.6%	38.2%	26.7%
Junior researchers	64	200	342	194
	8.0%	25.0%	42.8%	24.3%
Private researchers	18	26	34	22
	18.0%	26.0%	34.0%	22.0%

Note. Junior researchers are PhD students, post-docs, and all those not yet having a permanent academic position. Private researchers are all referees doing research in the private and nonacademic sector, for example, researchers in business companies, consultants, or research-based entrepreneurs.

chi-square 14.8, $p = .022$). Academic researchers were less likely to recommend acceptance than nonacademic researchers. While referees doing research in the private sector recommended acceptance of 18% of submissions, junior academic researchers and professors recommended acceptance in only the 8% of the cases. While junior academic researchers were more inclined towards recommending major revisions, full/associate professors were more severe and tended to reject more submissions than their colleagues. There is a general pattern that links seniority and selectivity, irrespective of the disciplinary context of the referees (Table 5).

Differences were also found in the lengths of reports, which were strongly correlated with the referee's status (Kruskal–Wallis $p < .05$). Junior researchers tended to write longer reports. In this respect, there is a robust statistical difference between professors and junior researchers in terms of report length (Dunn's post-hoc $p < .05$), as well as between private and junior researchers (Dunn's post-hoc $p < .05$), whereas the difference between private researchers and professors was not statistically significant. The standard error of the length of reports by private researchers was higher due to the heterogeneity of this category, which includes all researchers not performing research in an academic institute or a public research center, for example, researchers in private companies or foundations, consultants, and research-based entrepreneurs (Table 6).

TABLE 6. Status and report length.

Referee status	Mean	Std. Err.	[95%	Conf. Interval]
Full/associate professors	534.20	14.76	505.25	563.15
Junior researchers	622.98	16.17	591.27	654.70
Private researchers	532.57	36.23	461.50	603.64

TABLE 7. Referee status and revision time.

Referee status	Mean	Std. Err.	[95%	Conf. Interval]
Full/associate professors	**28.21**	0.94	26.37	30.05
Junior researchers	**25.23**	0.74	23.78	26.69
Private researchers	**27.63**	2.09	23.53	31.73

TABLE 8. Decisions on reviewing requests by status ("refusals" includes those who were asked but did not reply).

Status	Number of requests	Number of refusals	% Refusals
Professors	1,178	145	12.3
Junior researchers	1,222	112	9.7
Private researchers	162	19	11.7

Another status-influenced difference was found in the time taken to report. While academic professors took an average of 28 days to report and private researchers took 27 days, juniors took 25 days. However, only the difference between professors and junior researchers was statistically significant (Dunn's post-hoc $p < .05$). Combined with the previous finding, this means that junior researchers tended to complete their reports more quickly and with more content than more senior colleagues. This may be because junior researchers are motivated to take the reviewing task more seriously both as a means for learning and for building a reputation with the journal's editor for future submissions (Table 7).

Junior researchers tended to refuse requests to review less frequently than academic professors and private researchers, although they had more requests (Table 8).

Furthermore, there is a significant effect of the disciplinary background of the referees on agreeing to review (Table 9). More specifically, refusals were less frequent when referees were experts in medicine, computer science, or humanities, whereas they were more frequent for economists and physicists.

Finally, we measured the degree of alignment between referees who were assigned to the same submission. We assigned a number to recommendations: accept (1), minor revision (2), major revision (3), and reject (4), and calculated the standard deviation of the referees' recommendations. The principle was that the higher the standard deviation between the numbers of the n recommendations ($n =$ the

DOI: 10.1002/asi

TABLE 9. Decisions on reviewing requests by disciplinary background ("refusals" includes those who were asked but did not reply).

Disciplinary background of the referees	Number of requests	Number of refusals	% Refusals
Humanities	152	12	7.9
Social sciences	430	46	10.7
Behavioral sciences	199	26	13.1
Physics	162	26	16.1
Environmental sciences	122	15	12.3
Computer sciences/ engineering	742	53	7.1
Math	42	4	9.5
Geography	76	8	10.5
Medicine	35	1	2.9
Economics	400	64	16.0
Management	199	21	10.6

number of referees who were assigned to the same submission), the more the reviews were misaligned. We excluded all submissions assigned to only one referee.

Table 10 shows the disciplinary composition of the referee pools asked to evaluate the journal submissions. Eighty-two percent of the referee pools included referees coming from at least two different disciplinary backgrounds; 28% had referees from three different disciplines.

