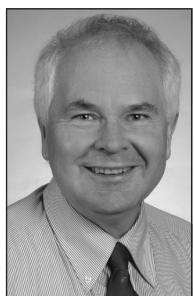

Reliability of reviewers' ratings when using public peer review: a case study

L. BORNMANN and H.-D. DANIEL
ETH Zurich

ABSTRACT. *If a manuscript meets scientific standards and contributes to the advancement of science, it can be expected that two or more reviewers will agree on its value. Manuscripts are rated reliably when there is a high level of agreement between independent reviewers. This study investigates for the first time whether inter-rater reliability, which is low with the traditional model of closed peer review, is also low with the new system of public peer review or whether higher coefficients can be found for public peer review. To investigate this question we examined the peer-review process practiced by the interactive open access journal Atmospheric Chemistry and Physics (based on 465 manuscripts submitted between 2004 and 2006 receiving 1,058 reviews in total). The results of the study show that inter-rater reliability is low (kappa coefficient) or reasonable (Intraclass Correlation Coefficient) in public peer review.*



L. Bornmann



H.-D. Daniel

© L. Bornmann and H.-D. Daniel 2010

Introduction

Print journals use the traditional system of closed peer review. Using modern information technology, in particular the Internet, numerous interactive open access journals have now become established in science that work either with the traditional peer review system or with the 'new' system of public peer review.^{1,2} In April 2009, the Directory of Open Access Journals (<http://www.doaj.org/>) listed 4,054 open access journals. Compared to the traditional system, the new system of peer review in an electronic environment is seen to have the following advantages, among others: (i) submitted manuscripts are immediately published as 'discussion papers' on the journal's website; (ii) reviewers' comments on the quality of the content of the manuscript and authors' replies to the reviewers' critical comments are publicly exchanged; and (iii) reviewers' arguments are publicly heard, and, if comments are openly signed, reviewers can also claim authorship for their contributions.³

The disadvantages of public peer review in interactive open access journals (or, in other words, the advantages of the traditional system of closed peer review used by print journals) are: (i) researchers' rather low acceptance of open access journals up to now;⁴ and (ii) the possible 'accumulation of "enemies" who may later try to torpedo one's own manuscripts or grant applications'.⁵ However, the strongest reservation about open access journals is doubt as to whether they achieve sufficient quality control.^{6,7} According to Weller,⁸ 'no studies have been conducted yet [i.e. the year 2002] that have comprehensively investigated models of peer review in an electronic environment' (p. 303). That is still the case today.

If a manuscript meets scientific standards and contributes to the advancement of sci-

ence, it can be expected that two or more reviewers will agree on its value. Manuscripts are rated reliably when there is a high level of agreement between independent reviewers. Cicchetti⁹ defines inter-rater reliability 'as the extent to which two or more independent reviews of the same scientific document agree' (p. 120). Even though a high level of inter-rater reliability is often seen as desirable, the results of a meta-analysis for the inter-rater reliability of peer reviews for traditional print journals¹⁰ confirm the findings of the narrative reviews published so far:^{11,12} a low level of inter-rater reliability. (However, the study found also a statistically significant study-to-study variation of the inter-rater reliability coefficients.) In one of the most comprehensive single primary studies on this issue, Bornmann and Daniel¹³ determined for the journal *Angewandte Chemie International Edition*, for example, that 'referees show agreement in their responses on 10–21% more manuscripts than would have been expected by chance' (p. 7174).

This study investigates for the first time whether inter-rater reliability can be assumed to be low also with the new system of public peer review, or whether higher coefficients can be found. Public peer review is supposed to bring a new openness to the reviewing process that will enhance its accuracy and fairness.^{14,15} Although a Cochrane review¹⁶ found no evidence of reviewer and/or author concealment on outcome of quality assessment, no evidence of the value of reviewer training and no effect on quality of different methods of communicating with reviewers and means of dissemination (see here also van Rooyen et al.¹⁷), the most comprehensive review of research on traditional journal peer review⁸ comes to the conclusion that 'it has been shown that "bias" on the part of reviewers is part of any review process' (p. 308). Publishing reviews is supposed to lead to reviewers using argumentation and judging solely on the basis of scientific criteria. If the reviewer's ratings are not to be influenced by potential sources of bias (such as the author's gender or institutional affiliation, which can endanger accuracy and fairness), a higher level of agreement between reviewers is to be

expected. We investigated the extent to which this expectation can be confirmed, taking the example of the public peer-review process practiced by the interactive open access journal *Atmospheric Chemistry and Physics* (ACP).

