Scientometrics (2017) 113:1059–1092
DOI 10.1007/s11192-017-2516-6

Inter-rater reliability and validity of peer reviews in an interdisciplinary field

Jens Jirschitzka[1] · Aileen Oeberst[2,3] · Richard Göllner[1] · Ulrike Cress[1,3]

Received: 5 April 2017 / Published online: 11 September 2017

Abstract Peer review is an integral part of science. Devised to ensure and enhance the quality of scientific work, it is a crucial step that influences the publication of papers, the provision of grants and, as a consequence, the career of scientists. In order to meet the challenges of this responsibility, a certain shared understanding of scientific quality seems necessary. Yet previous studies have shown that inter-rater reliability in peer reviews is relatively low. However, most of these studies did not take ill-structured measurement design of the data into account. Moreover, no prior (quantitative) study has analyzed inter-rater reliability in an interdisciplinary field. And finally, issues of validity have hardly ever been addressed. Therefore, the three major research goals of this paper are (1) to analyze inter-rater agreement of different rating dimensions (e.g., relevance and soundness) in an interdisciplinary field, (2) to account for ill-structured designs by applying state-of-the-art methods, and (3) to examine the construct and criterion validity of reviewers' evaluations. A total of 443 reviews were analyzed. These reviews were provided by $m = 130$ reviewers for $n = 145$ submissions to an interdisciplinary conference. Our findings demonstrate the urgent need for improvement of scientific peer review. Inter-rater reliability was rather poor and there were no significant differences between evaluations from reviewers of the same scientific discipline as the papers they were reviewing versus reviewer evaluations of papers from disciplines other than their own. These findings extend beyond those of prior research. Furthermore, convergent and discriminant construct validity of the rating

Jens Jirschitzka and Aileen Oeberst have shared first authorship.

Electronic supplementary material The online version of this article (doi:10.1007/s11192-017-2516-6) contains supplementary material, which is available to authorized users.

✉ Aileen Oeberst
aoeberst@uni-mainz.de

[1] Eberhard Karls Universität Tübingen, Tübingen, Germany

[2] Johannes Gutenberg-Universität Mainz, Binger Str. 14-16, 55122 Mainz, Germany

[3] Leibniz-Institut für Wissensmedien, Tübingen, Germany

dimensions were low as well. Nevertheless, a multidimensional model yielded a better fit than a unidimensional model. Our study also shows that the citation rate of accepted papers was positively associated with the relevance ratings made by reviewers from the same discipline as the paper they were reviewing. In addition, high novelty ratings from same-discipline reviewers were negatively associated with citation rate.

Keywords Peer review · Inter-rater reliability · Construct validity · Criterion validity · Interdisciplinary research · Citation rate

Introduction

Scientific progress is a fundamentally social process. Research not only always builds on the work of others, their ideas and findings (Hardwig 1985), but it also particularly benefits from different viewpoints and strategies (Kitcher 1990), as well as from empirical investigation and critical discussion (Popper 1968). Peer review has introduced the social element into the publication process (Fiske and Fogg 1990). That is, examination and discussion of a researcher's work by colleagues has already been carried out before it is made accessible to the whole scientific community. Today, most scientific disciplines trust in this kind of quality control (Hemlin and Rasmussen 2006). The underlying rationale of this procedure was to avoid errors and to ensure a certain quality of the publications (Bailar and Patterson 1985; Church et al. 1996; Cornforth 1974). In order to meet the challenges of this responsibility, a certain shared understanding of what characterizes quality seems necessary. Much research has already examined inter-rater reliability, also called inter-rater agreement, and provided a rather pessimistic picture.

In an effort to expand upon prior research, the paper presented here has the following three major research goals: First, we investigate whether prior findings of poor inter-rater reliability are generalizable to the interdisciplinary context. Our second major goal is to discuss and apply adequate methods for ill-structured measurement designs and thereby take reviewer discipline into account. The third major goal is to examine the underlying structure of the ratings of specific paper characteristics and to explore their potential to predict the citation rate of published papers.

To accomplish these goals, the paper is structured as follows: First, we outline previous research on inter-rater reliability. Second, we turn to methods of analysis used in prior research, with their limitations and possible solutions. Third, we summarize research on dimensionality and construct validity of peer-review ratings. In addition, we also consider prior research into criterion validity with regard to the predictability of the citation rate by reviewer recommendations. Then, we describe the data on which our analyses are based. Finally, we report our results, which we subsequently discuss in the light of previous research and practical implications.