The number of referees with different disciplinary backgrounds assigned to the same submission had no impact on the probability of recommendations being aligned (Kruskal–Wallis $p = .58$). Similarly, having a group of referees with different status who were asked to evaluate the same submission had no impact on the consensus between referees (Kruskal–Wallis $p = .29$). This suggests that, by involving referees with different degrees of seniority and sector of specialization, a submission could benefit from more informative reports and quicker response times by junior researchers while at the same time taking advantage of the more learned judgment by senior researchers (e.g., see Tables 6 and 7).

The fact that a consistent degree of consensus was reached between referees on the microscale of the single submission, independently of their heterogeneous disciplinary backgrounds, would contradict recent findings on disagreement among referees in peer review (e.g., Bornmann & Daniel, 2009a, 2009b; Kravitz et al., 2010; Lee, 2012). The journal editor tended to accept or reject submissions only when the disagreement between referees was low. When referees disagreed about their recommendations, the editors usually opted for a minor or major revision and assigned a second round of reviews to the submission (Kruskal–Wallis $p < .05$; there is more detail on the post-hoc analysis in Table A3 in the Appendix). Therefore, the combination of referees of different disciplinary backgrounds asked to evaluate the same submission probably reduced disciplinary bias (e.g., Chubin & Hackett, 1990; Lee, 2012, Grimaldo & Paolucci, 2013).

By determining the number of citations received by accepted articles, as recorded by Google Scholar, we found

TABLE 10. Disciplinary composition of reviews.

Disciplinary composition	Number	Percentage
Mono-disciplinary referee pool	112	18.1
Multidisciplinary referee pool	506	81.8
Referees pool with two disciplinary backgrounds	362	71.5
Referees pool with three disciplinary backgrounds	141	27.8
Referees with four disciplinary backgrounds	3	0.5

TABLE 11. Relationship between number of citations received by published articles and mono versus multidisciplinarity of the referees.

Nature of the reviewing	Number	Mean citations	Std. Err.	[95%	Conf. Interval]
Mono-disciplinary	**129**	**18.46**	2.21	14.08	22.84
Multidisciplinary	**48**	**29.43**	6.49	16.36	42.50

Note. Mono-disciplinary reviewing refers to submissions being reviewed by two or more referees from the same discipline, while multidisciplinary reviewing refers to submissions being reviewed by two or more referees from different disciplines (data source: Google Scholar).

that articles that have been reviewed by a multidisciplinary referee group had more success in terms of citations. Not only was the multidisciplinary nature of the referee group helpful for editor's judgment, it also probably generated important knowledge that improved the quality of submissions (Table 11).

Discussion and Conclusions

These findings help to cast new light on old matters concerning the quality of peer review, the evaluation of multidisciplinary work, and the typical behavior of referees. Although limited to one case study, they help us to reconsider certain bad signals about the quality of peer review that have come from recent scandals where misbehavior by authors was combined with unreliability by referees (e.g., Alberts, Hanson, & Kelner 2008; Couzin, 2006; Crocker & Cooper, 2011).

This is important especially if the multiple functions of peer review are considered. Peer review is a means for selecting scientific work through criteria of excellence and avoiding publishing submissions that undermine the prestige and standards of a journal. However, it is also a way to improve the quality of scientific work through anonymous, decentralized collaboration (e.g., Jeggins, 2006). In our case, it is evident that these two functions, which often might create ambiguity in reviewers' interpretations of their task, harmonize positively.

First, the findings show that common standards of evaluation can be developed even in multidisciplinary journals that call upon expert referees from a variety of fields who might have different standards of judgment. Although there is evidence of disagreement among the referees, we found

DOI: 10.1002/asi

that the journal editor's tendency to match submissions with a diverse set of referees, for both disciplinary background and seniority, was instrumental in evaluating multidisciplinary submissions fairly. This would confirm the importance of carefully selecting referees in peer review and ensuring a diversity of criteria of judgment and opinions (e.g., Ferreira et al., 2015).

Furthermore, citation analysis confirmed that being subject to multidisciplinary reviews can be beneficial for the success of multidisciplinary articles. Obviously, the positive effect of diversity on judgment and impact of submissions could have been due to the multidisciplinary nature of submissions more than to the quality and fairness of the peer-review process they had been exposed to (e.g., Levitt & Thelwall, 2008). However, it is important that a journal's evaluation process reflects and magnifies, rather than suppresses multidisciplinary research (e.g., Huutoniemi, 2015).