Methods

Manuscript review at ACP

ACP was launched in September 2001. It is produced and published by the European Geosciences Union (EGU; www.copernicus.org/EGU/EGU.html) and the Copernicus Society (www.copernicus.org). ACP is freely accessible via the Internet (www.atmoschem-phys.org). It has the second highest annual Journal Impact Factor (JIF) (provided by Thomson Reuters, Philadelphia, PA) in the category 'Meteorology & Atmospheric Sciences' (at 4.927 in the 2008 Journal Citation Reports, Science Edition). ACP has a two-stage publication process, with a 'new' peer-review process consisting of public peer review and interactive discussion^{3,18} that is described on the ACP website as follows: in the first stage, manuscripts that pass a rapid pre-screening process (access review) are immediately published as 'discussion papers' on the journal's website. These discussion papers are then made available for 'interactive public discussion', during which the comments of designated reviewers (usually, reviewers that already conducted the access review), additional comments by other interested members of the scientific community, and the authors' replies are published alongside the discussion paper.

During the discussion phase, the designated reviewers are asked to answer to the following questions according to the ACP's principal evaluation criteria (see http://www.atmospheric-chemistry-and-physics.net/review/ms_evaluation_criteria.html, from which the following information is taken): (i) scientific significance ('Does the manuscript represent a substantial contribution to scientific progress within the scope of ACP (substantial new concepts, ideas, methods, or data?'); (ii) scientific quality ('Are the scientific approach and applied methods valid? Are the results discussed in an appro-

public peer review is supposed to bring a new openness to the reviewing process that will enhance its accuracy and fairness

*for the
investigation of
peer review we
had data
for 1,111
manuscripts*

appropriate and balanced way (consideration of related work, including appropriate references?); and (iii) presentation quality ('Are the scientific results and conclusions presented in a clear, concise, and well-structured way (number and quality of figures/tables, appropriate use of English language)?'). The response categories for the three questions are: (1) excellent, (2) good, (3) fair, and (4) poor. In addition to the principal evaluation criteria the reviewers are asked to give a final publication recommendation: 'Do you recommend acceptance of the manuscript?' Here, the response categories are: (1) yes, without alterations, (2) yes, after minor alterations, (3) yes, after major alterations, (4) no. In addition to giving the formal ratings to the four questions, the reviewers have the opportunity to write a commentary (nearly all manuscripts receive commentaries).

The ratings are submitted in parallel to the commentaries, but they are not open because they are meant to support the editorial decision rather than the scientific discussion. This policy has been introduced in 2001. According to the experiences and the philosophy of ACP's chief-executive editor Ulrich Pöschl, prescribed publication of formal ratings is likely to do more harm than good (e.g. initiation/escalation of unnecessary controversies). Most other journals pursuing public peer review do not prescribe publication of formal ratings either and some of them explicitly instruct reviewers not to include formal ratings in their public comments (see, e.g., <http://adv-model-earth-syst.org/index.php/JAMES/about/faq>). At ACP, the editors leave it up to the reviewers if they want to include ratings in their public comments, and sometimes they do (~30%). With increasing acceptance and spread of public review it may become beneficial and appropriate to prescribe publication of formal ratings. For now, however, the ACP editors prefer a mix of open commentaries and non-public ratings for the discussion phase.

After the end of the discussion phase every author has the opportunity to submit a revised manuscript taking into account the reviewers' comments and the comments of interested members of the scientific commu-

nity. Based on the revised manuscript and in view of the access peer review and interactive public discussion, the editor accepts or rejects the revised manuscript for publication in ACP. For this decision, further external reviewers may be asked to review the revision, if needed.

Database for the present study

For the investigation of peer review at ACP we had data for 1,111 manuscripts that went through the complete ACP selection process in the years 2001–2006. Of the 1,111 manuscripts, 1,032 (93%) manuscripts were published as discussion papers; 79 (7%) were rejected during access review for publication as discussion papers. Reviewers' ratings on the evaluation criteria and reviewers' publication recommendations were available for 552 (55%) of the 1,008 manuscripts. This reduction in number is due to the fact that the ratings have been stored electronically by the publisher only since 2004. Of the 552 manuscripts, 16% (n = 87) have one review, 64% (n = 356) have two, 17% (n = 92) have three, 3% (n = 15) have four, and two manuscripts have five independent reviews. Since to check the inter-rater reliability a manuscript must have at least two reviews, 465 manuscripts with a total of 1,058 reviews could be included in the analyses of this study.