Inter-rater reliability in interdisciplinary peer-review processes

To date, numerous studies have examined inter-rater reliability among reviews of different scientific products, such as abstracts (e.g., Blackburn and Hakel 2006; Cicchetti and Conn 1976; Rubin et al. 1993), grant proposals (e.g., Cole et al. 1981; Jayasinghe et al. 2003; Marsh et al. 2007) and scientific papers (e.g., Gottfredson 1978; Marsh and Ball 1981; Wood et al. 2004), as well as applications to scholarships (Bornmann and Daniel 2005). Research clearly indicates that evaluations of the same scientific manuscript differ substantially among reviewers. That is, the level of inter-rater reliability is quite low (Bornmann et al. 2010; Campanario 1998; Cicchetti 1991; Lindsey 1988). Accordingly, and consistent with previous literature reviews (Cicchetti 1991; O'Brien 1991), a recent meta-analysis based on 48 studies (and 70 reliability coefficients) concluded that inter-rater reliability "is quite limited and needs improvement" (mean ICC = .34, mean Cohen's Kappa = .17; Bornmann et al. 2010, p. 1; for a critical view on the Kappa coefficient, see Baethge et al. 2013).[1]

Although different scientific disciplines have been investigated (e.g., medicine, psychology, and sociology) and partially compared with each other (e.g., Kemper et al. 1996; Mutz et al. 2012), research on inter-rater reliability in an interdisciplinary field is scarce. One notable exception was the study by Langfeldt (2001). This study, however, applied mainly qualitative methods and did not provide quantitative indices of inter-rater reliability. In some other studies it is not entirely clear whether their analyses involved different disciplines (e.g., Herzog et al. 2005; Marsh et al. 2007). In any case, none of them has examined whether a match or a mismatch between reviewer discipline and paper discipline mattered with regard to the inter-rater reliability. This is, however, highly relevant with regard to the rapid increase of interdisciplinary research (Qiu 1992; van Noorden 2015). The corresponding challenge is that different disciplines may utilize different methodological approaches (e.g., quantitative vs. qualitative methods; hypothesis guided experiments vs. explorative descriptions; Platt 1964). Ultimately, such different approaches may be reflected in different standards for the evaluation of scientific contributions. Therefore, it was the first main objective of the present paper to investigate inter-rater reliability in an interdisciplinary scientific context. To this end, we analyzed proceedings submitted to an international interdisciplinary conference at the interface of social sciences (e.g., education), natural sciences (e.g., psychology) and technological sciences (e.g., information technology).[2]

Methods for assessing inter-rater reliability

Complex data structures are typical for the analysis of inter-rater reliability in peer-review contexts. The second major goal of our study was, therefore, to discuss and apply an adequate method for dealing with such highly complex data structures. Prior research has mainly relied on the classical multitrait-multimethod (MTMM) approach by Campbell and Fiske (1959). In such designs, each target needs to be measured by each method. In reviewing papers, that means that each paper submission should be rated by every reviewer. Consequently, papers and reviewers had to be fully crossed (see Putka et al. 2011). Such fully crossed designs, however, are rare in peer-review contexts.

[1] With regard to dichotomous nominal data (e.g., "accepted" vs. "rejected"), it should be noted that Cohen's Kappa (Cohen 1960), although often used, is by far not a reliable measurement of agreement, especially in cases of imbalanced marginal totals (e.g., see Baethge et al. 2013; Feinstein and Cicchetti 1990; Gwet 2008, 2014; Uebersax 1982–1983). Accordingly, Baethge et al. (2013) applied the agreement coefficient AC_1 for two raters proposed by Gwet (2008) to dichotomized reviewer evaluations and found a chance-corrected agreement estimation of .63. Cohen's Kappa statistic reached only a value of .16 in the study of Baethge et al. (2013).

[2] The investigated international conference took place within the last two decades. All of the reviewers were aware of the fact that others could access their evaluations of the papers. For the present study, the reviewers and their evaluations were fully anonymized and were analyzed in an aggregated way. Moreover, in order to protect the reviewers' privacy and anonymity as far as possible, we have omitted the mentioning of the name and the year of the conference. The same is applied to the conference proceedings.

Studies in which submissions to scientific journals have been analyzed are common (e.g., Bornmann and Daniel 2008b; Howard and Wilkinson 1998; Kirk and Franke 1997; Petty et al. 1999; Scarr and Weber 1978; Scott 1974). In such scenarios, editors typically prefer reviewers that are experts in the field of a submitted manuscript. As a consequence, the overall design is far from a fully-crossed design. Nested designs, in which "each target is rated by a unique, non-overlapping set of raters" (Putka et al. 2008, p. 960) might come closer to reality. However, even this is more an exception than the rule. In most cases, some reviewers evaluate more than one submission. This is especially true for conferences where a limited number of reviewers typically evaluate a subset of all submissions (e.g., Rubin et al. 1993). Thus, many practical scenarios only provide *ill-structured measurement designs* (ISMDs) in which "ratees and raters are neither fully crossed nor nested" (Putka et al. 2008, p. 960). Similar problems associated with ISMD are also known with regard to traditional nominal scale agreement coefficients (Cohen 1960; Fleiss 1971; see also Baethge et al. 2013; Uebersax 1982–1983).

Some previous studies have tried to solve the ISMD problem by randomly selecting a certain number of reviewers per target and "arbitrarily identifying those raters as 'Rater 1' and/or 'Rater 2' for each ratee" (Putka et al. 2011, p. 506; e.g., see Marsh and Ball 1989, p. 157; Petty et al. 1999, p. 192). This procedure is associated with various problems, however, such as possible identification problems, inappropriate solutions, and data loss (e.g., see Brown 2015; Eid 2000). Moreover, it raises the question about the meaning of "Rater i" and "Rater j" in such models. Most critical, however, is that researchers can subsequently come to different findings and conclusions simply as a result of differences in the rater selection and assignment (Putka et al. 2011).