Second, previous studies suggested the importance of motivations of reciprocity in explaining referee behavior. Reciprocity may have a bright or a dark side. For instance, in a laboratory experiment, Squazzoni, Bravo, and Takács (2013) found that reciprocity can cause referees to behave fairly and keep evaluation standards high, in the hope of future benefits from the quality of the process when they become authors. On the other hand, Squazzoni and Gandelli (2013) found that reciprocity motives by referees can be beneficial only when referees consider the quality of peer review they have been exposed to more than the fact that they have been previously published or rejected.

However, in this study we found no trace of such strategic behavior by referees. Previous positive or negative experiences as authors did not help to predict subsequent referee behavior and so did not affect the recommendations and time to report. However, the specificity of *JASSS*, which includes scientists who have different backgrounds but share a common approach and method of analysis, that is, agent-based models of social interaction, and which is targeted to a relatively small and so probably cohesive scientific community must be borne in mind. Moreover, not only do researchers from social simulation constitute a relatively small scientific community; the vivid associational life that characterizes the community, with three large associations, that is, the European Social Simulation Association, the Computational Social Science Society of the Americas, and the Pacific-Asian Association for Agent-based Approach in Social Systems Sciences, each quite active in organizing events and creating a cohesive community worldwide, could explain the cooperative tendency and the lack of misbehavior by referees. This could indicate that features of the organization of the scientific community could have a significant effect on scientists' behavior in peer review and so have important implications for the quality of the process. On the other hand, it must be said that all empirical analyses of peer review in scholarly journals are context-dependent, due to the complexity and variety of scientific research and the lack of large-scale, systematic, or comparative studies that could help with identifying general trends (Siler, Lee, & Bero, 2015).

Finally, it is important to note that empirical analysis of peer review is still in its infancy. This is due to some resistance to scrutiny, accountability, and openness by many of the stakeholders involved, for example, funding agencies, publishers, and journal editors (e.g., Couzin-Frankel, 2013). Establishing the sharing of data on peer review at a large scale is the only means to make systematic analysis of this important institution possible. While improvements in transparency and ethical standards are important to reduce potential bias and increase the accountability and legitimacy of peer review (e.g., Bosch, Hernandex, Pericas, Doti, & Marusic, 2012; Wager, 2010; Wager, Fiack, Graf, Robinson, & Rowlands, 2009), the degree of openness of journals towards sharing internal data and performing systematic analyses of their internal procedures should become a required element to certify their quality. This would stimulate scientists to look at journals not only through the lens of their impact factor, which is often a misleading guide to judging journal quality, but also their contributions to preserving the normative foundations of science as an open, transparent, and civilizing system in a world of increasing competitive pressures.

Acknowledgment

This publication is supported by the COST Action TD1306 "New frontiers of peer review" (www.peere.org).

References

Alberts, B., Hanson B., & Kelner K. L. (2008). Reviewing peer review. Science, 321, 15.

Azar, O. H. (2008). Evolution of social norms with heterogeneous preferences: A general model and an application to the academic review process. Journal of Economic Behavior and Organization, 65, 420–435.

Bornmann, L., & Daniel, H.-D. (2009a). The luck of the referee draw: The effect of exchanging reviews. Learned Publishing, 22(2), 117–125.

Bornmann, L., & Daniel, H.-D. (2009b). Reviewer and editor biases in journal peer review: An investigation of manuscript refereeing at Angewandte Chemie International Edition. Research Evaluation, 18(4), 262–272.

Bornmann, L., Wheymuth, C., & Daniel, H.-D. (2010). A content analysis of referees' comments: How do comments on manuscripts rejected by a high-impact journal and later published in either a low-or high-impact journal differ? Scientometrics, 83, 493–506.

Bosch, X., Hernandex, C., Pericas, J. M., Doti, P., & Marusic, A. (2012). Misconduct policies in high-impact biomedical journals, PLoS One, 7, e51928.

Chubin, D. E., & Hackett, E. J. (1990). Peerless science: Peer review and U.S. science policy. Stony Brook, NY: State University of New York Press.

Couzin, J. (2006) . . . . and how the problems eluded peer reviewers and editors. Science, 311, 614–615.

Couzin-Frankel, J. (2013). Secretive and subjective, peer review proves resistant to study. Science, 314(6152), 1331.

Crocker, J., & Cooper, M. L. (2011). Addressing scientific fraud. Science, 334(6060), 1182.