Statistical procedures

In this study we used two approaches to the statistical analysis of inter-rater reliability: kappa (unweighted, weighted, conditional) and Intraclass Correlation Coefficient (random-effects ANOVA model, random-intercept proportional odds model).

According to von Eye and Mun,¹⁹ kappa (κ) is 'one of the most widely employed coefficients in the social sciences' to determine inter-rater reliability (p. 1). If the raters are in complete agreement, then $\kappa = 1$; if κ is near 0, the observed level of agreement is not higher than a chance level: 'Multiplied by 100, κ indicates the percentage by which two raters' agreement exceeds the agreement that could be expected from chance' (p. 5). In contrast to unweighted κ , weighted κ additionally takes into account that where

there is a lack of agreement between the ratings of two raters, there can be different degrees of disagreement. In the analysis, a weight of 0.6667 was assigned to those manuscripts where the reviewers show 'two-thirds' agreement (i.e. the reviewers' chose neighboring response categories, such as 'excellent' and 'good'). A weight of 0.3333 was assigned in the case of one-third agreement (e.g. 'excellent' and 'fair'). In the analysis, a weight of 0 (i.e. no weight) was used when the reviewers' ratings were completely contrary, and a weight of 1 was used when the reviewers' ratings agreed completely. κ was calculated by using the Stata,²⁰ kappa,²¹ and kapci²² syntaxes.

In addition to the calculation of κ as a summary statement for the entire agreement table, it is also possible to calculate the inter-rater reliability for individual rating categories. Von Eye and Mun¹⁹ call this coefficient conditional κ (section 1.1.2). In the present study we used conditional κ to test whether the inter-rater reliabilities differ for positively and negatively reviewed manuscripts. Weller's⁸ overview of studies on the reliability of journal peer review shows that

on the average, reviewers are twice as likely to agree on rejection than on acceptance. An average of 44.9 percent of the reviewers agree when they make a rejection recommendation while an average of 22.0 percent agree when they make an acceptance recommendation. (p. 193)

The Intraclass Correlation Coefficient (ICC)

Whereas κ compares the observed proportion of agreement with the expected proportion, in the context of inter-rater reliability a correlation coefficient, according to von Eye and Mun,¹⁹ 'can be interpreted as the expected agreement in relative standings of objects on a measure of random cases between two sources' (p. 116). ICC 'is a variance decomposition method to assess the portion of overall variance attributable to between-subject variability' (p. 116). A one-way random ICC is appropriate for a data set, if the ordering of raters is irrelevant and the raters are different for different manuscripts. That means only one systematic source of variation exists: the manu-

scripts. ICC will approach 1.0 when there is no variance within manuscripts, indicating that total variation in measurements on the rating scale is due solely to the manuscript (the between-manuscripts effect is very large relative to the within-manuscripts effect). ICC is 0 when within-manuscripts variance equals between-manuscripts variance, indicative of the manuscript variable having no effect. Multiplied by 100, the ICC shows the ratio of the variability in reviewers' ratings occurring between manuscripts, with the remaining rate (1 - ICC) occurring within manuscripts.

The random-effects ANOVA model is the traditional way to calculate the ICC. For that we used the Stata loneway syntax.²¹ However, the ICC also finds use in multi-level regression analysis.²³ Unlike the random-effects ANOVA model, here the assumption can be relaxed that the dependent variable has an interval level and is normally distributed.²¹ In the case of a categorical variable with ordered categories (here: the reviewers' ratings) a random-intercept proportional odds model can be used.²⁴ With the estimated random-intercept variance, the unconditional intraclass correlation can be calculated. For calculation of the random-intercept proportional odds models, we used the gllamm syntax of Rabe-Hesketh et al.²⁵ developed for Stata.

Results

Table 1 presents the coefficients (κ and ICC) of the reviewers' ratings on three evaluation criteria and the final publication recommendation for the manuscripts published as discussion papers in ACP.