There are also drawbacks with more traditional estimators of inter-rater reliability when it comes to ISMD scenarios. Putka et al. (2008) showed that for Pearson correlation approaches as well as for conventional intraclass correlation coefficients (ICCs; e.g., McGraw and Wong 1996; Shrout and Fleiss 1979), each of these methods may systematically underestimate inter-rater reliability in ISMD scenarios. The magnitude of this bias depends on the specific design conditions (Putka et al. 2008).

With reference to the generalizability (G) theory (e.g., Brennan 2001) and as an alternative to the strategies described above, Putka et al. (2008) offer the G-coefficient $G(q, k)$ as a "new interrater estimator that can be used regardless of whether one's design is crossed, nested, or ill-structured" (p. 977). Parameter k is the number of reviewers per paper. In ISMD scenarios, k can be estimated by the harmonic mean (HM) of the number of raters per rate. Parameter q scales the variance proportion that is related to the rater main effects (Putka et al. 2008, p. 963). In fully nested designs, parameter q equals $1/k$, in ISMD scenarios it is always smaller than 1/k, and in fully crossed designs it equals zero. With regard to the average ICCs (see Shrout and Fleiss 1979; McGraw and Wong 1996), $G(q = 1/k, k)$ equals ICC(1, k) and $G(q = 0, k)$ equals ICC(C, k). All these coefficients estimate the reliability of target scores which are derived by aggregating the ratings of k raters.

Analogously, setting k to the value of one in the equations described by Putka et al. (2008, p. 963), $G(q, 1)$ allows for estimating a single-rater reliability (D. J. Putka, personal communication, December 15, 2015). Single-rater coefficients refer to the reliability of a target score that is derived from only one rater. Such single-rater ICCs were used in Bornmann et al.'s (2010) meta-analysis for inter-rater reliability of journal peer reviews. Usually, a single-rater reliability should be smaller than $G(q, k \geq 2)$. This is analogous to the phenomenon that adding items to a test can, if certain assumptions are met, improve the test reliability (e.g., Raykov and Marcoulides 2011; Wirtz and Caspar 2002; Yousfi 2005).

Putka et al. (2008) recommend, however, that coefficients $G(q, 1)$ and $G(q, k \geq 2)$ should be separately estimated, that is, without using the Spearman-Brown prophecy formula.[3]

Based on a Monte Carlo simulation, Putka et al. (2008) conclude that "traditional estimators are either inappropriate or do not provide the most accurate result" (p. 980). They recommend the G-coefficient as "an attractive option relative to traditional methods" (Putka et al. 2008, p. 978). In our paper we therefore made use of this method, which perfectly applies to our ill-structured data set (see below).

Validity of peer-review ratings

Despite the fact that several studies have assessed inter-rater reliability for various different measures of paper quality (e.g., originality and relevance), little research has addressed the *dimensionality* of the quality judgements itself. This is important with regard to two major aspects. On the one hand, there are questions about the paper characteristics that should be considered in peer-review processes. On the other hand, it tackles the issue of whether reviewers are able to differentiate among different aspects of paper quality or whether they are instead driven by an overall, general impression (e.g., a halo-effect, Thorndike 1920; see also Pulakos et al. 1986). Despite the practical relevance and the fact that many journals (and conferences) provide multiple rating dimensions for peer-review evaluations, the issue of dimensionality itself has received much less scientific attention than inter-rater reliability (Marsh et al. 2008).

When a scientific contribution is evaluated in terms of quality, the question arises as to how quality can be defined. This becomes even more important in interdisciplinary contexts, as scientific disciplines may differ substantially in what they value. At the same time, it becomes relevant whether the instruments that are employed (e.g., rating scales) indeed measure the hypothetical construct that was targeted (e.g., originality). It is also relevant whether multiple measures actually assess different constructs or whether these overlap to an extent that they could be subsumed under the same label. This is the matter of *construct validity* (Messick 1995; Strauss and Smith 2009), which is the "overarching principle of validity, referring to the extent to which a psychological measure in fact measures the concept it purports to measure" (Brown 2015, p. 187).

Two concepts of construct validity are important in this context: *convergent validity* and *discriminant validity* (Campbell and Fiske 1959; see also Brown 2015). Convergent validity means that measures (different methods or indicators) of the same construct should be highly interrelated. Discriminant (or divergent) validity means that measures of different constructs should not be interrelated. Hence, an important question is whether the quality of a scientific contribution is a unidimensional construct that can be summarized in one global evaluation score. The alternative view would argue that quality comprises multiple dimensions which should be considered separately. In other words, if reviewers are asked to rate the quality of a scientific contribution on various dimensions (e.g., relevance, soundness, and novelty), the question is, whether these dimensions indeed represent distinct constructs, which would suggest a multidimensional structure, or whether they all converge, suggesting a unidimensional structure.

[3] The Spearman–Brown prophecy formula can be used to predict the reliability of a test or target score after increasing (or decreasing) the corresponding number of items, observations, or raters. It can also be used to determine the necessary number of items, observations, or raters for obtaining a certain reliability value (e.g., see Shrout and Fleiss 1979, p. 426).