DOI: 10.1002/asi

Demarest, B., Zhang, G., & Sugimoto, C. R. (2014). The reviewer in the mirror. Examining gendered and ethnicized notions of reciprocity in peer review. Scientometrics, 101(1), 717–735.
Dunn, O. J. (1964). Multiple comparisons using rank sums. Technometrics, 6(3), 241–252.
Ellison, G. (2002). The slowdown of the economics publishing process. Journal of Political Economy, 110(5), 947–993.
Ferreira, C., Bastille-Rousseau, G., Bennett, A. M., Hance Ellington, E., Terwissen, C., Austin, C., Borlestean, A., Boudreau, M. R., Chan, K., Forsythe, A., Hossie, T. J., Landolt, K., Longhi, J., Otis, J.-A., Peers, M. J. L., Rae, J., Seguin, J., Watt, C., Wehtje, M., & Murray, D. L. (2015). The evolution of peer review as a basis for scientific publication: Directional selection towards a robust discipline? Biological Reviews, DOI: 10.1111/brv.12185.
García, J. A., Rodriguez-Sánchez, R., & Fdez-Valdivia, J. (2015). The principal-agent problem in peer review. Journal of the American Society for Information Science and Technology, 66(2), 297–308.
Grimaldo, F., & Paolucci, M. (2013). A simulation of disagreement for control of rational cheating in peer review. Advances in Complex Systems, 16(7), 1350004.
Hartonen, T., & Alava, M. J. (2013). How important tasks are performed: Peer review. Scientific Reports, 3, 1679, doi:10.1038/srep01679
Hochberg, M. E., Chase, J. M., Gotelli, N. J., Hastings, A., & Naeem, S. (2009). The tragedy of the reviewer commons. Ecology Letters, 12, 2–4.
Huutoniemi, K. (2015). Peer review: Organized skepticism. In Wright, J. D. (Ed.), International encyclopedia of the social and behavioral sciences (2nd ed.). Elsevier, 685–689.
Katri Huutoniemi, K. (2015). Interdisciplinarity as academic accountability: Prospects for quality control across disciplinary boundaries. Social Epistemology. DOI: 10.1080/02691729.2015.1015061
Huutoniemi, K., Thompson Klein, J., Bruun, H., & Hukkinen. J. (2010). Analyzing interdisciplinarity: Typology and indicators. Research Policy, 39(1), 79–88.
Jeggins, C. G. (2006). Quality and value: The true purpose of peer review. Nature, DOI: 10.1038/nature05032.
Kravitz, R. L., Franks, P., Feldman, M. D., Gerrity, M., Byrne, C., & Tierney, W. M. (2010). Editorial peer reviewers' recommendations at a general medical journal: Are they reliable and do editors care? PLoS One, 5(4), e10072.
Lamont, M. (2009). How professors think. Inside the curious world of academic judgment. New York: Harvard University Press.
Lee, C. J. (2012). A Kuhnian critique of psychometric research on peer review. Philosophy of Science, 79, 859–870.
Lee, C. J., Sugimoto, C. R., Zhang, G., & Cronin, B. (2013). Bias in peer review. Journal of the American Society for Information Science and Technology, 64(1), 2–17.
Levitt, J. M., & Thelwall, M. (2008). Is multidisciplinary research more highly cited? A macrolevel study. Journal of the American Society for Information Science and Technology, 59(12), 1973–1984
Mahoney, M. J. (1977). Publication prejudices: An experimental study of confirmatory bias in the peer review system. Cognitive Therapy and Research, 1(2), 161–175.
Merton, R. K. (1942). The normative structure of science. Journal of Legal and Political Sociology, 1, 115–126 (Reprinted in *The Sociology of Science: Theoretical and Empirical Investigations*, Chicago, University of Chicago Press, 1973, 267-278).
Meyer, M., Lorscheid, I., & Troitzsch, K. G. (2009). The development of social simulation as reflected in the first ten years of *JASSS*: A citation and co-citation analysis. Journal of Artificial Societies and Social Simulation, 12(4), 12 <http://jasss.soc.surrey.ac.uk/12/4/12.html>.
Mulligan, A., Hall, L., & 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 and Technology, 64(1), 132–161.
Richiardi, M., Leombruni, R., Saam, N., & Sonnessa, M. (2006). A common protocol for agent-based social simulation. Journal of Artificial Societies and Social Simulation, 9(1), 15. <http://jasss.soc.surrey.ac.uk/9/1/15.html>.
Siler, K., Lee K., & Bero, L. (2015). Measuring the effectiveness of scientific gatekeeping. PNAS, 112(2), 360–365.
Squazzoni, F., & Casnici, N. (2013). Is social simulation a social science outstation? A Bibliometric analysis of the impact of JASSS. Journal of Artificial Societies and Social Simulation, 16(1), 10 <http://jasss.soc.surrey.ac.uk/16/1/10.html>.
Squazzoni, F., & Gandelli, C. (2013). Opening the black-box of peer review: An agent-based model of scientist behavior. Journal of Artificial Societies and Social Simulation, 16(2), 3 <http://jasss.soc.surrey.ac.uk/16/2/3.html>
Squazzoni, F., Bravo, G., & Takács, K. (2013). Does incentive provision increase the quality of peer review? An experimental study. Research Policy, 42(1), 287–294.
Sugimoto, C. R., & Cronin, B. (2013). Citation gamesmanship: Testing for evidence of ego bias in peer review. Scientometrics, 95(3), 851–862.
Wager, E. (2010). Editorial code of conduct. The Lancet, 379, 9814.
Wager, E., Fiack, S., Graf, C., Robinson, A., & Rowlands, I. (2009). Science journal editors' view on publication ethics: Results of an international survey. Journal of Medical Ethics, 35(6), 348–353.

