Interrater Reliability of the Peer Review Process in Management Journals

Alexander T. Nicolai, Stanislaw Schmal, and Charlotte L. Schuster

Abstract Peer review is an established method of assessing the quality and contribution of academic performance in most scientific disciplines. Up to now, little is known about interrater agreement among reviewers in management journals. This paper aims to provide an overview of agreement among the judgments of reviewers in management studies. The results of our literature review indicate a low level of agreement among reviewers in management journals. However, low consensus is not specific to management studies but widely present in other sciences as well. We discuss the consequences and implications of low judgment agreement for management research.

1 Introduction

In order to make a scientific contribution, research work has to be shared within the scientific community and come under discussion and scrutiny (Beyer et al. 1995). In management studies, as in any other academic discipline, scholarly journals serve as a platform for scientists to communicate their work to each other. Peer review is the predominant process through which manuscripts are evaluated prior to publication. It is a "key quality control mechanism" (Campanario 1998, p. 181; Rowland 2002) and has a "gatekeeping role" (Beyer et al. 1995, p. 1219). Publication in academic journals is regarded as an indicator of the "quality of role-performance in a social system" (Zuckerman and Merton 1971, p. 66). Review processes thus play a significant role in an author's career development, salary, and recognition within the scientific community (Frey 2003; Hunt and Blair 1987; Ketchen and Ireland 2010; Miller 2006). Since the rankings of departments and universities are also frequently based on publications in peer-refereed journals (Frey 2003, p. 211), the decisions of reviewers have a significant impact on academic systems in general.

A.T. Nicolai (✉) • S. Schmal • C.L. Schuster
Carl von Ossietzky Universität, Oldenburg, Germany
e-mail: alexander.nicolai@uni-oldenburg.de; stanislaw.schmal@uni-oldenburg.de; charlotte.schuster@uni-oldenburg.de

I.M. Welpe et al. (eds.), *Incentives and Performance*,
DOI 10.1007/978-3-319-09785-5_7

Given the importance of peer review, it is not surprising that this method "arouses very diverse emotions, beliefs, and ambitions. It angers, it reassures, it intimidates, it tramples egos, and it puffs them up" (Starbuck 2003, p. 348). At the same time, peer review is a heavily discussed topic, also in management studies (Miller 2006; Nicolai et al. 2011; Starbuck 2003). The most widely discussed aspects of peer review include validity, generalizability, accuracy, and bias (Campanario 1998; Marsh et al. 2008). In particular, the biases that affect peer review receive a lot of attention. Scholars from various disciplines, including the sociology of science, have named up to 25 different biases that can affect the fairness of peer review.[1] Campanario (1998), for instance, discusses evidence for bias towards positive and statistically significant results, as well as for bias against replication (Hubbard et al. (1998) show that bias against replication also applies to the area of Strategic Management). Gans and Shepherd (1994) discuss bias against fundamentally new ideas and refer to later Nobel Laureates whose manuscripts were often rejected in the first instance. Other scholars argue that some authors are favored over others who produce the same or even better quality which results from biases of reputation (Beyer et al. 1995; Merton 1968; Miller 2006). Other authors discuss a possible gender bias (Bornmann 2008).

Biased judgments and, accordingly, a lack of fairness are certainly among the most discussed issues in peer review. Still, the proponents of peer review argue that, although this method is imperfect, it is "more effective than any other known instrument for self-regulation in promoting the critical selection that is crucial to the evolution of scientific knowledge" (Bornmann 2011, p. 202).

One of the "most basic", "damning", and "broadly supported" criticism of peer review is "its failure to achieve acceptable levels of agreement among independent assessors", which makes peer review unreliable (Marsh et al. 2008, p. 161f). According to Mutz et al. (2012, p. 1), differing reviewer judgments of manuscripts are "[o]ne of the most important weaknesses of the peer review process." The reliability of peer review is typically studied by measuring "interrater reliability" (Bornmann 2008, 2011). Cicchetti (1991, p. 120) defines interrater or interreferee reliability as "the extent to which two or more independent reviews of the same scientific document agree."