Naturally, the answer to this question depends upon the rating dimensions employed. Unfortunately, to date there is hardly any universal consensus on which dimensions a review should be based (for some elaborations, see Chase 1970; Hemlin and Montgomery 1990). Cicchetti (1991) identified two aspects which are broadly accepted: the importance of the study to the field and the perceived adequacy of the research design. Journals, conferences, and funding agencies, however, often ask their reviewers to evaluate papers on many more rating dimensions, as can be seen in the studies on inter-rater agreement (e.g., Cicchetti and Conn 1976; Marsh and Ball 1989; Montgomery et al. 2002; Rubin et al. 1993; Scott 1974; Whitehurst 1983).

Unfortunately, however, only one of these studies has examined the dimensional structure of these ratings as well as other aspects of construct validity. Marsh and Ball (1989) found modest support for the distinctiveness of the four rating dimensions that they had extracted from a 21-item instrument (research methods, relevance to readers, writing style and presentation clarity, and significance/importance). Their analysis favored the multidimensional model over an alternative unidimensional model. Similarly, Petty et al. (1999) reported a better fit when a model was based on five dimensions (literature, theory, methodology, importance, and recommendation) compared to an alternative unidimensional model. On the other hand, Cicchetti and Conn (1976) found that certain single dimensions (originality, design-execution, importance, and overall scientific merit) correlated strongly with an overall score (.55–.96). However, they did not directly compare a multidimensional model with a unidimensional one.

It is clear there is still little evidence with regard to the question of whether the use of multiple dimensions in fact adds something unique to a general evaluation. Moreover, there is no evidence at all when it comes to the interdisciplinary context. In our study, we have addressed this gap and analyzed reviews of submissions to an interdisciplinary conference. Here, reviewers provided both an overall evaluation and ratings with respect to four specific rating dimensions (relevance, novelty, significance, and soundness). These dimensions have been employed or suggested in previous publications as well (e.g., Beyer et al. 1995; Campion 1993; Cicchetti and Conn 1976; Gilliland and Cortina 1997; Gottfredson 1978).

Another way of investigating the validity of the peer-review ratings is to look at the potential of the rating dimensions to predict the citation rate of the accepted papers. For example, Bornmann and Daniel (2008a) showed that papers accepted by a high-impact chemistry journal would get more citations than papers that were rejected and published elsewhere. Opthof et al. (2002) showed a positive and significant relationship between reviewers' priority recommendations and papers' citation count for three years after publication in a cardiology journal. A positive relationship between peer-review scores and citations rates also exists in the field of research project grants (Li and Agha 2015). However, in another medical journal, Baethge et al. (2013) did not find a significant relationship between reviewer recommendations and citation rate. A possible explanation could be that "accepted versions of manuscripts differ considerably from submitted versions" (Baethge et al. 2013, p. 6). In any case, it seems worth investigating the *predictive criterion validity* of different rating dimensions in an interdisciplinary peer-review context. For this investigation, we took papers into consideration that had been published in the conference proceedings (see footnote 2).

Taken together, our paper has the following three major goals: (1) analyze inter-rater reliability in an interdisciplinary context, across all paper-reviewer-combinations and separated for same-discipline versus different-discipline reviewers, (2) apply state-of-the-art methods of analysis to account for ill-structured measurement design, and (3) examine the dimensionality and validity of the different rating dimensions.

Methods

Our study analyzed the reviews of conference proceeding papers submitted to an international interdisciplinary conference (see footnote 2). The conference takes place annually and is interdisciplinary, with researchers from computer science, education, psychology, and communication science. It is of medium size (about 200–300 participants, about 100–200 submissions) and has an acceptance level of less than 30%. As such, the conference is a competitive, typically mid-size conference with an interdisciplinary topic. Papers which are accepted appear in a Springer book series (see footnote 2).

Papers and reviewers

A total of one hundred and seventy-four submissions (including keynotes) were listed in the conference system. For our analyses, we considered only those $n = 145$ submissions which had been rated by at least two reviewers. From these, a total of $n_{\mathrm{ap}} = 82$ submissions that had been accepted were later published in the conference proceedings (see footnote 2). Overall, $m = 130$ reviewers conducted reviews of the $n = 145$ submissions. This resulted in a total of $v = 443$ reviews.

Due to the fact that reviewers could opt for the papers they would like to review, the number of reviewers per paper varied. Each selected paper, on average, received $M = 3.06$ reviews ($SD = 0.40$), with a minimum of 2 and a maximum of 5 reviews. Each of the corresponding reviewers ($m = 130$) reviewed, on average, $M = 3.41$ papers ($SD = 1.90$), with a minimum of 1 and a maximum of 6 reviewed papers.

Papers as well as reviewers were categorized by two independent raters into one of three different disciplines: (a) psychological-experimental, (b) empirical-social, or (c) information technological (Cohen's Kappa = .86, disagreements were solved by discussion). Of all the papers, $n_{\mathrm{psy.exp}} = 14$ (9.7%) were regarded as psychological-experimental, $n_{\mathrm{emp.soc}} = 51$ (35.2%) were regarded as empirical-social, and $n_{\mathrm{it}} = 80$ (55.2%) were regarded as information technological. A similar distribution appeared on the reviewers' side: $k_{\mathrm{psy.exp}} = 13$ (10.0%) were considered as psychological-experimental, $k_{\mathrm{emp.soc}} = 32$ (24.6%) as empirical-social, and $k_{\mathrm{it}} = 85$ (65.4%) as information technological. Altogether, $n_{\mathrm{both}} = 93$ (64.1%) papers were reviewed by both same-discipline and different-discipline reviewers, $n_{\mathrm{same}} = 26$ (17.9%) papers were reviewed only by same-discipline reviewers, and $n_{\mathrm{diff}} = 26$ (17.9%) papers were reviewed only by different-discipline reviewers. Again, a similar pattern appeared on the reviewer's side: $k_{\mathrm{both}} = 72$ (55.4%) reviewed both same-discipline and different-discipline papers, $k_{\mathrm{same}} = 31$ (23.8%) reviewed only same-discipline papers, and $k_{\mathrm{diff}} = 27$ (20.8%) reviewed only different-discipline papers.