Results with regard to κ as coefficient for the inter-rater reliability

With regard to κ as coefficient for the inter-rater reliability, Table 1 shows unweighted and weighted κ as well as conditional κ for the partial reviewer agreement. Two guidelines for the interpretation of κ have been published to date. According to Landis and Koch,²⁶ κ should be interpreted as follows:

$\kappa < 0.00$	poor agreement
$0.00 \leq \kappa \leq 0.20$	slight

only one systematic source of variation exists: the manuscripts

Table 1. Reliability of the reviewers' ratings on three evaluation criteria and of the reviewers' final publication recommendations

Evaluation criteria and final publication recommendation	Observed agreement (%) for two reviewers (n = 356 manuscripts)	Expected agreement (%) for two reviewers (n = 356 manuscripts)	No. of reviewers (no. of manuscripts)			
			2 (n = 356)	3 (n = 92)	4 (n = 15)	2–5 (n = 465)
Scientific significance: Does the manuscript represent a substantial contribution to scientific progress within the scope of ACP (substantial new concepts, ideas, methods, or data)? ^a						
Unweighted κ	55.1	42.4	0.22	0.16	0.07	0.19
Weighted κ	83.0	76.7	0.27			
Conditional κ ('excellent' and 'good') ^b			0.28			
Conditional κ ('fair' and 'poor') ^b			0.36			
ICC (random-effects ANOVA model)			0.33	0.32	0.47	0.34
Unconditional ICC (random-intercept proportional odds model)			0.41	0.39	0.44	0.41
Scientific quality: Are the scientific approach and applied methods valid? Are the results discussed in an appropriate and balanced way (consideration of related work, including appropriate references)? ^a						
Unweighted κ	46.1	37.5	0.14	0.08	-0.01	0.12
Weighted κ	78.8	73.6	0.20			
Conditional κ ('excellent' and 'good') ^b			0.24			
Conditional κ ('fair' and 'poor') ^b			0.20			
ICC (random-effects ANOVA model)			0.27	0.29	0.06	0.26
Unconditional ICC (random-intercept proportional odds model)			0.32	0.32	0.01	0.30
Presentation quality: Are the scientific results and conclusions presented in a clear, concise, and well-structured way (number and quality of figures/tables, appropriate use of English language)? ^a						
Unweighted κ	49.2	39.7	0.16	0.18	0.03	0.15
Weighted κ	81.4	75.5	0.24			
Conditional κ ('excellent' and 'good') ^b			0.29			
Conditional κ ('fair' and 'poor') ^b			0.22			
ICC (random-effects ANOVA model)			0.34	0.32	0.07	0.31
Unconditional ICC (random-intercept proportional odds model)			0.41	0.36	0.06	0.36
Publication recommendation: Do you recommend acceptance of the manuscript? ^c						
Unweighted κ	58.7	48.4	0.20	0.20	0.25	0.20
Weighted κ	85.7	80.4	0.22			
Conditional κ ('yes, without alterations' and 'yes, after minor alterations') ^b			0.29			
Conditional κ ('yes, after major alterations' and 'no') ^b			0.25			
ICC (random-effects ANOVA model)			0.24	0.31	0.44	0.26
Unconditional ICC (random-intercept proportional odds model)			0.31	0.37	0.51	0.33

^aResponse categories: (1) 'excellent'; (2) 'good'; (3) 'fair'; (4) 'poor'.

^bTo calculate conditional κ , the two categories were combined in one category.

^cResponse categories: (1) 'yes, without alterations'; (2) 'yes, after minor alterations'; (3) 'yes, after major alterations'; (4) 'no'.

$0.21 \leq \kappa \leq 0.40$	fair
$0.41 \leq \kappa \leq 0.60$	moderate
$0.61 \leq \kappa \leq 0.80$	substantial
$0.81 \leq \kappa \leq 1.00$	almost perfect agreement.

Fleiss²⁷ names the following categories:

$\kappa < 0.40$	poor agreement
$0.40 \leq \kappa \leq 0.75$	good
$\kappa > 0.75$	excellent agreement.

Using Fleiss's²⁷ categories for the interpretation of κ , all in all the inter-rater reliability at ACP can be called poor. All κ coefficients in Table 1 are lower than 0.40 – independently of whether weighted or unweighted κ , of whether favorable or non-favorable reviewers' ratings were the basis for the calculation of κ (see the conditional κ in the table), and of whether the ratings concerned an evaluation criteria or the final publication recommendation. However, applying the finer-grade categories proposed by Landis and Koch²⁶ reveals differences in the inter-rater reliabilities. Fifteen κ coefficients (mostly unweighted κ) are in the range of slight agreement ($0.00 \leq \kappa \leq 0.20$), whereas 12 κ coefficients (mostly weighted κ) are in the fair agreement range ($0.21 \leq \kappa \leq 0.40$). Only one coefficient lays in the poor category ($\kappa < 0.00$). As the coefficients in Table 1 show, there are no systematic differences in the inter-rater reliability between the evaluation criteria and the final publication recommendation.