This article aims to discuss dissensus among reviewer judgments regarding the acceptance, revision or rejection of a manuscript. We specifically provide an updated overview of studies on the degree of consensus among reviewers in the peer review process of papers for publication in management journals. We compare the empirical results of management studies with those of studies from other disciplines. Finally, consequences of high dissensus that is observed among

[1] See Bornmann (2008, p. 26) and Cicchetti (1991, p. 129) for a list of literature on peer review research discussing different biases. See also Campanario (1998) who discusses fraud, favoritism, self-interest, the connections among authors, reviewers, and editors, as well as the suggestibility of particularistic criteria in the context of double-blind reviewing.

reviewers and its implications for management studies and the academic system are discussed.

2 Interrater Reliability in Management Studies

There is a controversial debate on the most appropriate statistical measure for interrater reliability (see for an overview Conger and Ward 1984; or Whitehurst 1984). Cicchetti (1991, p.120) points out that an appropriate statistical method should account for the number of referees, the matching procedure, and the degree of reviewer agreement that can be attributed to chance alone. Considering the points mentioned by Cicchetti (1991), interclass correlation (ICC) and Cohen's kappa (k) are argued to be the most appropriate measures of interrater reliability (Bartko 1976; Cicchetti 1980; Spitzer and Fleiss 1974; Tinsley and Weiss 1975). Interclass correlation provides reliability estimates of assessments made by two or more referees of the same manuscript. Full agreement between reviewers is indicated by the value 1.0 (Shrout and Fleiss 1979; Whitehurst 1984). Cohen's kappa is a statistical method for identifying the degree of agreement between two or more raters that is above the agreement that could be expected by chance (Fleiss and Cohen 1973). It ranges from −1 to 1. Negative values indicate poorer agreement than would be expected by chance, 0 indicates chance agreement, and positive values are interpreted as chance-corrected percentage agreement (Landis and Koch 1977; Weller 2001).

Interrater reliability in management journals is a seldom analyzed issue. Our systematic research of the relevant literature identified five studies that analyze the level of agreement among reviewers in management studies. Table 1 presents the results of five studies including three management journals: *Academy of Management Journal* (*AMJ*), *Administration Science Quarterly* (*ASQ*), and a German journal, *Zeitschrift für Führung und Organisation* (*ZFO*). The table shows the methods each study applied, the results it obtained, and the qualitative interpretation of the authors.

The studies presented here cover the period between 1976 and 2005. Cummings et al. (1985) initiated the debate on disagreement among referees in the management discipline. They analyzed the statements of reviewers on manuscripts submitted to the *AMJ* between 1976 and 1978. In 34 of 81 cases the authors found that the reviewers' recommendations were inconsistent. This corresponds to a disagreement rate of almost 42 %, indicating that there was no common ground among the reviewers' evaluations.

The next study, which was published 10 years later by Beyer et al. (1995), picks up the issue of interrater agreement in management sciences. Using a data sample of 400 manuscripts submitted to the *AMJ* from 1984 to 1987, Beyer et al. analyzed the effects of author, manuscript and review process characteristics on the publication decisions in the first and final review. They calculated an indicator for reviewer disagreement as the standard deviation of the five-point scaled submission

Table 1 Interrater agreement in management studies

Journal/Authors	Agreement	Sample size	Categories	Author's interpretation
Academy of Management Journal (AMJ) (Cummings et al. 1985)	42 % disagreement[a]	81		None
Academy of Management Journal (AMJ) (Beyer et al. 1995)	$SD = 0.69$	400	5	None
Administrative Science Quarterly (ASQ) (Starbuck 2003)	$\rho = 0.12$ Pearson product–moment correlation	~500	3	"Little agreement among reviewers" (p. 348)
Academy of Management Journal (AMJ) (Miller 2006)	37 % disagreement	68	5	"Dissensus is present at AMJ but certainly not to the extent that it could be" (p. 429)
Zeitschrift Führung + Organisation (ZFO) (Nicolai et al. 2011)	ρ(full sample) = 0.19 ρ(rejected) = 0.02 ρ(accepted) = −0.25[b]	142	5	"Weakly positive relationship between the evaluations of academics and practitioners" (p. 60)