Units of analysis

The primary units of analysis were reviews and, on an aggregate level, papers. *Review scores* and *paper scores* were calculated for the following five rating dimensions: (a) overall evaluation, (b) relevance to the conference (c) novelty, (d) significance, and (e) soundness. For each paper and each dimension, the paper score was estimated by averaging the ratings (a) of all reviewers, (b) only of same-discipline reviewers (same-discipline paper scores), and (c) only of different-discipline reviewers (different-discipline paper scores). Thus, the paper score for a certain paper for a certain dimension was the mean of the ratings over all reviewers who had rated that paper (for the concept of target and rater scores, see, for example, Hönekopp 2006; Hönekopp et al. 2006).

Measures and variables

A review form guided reviewers in their evaluations. They were asked to provide a detailed review including justification for their scores. They were urged to be constructive and to answer first some open-ended questions (see Table 1). More important to the purpose of this study, reviewers were then asked to fill out several rating scales (see Table 1).

Our analyses focused on the following five variables: overall evaluation, relevance, novelty, significance, and soundness. The values for the overall evaluation ranging from −2 to 2 and were recoded by adding the value of 3 to the range from 1 to 5. In order to analyze only genuine evaluations, we eliminated all of the ratings which fell into the

Table 1 Guide for reviewers' evaluations

Item	Response format
What is the contribution of the paper [to the conference's topic]; does the paper make this contribution clear?[a]	Open
What are the strong points of this work?	Open
What are the weak points of this work?	Open
Are the major claims and conclusions substantiated?	Open
Is the paper clear, explicit, and well-organized? Can you suggest improvements (e.g., body, title, abstract)?	Open
Does the paper adequately refer to related work? Suggest further references, if appropriate.	Open
Overall evaluation	−2 (*strong reject*) to 2 (*strong accept*)
Confidence	0 (*null*) to 4 (*expert*)
Relevance[a]	1 = *not relevant*, 2 = *cannot judge relevance*, 3 = *some relevance*, 4 = *highly relevant*
Novelty	1 = *not novel*, 2 = *cannot judge novelty*, 3 = *some novelty*, 4 = *highly novel*
Significance[a]	1 = *not significant*, 2 = *cannot judge significance*, 3 = *some aspects can be significant*, 4 = *highly significant*
Soundness[a]	1 = *unacceptable—major flaws*, 2 = *cannot judge soundness*, 3 = *good, but some flaws*, 4 = *excellent*
Recommended category	1 = *not suitable for the conference*[a], 2 = *recommend doctoral consortium submission*, 3 = *recommend workshop submission*, 4 = *poster*, 5 = *full paper*
Best paper candidate (top 5% of the papers)	1 = *no*, 2 = *yes*

[a] References to the conference were removed (see footnote 2)

cannot judge response category for the variables relevance, novelty, significance, and soundness. Thus, for these variables, the analyses are based on recoded 3-point scales (e.g., 1 = *not relevant*, 2 = *some relevance*, 3 = *highly relevant*).[4] Due to the resulting missing values, this step resulted in an even more complicated data structure. Nevertheless, the analyses we applied were still appropriate for the remaining data.

A further aim of the study was to estimate the relationship between the rating dimensions and the citation rate of the papers published in the subsequent conference proceedings (see footnote 2). From the $n = 145$ submissions which had been rated by at least two reviewers, a total of $n_{\mathrm{ap}} = 82$ submissions were published. We counted the number of citations per paper by searching all databases of the Thomson Reuters Web of Science citation index for each published conference paper for a time frame of roughly three years. Thus, the citation window includes the first three years after publication (in addition to the end of the year in which the conference proceedings had been issued). This time frame is comparable to citations windows which were used in other studies (e.g., see Bornmann and Daniel 2008a).

Analyses

For estimating the G-coefficients, it was necessary to estimate the following variance components: (a) the paper main effect, (b) the reviewer main effect, and (c) the combination of error variance with paper × reviewer interaction effects. We used IBM SPSS Statistics Version 20.0 (2011) for this task. Variance components were estimated with the restricted (or residual) maximum likelihood (REML) estimator (e.g., O'Neill et al. 2012; Putka et al. 2008; see also Searle et al. 1992). Estimation of $G(q, k)$ also requires estimates of the parameters k and q.