Results with regard to ICC as coefficient for the inter-rater reliability

In addition to the κ coefficients, Table 1 shows the ICCs that result from the ANOVA and multi-level models. In contrast to κ , for ICC there exist no guidelines for the interpretation. In the literature we find only indications, such as that by Hargens and Herting,²⁸ concerning the ICC that can be expected in journal peer review:

It seems likely that the highest attainable intra-class correlation coefficient between reviewers' overall evaluations of manuscripts is far below 0.9; perhaps 0.5 or 0.6 is the upper bound. If so, observed coefficients between 0.2 and 0.3 may indicate reasonable levels of agreement. (p. 94)

Taking this advice as a guideline for interpretation of the ICCs in Table 1, the ICCs for the evaluation criteria and the publication recommendations of the reviewers lie in most cases (28 out of 32 coefficients) around 0.30 – independently of whether the ICC was calculated using the random-effects ANOVA model or the random-intercept proportional odds model. Only four coefficients are lower – namely, <0.20 . However, these latter coefficients are based on only 15 manuscripts or 60 reviews, respectively (four reviews per manuscript).

Discussion

According to Marsh, Bond, and Jayasinghe,²⁹ the most important weakness of the traditional peer-review process is that the ratings of the same manuscript by different reviewers typically differ. This results in a lack of inter-rater reliability. As the results of the present study on public peer review in the case of ACP show, inter-rater reliability has to be seen as low (κ) or reasonable (ICC) here as well: all κ coefficients are in the range of poor, slight, or fair agreement (26, 27); most of the ICCs shows a reasonable level of agreement (28). Hence, public peer review does not have the effect mentioned above of raising inter-rater reliability as compared to traditional closed peer review. Another step towards transparency in public peer review has not been shown to be effective to increase the agreement among reviewers.

Although a high level of inter-rater reliability is generally seen as desirable, when it comes to peer review, some researchers, such as Bailar,³⁰ view agreement as detrimental to the review process: 'Too much agreement is in fact a sign that the review process is not working well, that reviewers are not properly selected for diversity, and that some are redundant' (p. 138). Although selecting reviewers according to the principle of complementarity (e.g. choosing a generalist and a specialist) will lower inter-rater reliability, the validity of the process can gain, according to Langfeldt:³¹

Low inter-reviewer agreement on a peer panel is no indication of low validity or low legitimacy of the assessments. In fact,

public peer review does not have the effect mentioned above of raising inter-rater reliability as compared to traditional closed peer review

it may indicate that the panel is highly competent because it represents a wide sample of the various views on what is good and valuable research. (p. 821)

The task of reviewers is to guide editors in making their decisions on publication – not to decide themselves by a majority vote what should or should not be published.³²

the task of reviewers is to guide editors in making their decisions on publication

Differing recommendations in manuscript reviewing are not necessarily a sign of disagreement and can be due to the differing paradigmatic positions ('schools'), approaches, and mentalities of the reviewers.³³ In addition, reviewers can tend to be more critical or more lenient in their judgments;³⁴ they direct their attention, writes Eckberg,³⁵ to 'different points, and may draw different conclusions about "worth"' (p. 146). According to Starbuck,¹² 'some editors seek highly qualified reviewers, others . . . use inexperienced reviewers such as doctoral students' (p. 184).

Building on these statements, the finding of this study – low (κ) or reasonable (ICC) inter-rater reliability – should be interpreted not as an indication that the quality of the ACP peer review process is low but instead as a general characteristic of journal manuscript reviewing¹⁰ that possibly has a positive effect on the predictive validity of manuscript selection decisions. In recent years, some ways to increase inter-rater reliability in journal peer review have been suggested, such as including a greater number of reviewers.³⁶ Another suggestion is to make use of a structured form that specifies unambiguously the dimensions that a review should address, thereby introducing uniformity into the criteria and limiting the scope of reviewer subjectivity. However, as Cicchetti and Eron³⁷ demonstrated with manuscript reviews for the *Journal of Abnormal Psychology*, explicit definition of the reviewing criteria does not necessarily result in an increase in the reliability of reviewer judgments:

The reviewer agreement levels for 1973–1975 (the years when the manuscript attribute rating form (MARF) was not available) were, in fact, slightly higher than agreement levels for 1976–1977 (when the

MARF was first applied by reviewers). (p. 596)

Acknowledgements

The research project, which is investigating quality assurance of interactive open access journals, is supported by a grant from the Max Planck Society (Munich, Germany).