ρ = Pearson product–moment correlation; *SD*=standard deviation
[a]In 34 of 81 cases the authors found that the reviewers' recommendations were inconsistent
[b]"Full" indicates a sample consisting of rejected and accepted manuscripts; "Rejected" indicates a sample consisting only of rejected manuscripts; "accepted" indicates a sample consisting only of accepted manuscripts

ratings. Even though this was not their main focus, some of their results provide interesting insights into interrater agreement in the *AMJ*. For new submissions the authors report a disagreement of 0.69 (averaged intra-paper *SD*). One may argue that this is not an extremely high value for a five-point Likert scale (Miller 2006, p. 429). However, a closer look at the results indicates an almost equivalent averaged dispersion from the mean value by 0.62, which may lead to a difference of more than one level on the scale.

Starbuck (2003) evaluated the review process from the editorial point of view. Drawing on his own experience as an editor for the *ASQ* in the late 1960s, he calculated the correlation among three-point scaled peer review recommendations of around 500 manuscripts. It resulted in a significant but almost negligible coefficient of 0.12. This value indicates a very low level of agreement among reviewers.

Further results on dissensus are discussed by Miller (2006). He mostly agrees that there is low agreement among reviewers in sociological and psychological research as a whole, but critically questions whether this is the case in the *AMJ*. Using a randomly drawn sample of 68 cases from Rynes editorship, he calculated the disagreement rate and the standard deviation of recommendations. He comes up with a disagreement rate of 37 %, which is similar to the result reported by

Cummings et al. (1985). Likewise his results on within-paper standard deviation[2] correspond to Beyer et al. (1995). As mentioned earlier, the level of both results indicates the existence of considerable dissensus among *AMJ* reviewers.

Nicolai et al. (2011) examined disagreement among the reviewers of a German management journal. This study is a special case in that the authors analyze a so-called bridge journal, the *ZFO*, which uses double-blind reviews conducted by one academic and one practitioner. The study's sample consists of 142 manuscripts submitted to the *ZFO* between 1995 and 2005. All examined recommendations are based on a five-point Likert scale. Correlation analysis reveals a significant ($p < 0.05$) but relatively low correlation (0.19), which implies an almost negligible relationship among the reviewers. Further analyses indicate a substantial difference in agreement between accepted and rejected manuscripts. The correlation coefficient of reviews for rejected papers was insignificant different from zero. In contrast, the coefficient for accepted papers is significant ($p < 0.10$) but negative (-0.25), which suggests a weak inverse relationship between the recommendations of academics and of practitioners.

Another study, which analyzes authors' opinions instead of directly comparison of review judgments, is not presented in Table 1. Bedeian (2003) analyzed the experience of 173 authors whose manuscripts were accepted for publication in the *AMJ* or the *AMR* between 1999 and 2001. In particular, Bedeian asked the leading authors how satisfied they were with the consistency of the different referees' comments (Bedeian 2003, p. 334). More than half of the respondents—a total of 107 (62 %)—were satisfied with the uniformity of the reviewers' statements. Another 34 (20 %) were neither satisfied nor dissatisfied, and 31 (18 %) were unsatisfied with the degree of agreement among the reviewers' recommendations. Bedeian (2003, p. 333) concluded that the "authors expressed satisfaction with the consistency of the editor and referee comments among one another." However, this result should be interpreted with caution. As Miller (2006, p. 429) points out, the nature of Bedeian's sample is biased towards successful authors. Indeed, Bedeian (2003, p. 335) himself states that as "several authors noted, they would have responded to various survey items differently had their manuscripts been rejected."

The results of all studies presented here that directly analyzed interrater agreement mostly indicate low consensus among reviewers in management studies. Moreover, the works indicate that dissensus in management studies seems to be independent of editor, journal, or period covered by the data. However, the methods these studies use (disagreement rate, correlations etc.) hardly make it possible to qualify the level of agreement or make comparisons between different studies and different measures. Furthermore, frequency statistics and correlation measures do not consider chance agreement among raters and present an additive bias (Whitehurst 1984). For this reason, in our literature review we also gathered the qualitative statements that the authors of these works have made (see column 5 in Table 1). All in all, these statements demonstrate that the degree of interrater

[2] The author or Miller (2006, p. 429) do not report numerical results.

agreement was smaller than expected. Starbuck (2003, p. 349) concludes that "knowing what one reviewer had said about a manuscript would tell me almost nothing about what a second reviewer had said or would say."