All other analyses were conducted with Mplus 7.3 (Muthén and Muthén 2012). With one exception, models and parameters were estimated with the robust maximum likelihood (MLR) estimator implemented in Mplus (for advantages of using a sandwich estimator, see Muthén and Muthén 2012; White 1980; Yuan and Bentler 2000). For significance testing purposes, several likelihood-ratio tests (LRTs) were also conducted. The reason for this choice was that, especially in small samples, the LRT is superior to the commonly used Wald test (e.g., Enders 2010). It must be noted, however, that LRTs based on the MLR estimator need to be corrected by special scaling factors (www.statmodel.com/chidiff.shtml; see also Enders 2010, p. 149). Missing data were dealt with by the full information maximum likelihood (FIML) method. This method uses all of the available information in the data (e.g., Enders 2001, 2010; Rubin 1976).

Beside the MLR estimator described above, a confirmatory factor analysis (CFA) was estimated with the *Bayes* estimator (e.g., see Kaplan and Depaoli 2013; Muthén 2010; van de Schoot et al. 2014; Zyphur and Oswald 2015). Missing data issues were dealt with in a similar fashion as with the FIML under the missing at random (MAR) assumption (Asparouhov and Muthén 2010; Enders 2010). For the Bayes estimation procedure, non-informative priors were used. The medians of the posterior distributions acted as point estimates of the parameters. The posterior distributions were estimated with the Markov

[4] It seems noteworthy that upon inspection, the "cannot judge" responses did *not* indicate, at least after the correction proposed by Holm (1979), that reviewers who came from other disciplines than the paper (different-discipline reviewers) used this category more often than same-discipline reviewers (all Holm-adjusted $ps > .085$; see Online Resource 1).

chain Monte Carlo (MCMC) Gibbs sampler algorithm (e.g., see Brown 2015; Kaplan and Depaoli 2013; van de Schoot et al. 2014).[5]

Consistent with the Bayesian philosophy, the Bayes estimator does not produce *p* values but Bayesian credibility intervals (CIs), which are not necessarily symmetric. A 95% CI means that there is "a 95% probability that the population value is within the limits of the interval" (van de Schoot et al. 2014, p. 844). Furthermore, if the CI does not contain the value zero, the corresponding parameter can be interpreted as significant according to classical frequentist null hypothesis testing (van de Schoot et al. 2014).

To address the multiple testing problem (e.g., Shaffer 1995), the raw *p* values for a given meaningful family of tests (e.g., all coefficients in a correlation matrix) were adjusted. In almost all cases, this was done by the Holm-procedure (Holm 1979). For such a test family, the conditional probability of committing a type I error at least once, that is, the multiple (familywise) significance level, is restricted to .05. With regard to Bayesian analyses, the multiple testing problem was addressed by using more conservative 99% CIs instead of 95% CIs.

Results

In the following section, we present our results with regard to the key aspects of inter-rater reliability and validity. First, we estimated the inter-rater reliabilities of the five rating dimensions, taking the ill-structured measurement design and the interdisciplinary context into account. Then, we investigated the dimensionality and the construct validity of the rating dimensions. Finally, we examined the predictive criterion validity of the rating dimensions in order to analyze whether the rating dimensions had the potential to predict the citation rate of accepted papers. In each of these steps, we differentiated between ratings that came from same-discipline reviewers and those that came from different-discipline reviewers.

Inter-rater reliability

Here, we primarily focused on the estimation of single-rater reliabilities $G(q_k, k = 1)$ for each dimension. Single-rater reliabilities were chosen, as they are comparable with the coefficients reported in the meta-analysis of Bornmann et al. (2010). Furthermore, for each dimension, we differentiated between paper scores that were based on the ratings of all reviewers and paper scores that resulted as a function of the match/mismatch between the discipline of the paper and the discipline of the reviewer.

[5] Two chains per model were used whereby a minimum of 30,000 iterations and a maximum of 200,000 iterations were specified for each chain. The convergence criterion was repeatedly assessed each time after 100 iterations, based on the final half of all iterations per chain. After reaching the criterion, the first half of all the iterations were dropped (burn-in phase). The posterior distributions were constructed with the remaining post-burn-in iterations (Brown 2015; Muthén and Muthén 2012). For determining the convergence, the Gelman-Rubin convergence criterion (Muthén and Muthén 2012; Gelman and Rubin 1992) was used for determining convergence. The parameter *b* in the formula of the Potential scale reduction (PSR) was set at the value of 0.001, which defines a very strict criterion (Brown 2015; Gelman et al. 2013; Muthén and Muthén 2012; van de Schoot et al. 2014; Zyphur and Oswald 2015).

Across all reviewers

Table 2 summarizes the data on which the G-coefficient estimations were based. Point estimates for the variance components are listed in Online Resource 2 and estimates for q and k are listed in Online Resource 3. Based on these values, the G-coefficients were calculated by applying the formula described in Putka et al. (2008). The single-rater reliabilities $G(q_k, k = 1)$ based on the ratings of all reviewers are shown in Table 3. As already mentioned, a single-rater reliability coefficient estimates the reliability of a paper score as if this score was based on only a single reviewer's evaluation.

The estimated single-rater reliability of the overall evaluation was .21 and the values for the other rating dimensions ranged from .17 to .28. Confidence intervals were estimated based on the Fisher z-transformation and on the corresponding back-transformation for ICCs (Fisher 1934; see McGraw and Wong 1996; Putka 2002). None of the Holm-adjusted confidence intervals (Holm 1979; see Altman and Bland 2011; Serlin 1993) contained the value of zero. Hence, agreement was significantly above chance. Nevertheless, agreement was low for all dimensions (below .40; e.g., see Cicchetti 1994, p. 286).