We thank Dr Ulrich Pöschl, Chief Executive Editor of *Atmospheric Chemistry and Physics*, the Editorial Board of *Atmospheric Chemistry and Physics*, and Copernicus Publications (Göttingen, Germany) for permission to conduct the evaluation of the selection process of the journal, and thank the members of Copernicus Systems + Technology (Berlin, Germany) for their generous technical support during the carrying out of the study.

References

1. Bailey, C.W. *Open Access Bibliography*. Washington DC, Association of Research Libraries, 2005.
2. Pöschl U. Interactive open access publishing and public peer review: the effectiveness of transparency and self-regulation in scientific quality assurance. submitted.
3. Koop, T. and Pöschl, U. 2006. Systems: an open, two-stage peer-review journal. The editors of *Atmospheric Chemistry and Physics* explain their journal's approach. 26 June 2006; available from: <http://www.nature.com/nature/peerreview/debate/nature04988.html>
4. Deutsche Forschungsgemeinschaft. *Publikationsstrategien im Wandel? Ergebnisse einer Umfrage zum Publikations- und Rezeptionsverhalten unter besonderer Berücksichtigung von Open Access*. Weinheim, Wiley-VCH, 2005.
5. DeCoursey, T. 2006. Perspective: the pros and cons of open peer review. Should authors be told who their reviewers are? Cited 20 June 2006; available from: <http://www.nature.com/nature/peerreview/debate/nature04991.html>
6. Joint Information Systems Committee. *Journal Authors Survey Report*. Truro, Key Perspectives Ltd, 2004.
7. Gesellschaft Deutscher Chemiker. *Diskussionspapier der Gesellschaft Deutscher Chemiker zum offenen Zugang zu wissenschaftlichem Wissen (Open Access)*. 2004. 21 June 2005; available from: <http://www.gdch.de/oearbeit/openaccess.pdf>.
8. Weller, A.C. *Editorial Peer Review: Its Strengths and Weaknesses*. Medford, NJ, Information Today, Inc., 2002.
9. Cicchetti, D.V. 1991. The reliability of peer review for manuscript and grant submissions: a cross-disciplinary investigation. *Behavioral and Brain Sciences*, 14(1): 119–135.
10. Bornmann, L., Mutz, R., and Daniel, H.-D. A reliability-generalization study of journal peer reviews – a multilevel meta-analysis of inter-rater reliability and its determinants. submitted.
11. Daniel, H.-D. 2005. Publications as a measure of scientific advancement and of scientists' productivity. *Learned Publishing*, 18: 143–148. <http://dx.doi.org/10.1087/0953151053584939>
12. Starbuck, W.H. 2005. How much better are the most-prestigious journals? The statistics of academic