Some authors explain dissensus among reviewers with the "low paradigm development" of an academic discipline (Beyer et al. 1995; Miller 2006; Starbuck 2003). "Low paradigm development" refers to a lack of common assumptions, terminology, theory, and methodology (e.g., Pfeffer 1993; Zammuto 1984). These authors follow Lodahl and Gordon (1972), who used Kuhn's concept (1962) of the scientific paradigm to classify the major sciences as "low-paradigm" or "high-paradigm" disciplines. On the basis of the state of technological development and the level of consensus within each field of research, the social sciences are classified as "low-paradigm" disciplines, whereas the natural sciences, such as physics or chemistry, are classified as "high-paradigm" disciplines (Lodahl and Gordon 1972).

According to this scheme, management studies can be classified as a "low–paradigm" field (Beyer et al. 1995, p. 1230; Pfeffer 1993). Beyer et al. (1995) argue that in the social sciences there are no universal scientific criteria on the basis of which generally accepted decisions could be made. Whitley (1984, p. 780) describes management studies as a "fragmented type of scientific field" with a high degree of technical and strategic task uncertainty. In contrast to mathematicians or economists, management researchers do not coordinate or control their work in keeping with a "global view" of their discipline (Whitley 1984, p. 799f). The level of agreement in management studies can be described as "multitude, vague, and ever changing. Researchers do not behave as if agreements exist about some beliefs and perceptions being correct" (Starbuck 2003, p. 348). Thus, "low-paradigm" development, as well as the immaturity of an academic discipline, can be a reason for the inconsistent (usage of) judgment criteria in the review process and dissensus among reviewers (Beyer et al. 1995).

3 Interrater Reliability in Other Sciences

If the "low-paradigm" character of management studies is the reason for dissensus, agreement should be higher in "high-paradigm" disciplines. On the basis of the classification metric of Lodahl and Gordon (1972) of "low-" and "high-paradigm" disciplines, we summarize in Table 2 the results on interrater reliability among reviewers in two disciplines classified as "high-paradigm": chemistry and physics.

We identified studies on referee agreement in journal submission and grant application processes. Zuckerman and Merton (1971) examined the referee system as a whole and analyzed the patterns of decision-making among editors and reviewers in *The Physical Review* (*PR*). Examining 172 manuscripts evaluated by two referees between 1948 and 1956, the authors calculated a rate on full

Table 2 Interrater overview in "high-paradigm" disciplines

Journal/Authors	Discipline	Agreement	Sample size	Categories	Author's interpretation
The Physical Review (PR) (Zuckerman and Merton 1971)	Physics	3[a]–67 %[b] disagreement rate	172	2–4	"Agreement was very high" (p. 67)
Physical Review Letters (Lazarus 1982)	Physics	85–90 % disagreement rate			None
Angewandte Chemie (Daniel 1993)	Chemistry	K = 0.2	856	4	"Reviewer agreement must be described as rather unsatisfying" (p. 23)
Angewandte Chemie (Bornmann and Daniel 2008)	Chemistry	K = 0.1–0.21	1,899	4	"Low level of agreement among referees' recommendations" (p. 7174)
Grants within Committee on Science and Public Policy (Cole et al. 1981)	Chemical dynamics	53 % (share of total variance)	50	12	"Substantial reviewer variance" (p. 884)
Grants within Committee on Science and Public Policy (Cole et al. 1981)	Solid-state physics	47 % (share of total variance)[c]	50	12	

K = Cohen's kappa
[a]Refers to full disagreement that one referee recommended acceptance and the other rejection
[b]Refers to minor differences between referees
[c]Percentage of total variance in reviewers' ratings accounted for by differences among reviewers of individual proposals (Cole et al. 1981, p. 884)

disagreement[3] of about 3 %. This finding indicates a very high level of agreement. However, other sets of results from this study draw a slightly different picture. As the authors state, two-thirds of these recommendation judgments reveal "minor differences in the character of proposed revisions" (Zuckerman and Merton 1971, p. 67, footnote 3). Unfortunately, no further information is given on the recommendation range or its variance. Thus, the evaluation of those results is hardly possible.