To compare our results for overall evaluation with those from Baethge et al. (2013; see footnote 1), we collapsed the five response categories into two categories: (a) *weak* or *strong acceptance* versus (b) all categories below weak acceptance (*borderline*, *weak reject*, and *strong reject*). Then, we estimated the AC_1 coefficient for multiple raters (Gwet

Table 2 Number of reviews, number of papers, number of different reviewers, and the harmonic mean (HM) of the number of reviewers per paper

Rating dimension	Class of reviewers	Number of reviews	Number of reviewers	Number of papers	k (HM)
Overall	All reviewers	443	130	145	3.00
	Same-discipline	216	99	119	1.48
	Different-discipline	227	103	119	1.58
Relevance	All reviewers	402	128	144	2.63
	Same-discipline	203	94	115	1.45
	Different-discipline	199	99	114	1.45
Novelty	All reviewers	344	120	143	2.11
	Same-discipline	168	86	107	1.35
	Different-discipline	176	89	109	1.34
Significance	All reviewers	369	124	143	2.33
	Same-discipline	184	89	110	1.39
	Different-discipline	185	96	108	1.41
Soundness	All reviewers	325	123	142	1.92
	Same-discipline	153	83	102	1.27
	Different-discipline	172	92	109	1.31

k = harmonic mean (HM) of the number of reviewers per paper. Differences in the frequencies between the rating dimensions and between the reviewer groups are due to the exclusion of the *cannot judge* responses

Table 3 Estimated single-rater reliabilities $G(q, k = 1)$ and confidence intervals as a function of the class of reviewers

Rating dimension	Kind of reviewers	$G(q_k, 1)$	95% CI	Holm-adj. CI
Overall evaluation	All reviewers	.21	[.10, .31]	[.07, .35]
	Same-discipline	.23	[.00, .46][a]	[.00, .52][a]
	Different-discipline	.18	[.00, .40][a]	[.00, .44][a]
Relevance	All reviewers	.21	[.09, .32]	[.06, .35]
	Same-discipline	.13	[.00, .39][a]	[.00, .46][a]
	Different-discipline	.37	[.10, .58]	[.01, .62]
Novelty	All reviewers	.28	[.13, .42]	[.09, .45]
	Same-discipline	.20	[.00, .48][a]	[.00, .55][a]
	Different-discipline	.35	[.01, .59]	[.00, .63][a]
Significance	All reviewers	.17	[.04, .30]	[.02, .32]
	Same-discipline	.14	[.00, .42][a]	[.00, .49][a]
	Different-discipline	.46	[.19, .65]	[.09, .70]
Soundness	All reviewers	.21	[.04, .37]	[.04, .37]
	Same-discipline	.13	[.00, .48][a]	[.00, .56][a]
	Different-discipline	.39	[.05, .63]	[.00, .66][a]

Variance components were estimated on the base of (a) the ratings of all reviewers, (b) the ratings of reviewers from the same discipline as the paper, and (b) the ratings from reviewers from different disciplines. Estimation of the reliability coefficients is described in Putka et al. (2008). Confidence intervals were estimated based on the Fisher z-transformation for intraclass correlation coefficients (Fisher 1934; see McGraw and Wong 1996; Putka 2002). This procedure was conducted three times (see the remarks above). For each class of reviewers a confidence interval (CI) adjustment was conducted based on the method proposed by Holm (1979) for $m = 5$ tests (see Serlin 1993; Altman and Bland 2011). The mean single-rater reliabilities based on all-reviewers, on same-discipline reviewers, and on different-discipline reviewers were .22, .16, and .35, respectively (estimated by the procedure of Bornmann et al. 2010, p. 3)

[a] Negative lower confidence bounds were set to zero

2008, 2014; see Fleiss 1971) as well as AC_1 coefficient and Cohen's Kappa as calculated by Baethge et al. (2013) for more than two raters. The corresponding values were .14, .15, and .15, respectively. These chance-corrected estimates underscored the conclusions already drawn from the G-coefficients.

As a function of the match/mismatch between paper and reviewer disciplines

Table 3 also displays the single-rater reliabilities $G(q_k, k = 1)$ based (a) on the ratings of reviewers from the same discipline as the paper and (b) on the ratings from reviewers from different disciplines. The estimated single-rater reliabilities for the overall evaluation were again poor (same-discipline reviewers: .23, different-discipline reviewers: .18). "Fair" agreement (Cicchetti 1994) was found only for some of the single-rater reliabilities, based on the ratings from the different-discipline reviewers (.35–.46). In contrast, the agreement for the same-discipline reviewers was rather poor (.13–.20) and even non-significant. In any case, single-rater reliabilities were generally low—regardless of whether reviewers matched or did not match the papers' scientific discipline.