Appendix

TABLE A1. Dunn's post-hoc analysis on report time and referee background.

	1	2	3	4	5	6	7	8	9	10
2	**−1.95**	**−2.39**	−0.87	**−2.93**	−1.59	−0.68	**−3.60**	**−2.44**	−1.49	−0.82
3	−2.39	−0.87	**−2.93**	−1.59	−0.68	**−3.60**	**−2.44**	−1.49	−0.82	−1.52
4	−2.93	−1.59	−0.68	**−3.60**	**−2.44**	−1.49	−0.82	−1.52	0.81	1.54
5	−3.60	**−2.44**	−1.49	−0.82	−1.52	0.81	1.54	**2.26**	**3.10**	1.52
6	−1.52	0.81	1.54	**2.26**	**3.10**	1.52	**2.84**	**3.14**	**3.52**	**4.01**
7	1.52	**2.84**	**3.14**	**3.52**	**4.01**	**2.58**	**−2.89**	**−1.79**	−1.09	−0.54
8	−2.89	**−1.79**	−1.09	−0.54	0.15	**−2.26**	**−3.55**	0.81	**1.99**	**2.33**
9	0.81	**1.99**	**2.33**	**2.70**	**3.19**	**1.72**	−0.52	**2.83**	**−2.72**	−1.07
10	−2.72	−1.07	0.00	0.77	**1.67**	**−2.00**	**−3.31**	1.18	**−2.44**	**−1.86**
11	−1.86	−0.14	0.65	1.32	**2.12**	−0.80	**−2.79**	1.60	−1.98	0.76

Note. Headers indicate the referee background: 1 = humanities, 2 = social sciences, 3 = behavioral sciences, 4 = physics, 5 = environmental sciences, 6 = computer sciences/engineering, 7 = mathematics, 8 = geography, 9 = medicine, 10 = economics, 11 = management.

TABLE A2. Dunn's post-hoc analysis on report length and referee background.

	1	2	3	4	5	6	7	8	9	10
2	**−2.15**	−0.85	1.24	−0.03	**2.09**	0.81	**−2.49**	−0.96	**−1.79**	**−2.45**
3	−0.85	1.24	−0.03	**2.09**	0.81	**−2.49**	−0.96	**−1.79**	**−2.45**	−0.35
4	−0.03	**2.09**	0.81	**−2.49**	−0.96	**−1.79**	**−2.45**	−0.35	**2.82**	0.76
5	**−2.49**	−0.96	**−1.79**	**−2.45**	−0.35	**2.82**	0.76	−0.30	**2.79**	0.61
6	−0.35	**2.82**	0.76	−0.30	**2.79**	0.61	**1.97**	1.19	0.63	**2.34**
7	0.61	**1.97**	1.19	0.63	**2.34**	0.88	−0.45	1.09	0.21	−0.42
8	−0.45	1.09	0.21	−0.42	1.60	−0.27	−0.88	−0.75	0.43	−0.23
9	−0.75	0.43	−0.23	−0.73	0.93	−0.63	−1.09	−0.35	−0.33	**2.39**
10	−0.33	**2.39**	0.68	−0.28	**2.59**	−0.01	−0.85	0.25	0.61	**−2.56**
11	**−2.56**	−0.79	**−1.76**	**−2.50**	0.28	**−3.11**	**−2.30**	−1.51	−0.80	**−2.78**

TABLE A3. Dunn's post-hoc analysis on average agreement between referees for the type of recommendation.

	Acceptance	Minor revision	Major revision
Minor review	−2.15		
	0.02		
Major review	−1.69	1.10	
	0.05	0.14	
Rejected	−0.90	2.33	1.54
	0.18	0.01	0.06

DOI: 10.1002/asi