Lazarus (1982) examined the issue of interrater agreement in physics. He claims that the agreement rate among reviewers on the question of accepting or rejecting a

[3] Full disagreement implies that one referee recommended acceptance and the other rejection.

manuscript in *Physical Review Letters* is about 10–15 %. This corresponds to a disagreement rate of about 85–90 %, which is more than twice as high as the disagreement rate in management journals. Unfortunately, Lazarus (1982) does not provide further details on this analysis.

Interrater reliability in chemistry was examined in Daniel (1993) and in Bornmann and Daniel (2008). Both studies used the referees' recommendations of the *Angewandte Chemie*, which is a journal published by the German Chemical Society (Gesellschaft Deutscher Chemiker, GDCh, Frankfurt am Main). As a measure of agreement, both studies used Cohen's kappa (Fleiss and Cohen 1973). In the earlier study, Daniel (1993) examined a sample of 856 manuscripts covering the mid-1980s. His calculations revealed a Cohen's kappa of 0.20, implying that agreement among the reviewers' judgments of these manuscripts is 20 % higher than would be expected by chance. Following the guidelines of Landis and Koch (1977) on how such measurements should be interpreted, this corresponds to a low level of agreement among reviewers.

In the subsequent study, Bornmann and Daniel (2008) used a larger sample of 1,899 manuscripts refereed in the year 2000. They applied a more advanced statistical method, weighted Cohen's kappa, which takes into account the different level of agreement between two or more referees. Depending on the weighting parameter, their kappa coefficients range between 0.10 and 0.21. That is, reviewers agreed on 10–21 % more manuscripts than could have been expected by chance. Thus, the more recent study shows an even lower degree of agreement among reviewers in the journal *Angewandte Chemie*.

Further results on interrater agreement in high paradigm disciplines can be derived from the analysis of Cole et al. (1981). The authors analyzed interrater reliability in a grant application process. Their data sample consisted of 150 proposals in chemical dynamics, solid-state physics, and economics—50 from each discipline—and about 12 reviewers for each proposal. In order to avoid making statistical assumptions, the authors used the analysis-of-variance approach to determine the level of consensus among referees. The variance in the ratings is decomposed into variation in the quality of a proposal, in the review procedure, and in the reviewers' judgments. The authors' results show that most of the variance in the ratings of the proposals is due to reviewer disagreement and not to differences in content or methodology. In fact, in chemistry dynamics the variance in the reviewers' ratings accounted for 53 % of the total variance. In solid-state physics this value reached 47 %. In both cases, the variance among the reviewers of the same proposal is almost twice as high as the variation in the quality of proposals. The authors conclude by saying that "[c]ontrary to a widely held belief that science is characterized by wide agreement [. . .] our research both in this and other studies in the sociology of science indicates that concerning work currently in process there is substantial disagreement in all scientific fields" (Cole et al. 1981, p. 885). Other articles presenting overviews of the reliability of peer review or meta-analysis also did not find "any effect of [the] discipline" (Bornmann et al. 2010, p. 7) on interrater agreement among reviewers (Cicchetti 1991; Weller 2001).

The results presented in Table 2 indicate that also in "high-paradigm" disciplines interrater agreement is low. In view of that, the argument that dissensus among reviewers is a consequence of the "low-paradigm" nature of management studies seems fragile. On the contrary, dissensus among reviewers who assess works submitted to academic journals appears to be a common issue in science.