- publication. *Organizational Science*, 16(2): 180–200.
<http://dx.doi.org/10.1287/orsc.1040.0107>
13. Bornmann, L. and Daniel, H.-D. 2008. The effectiveness of the peer review process: inter-referee agreement and predictive validity of manuscript refereeing at *Angewandte Chemie. Angewandte Chemie International Edition*, 47(38): 7173–7178.
<http://dx.doi.org/10.1002/anie.200800513>
 14. Bingham, C.M., Higgins, G., Coleman, R., and Van Der Weyden, M.B. 1998. The Medical Journal of Australia Internet peer-review study. *Lancet*, 352(9126): 441–445.
[http://dx.doi.org/10.1016/S0140-6736\(97\)11510-0](http://dx.doi.org/10.1016/S0140-6736(97)11510-0)
 15. Koonin, E. and Lipman, D. 2006. Systems: reviving a culture of scientific debate. Can 'open peer review' work for biologists? *Biology Direct* is hopeful. 21 June 2006; available from: <http://www.nature.com/nature/peerreview/debate/nature05005.html>
 16. Jefferson, T., Rudin, M., and Brodney Folse, S. F. D. 2007. Editorial peer review for improving the quality of reports of biomedical studies. *Cochrane Database of Systematic Reviews*, 2.
 17. van Rooyen, S., Godlee, F., Evans, S., Black, N., and Smith, R. 1999. Effect of open peer review on quality of reviews and on reviewers' recommendations: a randomised trial. *British Medical Journal*, 318(7175), 23–27.
 18. Pöschl, U. 2004 Interactive journal concept for improved scientific publishing and quality assurance. *Learned Publishing*, 17(2): 105–113.
<http://dx.doi.org/10.1087/095315104322958481>
 19. von Eye, A and Mun, E.Y. *Analyzing Rater Agreement. Manifest Variable Methods*. Mahwah, NJ: Lawrence Erlbaum Associates, 2005.
 20. StataCorp. Stata statistical software: release 11. College Station, TX, Stata Corporation, 2009.
 21. StataCorp. Stata base reference manual. Volume 2, release 11. College Station, TX, Stata Press, Stata Corporation, 2009.
 22. Reichenheim, M.E. 2004. Confidence intervals for the kappa statistic. *The Stata Journal*, 4(4): 421–428.
 23. Jayasinghe, U.W., Marsh, H.W., and Bond, N. 2003. A multilevel cross-classified modelling approach to peer review of grant proposals: the effects of assessor and researcher attributes on assessor ratings. *Journal of the Royal Statistical Society*, 166A: 279–300.
<http://dx.doi.org/10.1111/1467-985X.00278>
 24. Rabe-Hesketh, S. and Skrondal, A. *Multilevel and Longitudinal Modeling Using Stata*. 2nd edn. College Station, TX, Stata Press, 2008.
 25. Rabe-Hesketh, S., Skrondal, A., and Pickles, A. *GLLAMM Manual*. Berkeley, CA, University of California, 2004.
 26. Landis, J.R. and Koch, G.G. 1977. Measurement of observer agreement for categorical data. *Biometrics*, 33(1): 159–174.
<http://dx.doi.org/10.2307/2529310>
 27. Fleiss, J. *Statistical Methods for Rates and Proportions*. New York, Wiley VCH, 1981.
 28. Hargens, L.L. and Herting, J.R. 1990. Neglected considerations in the analysis of agreement among journal referees. *Scientometrics*, 19(1–2): 91–106.
<http://dx.doi.org/10.1007/BF02130467>
 29. Marsh, H.W., Bonds, N.W. and Jayasinghe, U.W. 2007. Peer review process: assessments by applicant-nominated referees are biased, inflated, unreliable and invalid. *Australian Psychologist*, 42(1): 33–38.
<http://dx.doi.org/10.1080/00050060600823275>
 30. Bailar, J.C. 1991 Reliability, fairness, objectivity, and other inappropriate goals in peer review. *Behavioral and Brain Sciences*, 14(1): 137–138.
 31. Langfeldt, L. 2001. The decision-making constraints and processes of grant peer review, and their effects on the review outcome. *Social Studies of Science*, 31(6): 820–841.
<http://dx.doi.org/10.1177/030631201031006002>
 32. Bornmann, L. Scientific peer review. *Annual Review of Information Science and Technology*, in press.
 33. Kostoff, R.N. 1995. Federal, research impact assessment – axioms, approaches, applications. *Scientometrics*, 34(2): 163–206.
<http://dx.doi.org/10.1007/BF02020420>
 34. Siegelman, S.S. 1991. Assassins and zealots – variations in peer review – special report. *Radiology*, 178(3): 637–642.
 35. Eckberg, D.L. 1991. When nonreliability of reviews indicates solid science. *Behavioral and Brain Sciences*, 14(1): 145–146.
 36. Daniel, H.-D. *Guardians of Science. Fairness and Reliability of Peer Review*. Weinheim, Wiley-VCH, 1993.
 37. Cicchetti, D.V. and Eron, L.D. 1979. The reliability of manuscript reviewing for the *Journal of Abnormal Psychology*. *Proceedings of the American Statistical Association (Social Statistics Section)*, 22: 596–600.

Dr Lutz Bornmann

ETH Zurich

Professorship for Social Psychology and

Research on Higher Education

Zähringerstr. 24

CH-8092 Zurich, Switzerland

Email: bornmann@gess.ethz.ch

Prof. Dr. Hans-Dieter Daniel

ETH Zurich, Professor for Social Psychology

and Research on Higher Education

and

University of Zurich, Evaluation Office

Mühlegasse 21

CH-8001 Zurich, Switzerland

Email: daniel@evaluation.uzh.ch