Conclusion

Interrater reliability is a topic that "goes to the heart of peer review" (Miller 2006, p. 426). The advancement of knowledge would be impossible if scientists were not able to reach a certain degree of consensus (Pfeffer 1993, p. 611). If a work of research is rigorous, it can be expected that two independent reviewers will agree on its quality (Bornmann 2011, p. 207). Our overview on interrater reliability in peer review conducted for management journals shows that this is often not the case: there seems to be little consensus among reviewers in management studies. Some authors attribute this tendency to the "low-paradigm" and fragmented nature of management research, as a result of which, inconsistent judgment criteria may be applied in the review process (Beyer et al. 1995, p. 1255). However, a low degree of interrater agreement is not specific to management studies. Also "high-paradigm" fields exhibit a high degree of reviewer disagreement. Thus, the hope that consensus might grow as the paradigm of management studies develops seems delusive. Dissensus could even increase if, as some authors (e.g., Cohen 2007) suggest, management journals integrate more and more science-external audiences into the peer review process (see Nicolai et al. 2011 for a critical discussion).

A high degree of dissensus illustrates the central role of journal editors for two reasons. First, the editor's opinion is given greater prominence if the reviewers' recommendations point in different directions. Second, the editor chooses the referees, and the result of a review might depend more on the selected reviewer than on the quality of the submitted manuscript. Kravitz et al. (2010, p. 4) found that, "recommendations were more consistent for multiple manuscripts assigned to the same reviewer (intra-class correlation coefficient rho = 0.23) than for multiple reviewers assessing the same manuscript (rho = 0.17)." Overall, low reliability implies a certain randomness of the peer review process. Consequently, publication in journals should not serve as the only measure of scientific performance. Instead, researchers should triangulate different kinds of evidence about scientific quality (Starbuck 2005, p. 197). An alternative form of measuring scholarly performance is to use its communications or hyperlinks on the World Wide Web. Webometrics is based on "link analysis, web citation analysis, search engine evaluation and purely descriptive studies of the web" (Thelwall 2008, p. 611). Examples of such measures are hyperlinks between pages (Aguillo

(continued)

et al. 2006) or numbers of external inlinks received by one's website (Tang et al. 2012). A further development on webometrics is provided by the so called altmetrics indicators. Those additionally take into account social bookmarking services and the number of downloads (see for an overview of altmetrics indicators Weller (2015)). Those measures are mainly unattached by reviewer judgments, but are available on a large scale and mostly immediately after publications. Moreover, these indices additionally assess the impact of readers and not only citers (Thelwall et al. 2013).

The statistical tests that the studies included in this overview applied to examine reviewer agreement are the subject of an ongoing methodological debate (Kravitz et al. 2010; Weller 2001). For example, percentages of agreement and Pearson product–moment correlation present numerous problems, while raw frequency counts do not distinguish agreement by chance alone and thus include both true and chance agreement (Watkins 1979). The correlation approach corrects for random agreement; however, as a measure of association rather than reliability, this statistic does not provide any information on the level of disagreement among reviewers. In fact, it can obtain perfect correlation even if the referees never agreed exactly but disagreed proportionally (Hendrick 1976; Whitehurst 1984). Kravitz et al. (2010, p. 4) criticize the methods that many studies apply to analyze the assessments of reviewers: they assume that judgments vary along one latent dimension of publishability and merit but that this "can hardly be tested by calculating kappa or intraclass correlation coefficients." In a similar vein, Hargens and Herting (1990) also criticized the latent-dimension approach. As Hargens and Herting (1990, p. 14) argue, by using the row–column association model of Goodman (1984), it is possible to "derive information about the distances between recommendation categories empirically rather than requiring arbitrary assumptions about those distances." The authors recommend that researchers should analyze these issues before calculating any disagreement measures.

It should be noted, however, that low interrater agreement is not necessarily a sign that the review process is not working well. Editors might deliberately choose reviewers with complementary expertise and opposing perspectives to obtain recommendations on different aspects of a piece of research (Hargens and Herting 1990, p. 2). It is open to debate whether the choice of reviewers with complementary competences can explain the low agreement rates observed among reviewers. So far, very few studies have analyzed comparatively the content of reviewers' comments to identify the reasons behind their disagreement (Bornmann 2011, p. 226). An analysis of this issue could shed light on an interesting discrepancy. Reviews, which usually are not publicly available, are characterized by very different points of view and often harsh criticism (Miller 2006). The published debate in

(continued)

management studies is much more consensual. Negational or critical citations in scholarly management articles are very rare (Schulz and Nicolai 2014). A better understanding of why exactly the opinions of reviewers differ could contribute not only to the improvement of the reviewing process, but also to the progress of the scholarly management debate in general.

References

Aguillo IF, Granadino B, Ortega JL, Prieto JA (2006) Scientific research activity and communication measured with cybermetrics indicators. J Am Soc Inf Sci Technol 57(10):1296–1302

Bartko JJ (1976) On various intraclass correlation reliability coefficients. Psychol Bull 83 (5):762–765

Bedeian AG (2003) The manuscript review process: the proper roles of authors, referees, and editors. J Manag Inq 12(4):331–338

Beyer JM, Roland GC, Fox WB (1995) The review process and the fates of manuscripts submitted to amj. Acad Manage J 38(5):1219–1260

Bornmann L (2008) Scientific peer review. An analysis of the peer review process from the perspective of sociology of science theories. Hum Archit 6(2):23–38

Bornmann L (2011) Scientific peer review. Ann Rev Inf Sci Technol 45:199–245

Bornmann L, Daniel H-D (2008) The effectiveness of the peer review process: inter-referee agreement and predictive validity of manuscript refereeing at angewandte chemie. Angew Chem Int Ed 47(38):7173–7178

Bornmann L, Mutz R, Daniel H-D (2010) A reliability-generalization study of journal peer reviews: a multilevel meta-analysis of inter-rater reliability and its determinants. PLoS ONE 5(12):e14331

Campanario JM (1998) Peer review for journals as it stands today-part 1. Sci Commun 19 (3):181–211

Cicchetti DV (1980) Reliability of reviews for the American psychologist: a biostatistical assessment of the data. Am Psychol 35(3):300–303

Cicchetti DV (1991) The reliability of peer review for manuscript and grant submissions: a cross-disciplinary investigation. Behav Brain Sci 14(01):119–135

Cohen DJ (2007) The very separate worlds of academic and practitioner publications in human resource management: reasons for the divide and concrete solutions for bridging the gap. Acad Manag J 50:1013–1019

Cole S, Cole RJ, Simon AG (1981) Chance and consensus in peer review. Science 214 (4523):881–886

Conger AJ, Ward DG (1984) Agreement among 2×2 agreement indices. Educ Psychol Meas 44 (2):301–314

Cummings LL, Frost PJ, Vakil TF (1985) The manuscript review process: a view from inside on coaches, critics, and special cases. In: Cummings LL, Frost PJ (eds) Publishing in the organizational sciences. Irwin, Homeland, pp 469–508

Daniel H-D (1993) Guardians of science: fairness and reliability of peer review. VCH, Weinheim

Fleiss JL, Cohen J (1973) The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. Educ Psychol Meas 33(3):613–619

Frey BS (2003) Publishing as prostitution? - choosing between one's own ideas and academic success. Public Choice 116(1–2):205–223

Gans JS, Shepherd GB (1994) How are the mighty fallen: rejected classic articles by leading economists. J Econ Perspect 8(1):165–179

Goodman LA (1984) The analysis of cross-classified data having ordered categories. Harvard University Press, Cambridge
Hargens LL, Herting JR (1990) A new approach to referees' assessments of manuscripts. Soc Sci Res 19(1):1–16
Hendrick C (1976) Editorial comment. Pers Soc Psychol Bull 2:207–208
Hubbard R, Vetter DE, Littel EL (1998) Replication in strategic management: scientific testing for validity, generalizability, and usefulness. Strateg Manage J 19:243–254
Hunt JG, Blair JD (1987) Content, process, and the Matthew effect among management academics. J Air Waste Manage Assoc 13(2):191–210
Ketchen D, Ireland RD (2010) From the editors upon further review: a survey of the academy of management journal's editorial board. Acad Manage J 53(2):208–217
Kravitz RL, Franks P, Feldman MD, Gerrity M, Byrne C, Tierney WM (2010) Editorial peer reviewers' recommendations at a general medical journal: are they reliable and do editors care? PLoS ONE 5(4):1–5
Kuhn T (1962) The structure of scientific revolutions, vol 2. University of Chicago Press, Chicago
Landis JR, Koch GG (1977) The measurement of observer agreement for categorical data. Biometrics 33(1):159–174
Lazarus D (1982) Interreferee agreement and acceptance rates in physics. Behav Brain Sci 5 (2):219–219
Lodahl JB, Gordon G (1972) The structure of scientific fields and the functioning of university graduate departments. Am Soc Rev 37(1):57–72
Marsh HW, Jayasinghe UW, Bond NW (2008) Improving the peer-review process for grant applications. Reliability, validity, bias, and generalizability. Am Psychol 63(3):160–168
Merton RK (1968) The Matthew effect in science. Science 159(3810):56–60
Miller CC (2006) From the editors: peer review in the organizational and management sciences: prevalence and effects of reviewer hostility, bias, and dissensus. Acad Manage J 49 (3):425–431
Mutz R, Bornmann L, Daniel H-D (2012) Heterogeneity of inter-rater reliabilities of grant peer reviews and its determinants: a general estimating equations approach. PLoS ONE 7(10):1–10
Nicolai AT, Schulz A-C, Göbel M (2011) Between sweet harmony and a clash of cultures: does a joint academic–practitioner review reconcile rigor and relevance? J Appl Behav Sci 47 (1):53–75
Pfeffer J (1993) Barriers to the advance of organizational science: paradigm development as a dependent variable. Acad Manage Rev 18(4):599–620
Rowland F (2002) The peer-review process. Learned Publish 15(4):247–258
Schulz A-C, Nicolai A (2014) The intellectual link between management research and popularization media: a bibliometric analysis of the harvard business review. Acad Manage Learn Educ. forthcoming
Shrout PE, Fleiss JL (1979) Intraclass correlations: uses in assessing rater reliability. Psychol Bull 86(2):420–428
Spitzer RL, Fleiss JL (1974) A re-analysis of the reliability of psychiatric diagnosis. Br J Psychiatry 125(587):341–347
Starbuck WH (2003) Turning lemons into lemonade: where is the value in peer reviews? J Manage Inq 12(4):344–351
Starbuck WH (2005) How much better are the most-prestigious journals? The statistics of academic publication. Organ Sci 16(2):180–200
Tang M-C, Wang C-M, Chen K-H, Hsiang J (2012) Exploring alternative cyberbibliometrics for evaluation of scholarly performance in the social sciences and humanities in Taiwan. Proc Am Soc Inf Sci Technol 49(1):1–1
Thelwall M (2008) Bibliometrics to webometrics. J Inf Sci 34(4):605–621
Thelwall M, Haustein S, Larivière V, Sugimoto CR (2013) Do altmetrics work? Twitter and ten other social web services. PLoS ONE 8(5):e64841

Tinsley HE, Weiss DJ (1975) Interrater reliability and agreement of subjective judgments. J Couns Psychol 22(4):358–376
Watkins MW (1979) Chance and interrater agreement on manuscripts. Am Psychol 34(9):796–798
Weller AC (2001) Editorial peer review: its strengths and weaknesses, Asist monograph series. Hampton Press, New Jersey
Weller K (2015) Social media and altmetrics: an overview of current alternative approaches to measuring scholarly impact. In: Welpe IM, Wollersheim J, Ringelhan S, Osterloh M (eds) Incentives and performance - governance of research organizations. Springer International Publishing AG, Cham
Whitehurst GJ (1984) Interrater agreement for journal manuscript reviews. Am Psychol 39 (1):22–28
Whitley R (1984) The development of management studies as a fragmented adhocracy. Soc Sci Inf 23(4–5):775–818
Zammuto RF (1984) Coping with disciplinary fragmentation. J Manage Educ 9(30):30–37
Zuckerman H, Merton RK (1971) Patterns of evaluation in science: institutionalisation, structure and functions of the referee system. Minerva 9(1):66–100