In order to test for significant differences between same-discipline and different-discipline reviews, we estimated confidence intervals for the differences (a) by the method of

Table 4 Two-sided 95% confidence intervals for the differences between single-rater reliabilities from same-discipline versus different discipline reviewers constructed (a) by the method of Donner (1986) for independent ICCs and (b) by the method of Ramasundarahettige et al. (2009) for dependent ICCs

Rating dimension	Difference	95% CI (Donner 1986)	95% CI (Ramasundarahettige et al. 2009)
Overall	0.05[a]	[.00, .36][c]	[.00, .39][c]
Relevance	0.25[b]	[.00, .57][c]	[.00, .63][c]
Novelty	0.16[b]	[.00, .53][c]	[.00, .61][c]
Significance	0.32[b]	[.00, .67][c]	[.00, .73][c]
Soundness	0.26[b]	[.00, .66][c]	[.00, .81][c]

Estimations were based on the Fisher *z*-transformation for ICCs (Fisher 1934) and on the corresponding back-transformation to the original scale

[a] Reliability coefficient derived from same-discipline reviewers acts as the minuend

[b] Reliability coefficient derived from different-discipline reviewers acts as the minuend

[c] Negative lower confidence bounds were set to zero

Donner (1986, p. 76) for independent ICCs and (b) by the method of Ramasundarahettige et al. (2009, p. 1043, pp. 1045–1046) for dependent ICCs (for another approach:[6]). All estimations were based on the Fisher *z*-transformation and its inverse (Fisher 1934). The confidence intervals for the differences are presented in Table 4. None of the comparisons yielded a significant difference, regardless of whether the two coefficients per rating dimension were defined as independent or as dependent. That is, although all same-discipline single-rater reliabilities were insignificant, and in most cases descriptively smaller than the corresponding different-discipline reliabilities, same- and different-discipline reliabilities did not significantly differ from one another.

In addition to the single-rater reliabilities we also estimated the *k*-rater reliabilities $G(q_{k=1}, k = \mathrm{HM})$, which were based (a) on the ratings from all reviewers, (b) on the ratings from same-discipline reviewers, and (c) on the ratings from different-discipline reviewers (see Fig. 1). Although the overall picture looks similar, it is obvious that the *k*-rater coefficients were larger than their corresponding single-rater counterparts (see Table 3). This is not surprising, because this phenomenon is analogous to the classical test theory, where more items usually result in higher reliability coefficients (e.g., see Yousfi 2005). Based on the ratings of all reviewers, reliability values can be regarded as “fair” (values above .40; Cicchetti 1994) for overall evaluation, relevance, and novelty. With regard to the *k*-rater reliabilities based on the ratings from different-discipline reviewers, coefficients can be regarded as “fair” for the dimensions relevance, novelty, significance, and soundness. In contrast, reliability values based on the ratings from same-discipline reviewers were poor for all dimensions (see Fig. 1). Finally, there was no evidence for a severity or a leniency bias in either the same-discipline sample or the different-discipline sample of our study.[7]

[6] We have also compared the variance components (e.g., the paper component) for each rating dimension between the same-discipline and the different-discipline paper × reviewer combinations as suggested by O’Neill et al. (2012). However, the LRTs based on the REML log-likelihoods in our study yielded several negative Chi square statistics which cannot be regarded as trustworthy (for a similar phenomenon in another context, see Satorra and Bentler 2010).

[7] We found no significant differences between (a) the paper scores from the same-discipline reviewers and (b) the paper scores from the different-discipline reviewers (all Holm-adjusted *p*s $> .147$; see Online Resource 4).

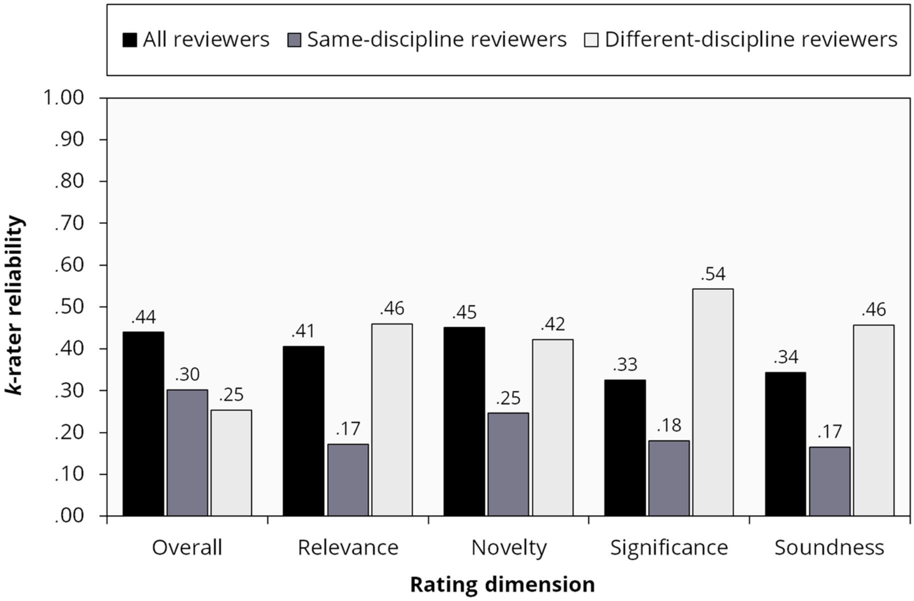

Fig. 1 Estimated *k*-rater reliabilities $G(q_{k=1}, k = \text{HM})$ for the rating dimensions

In sum, the inter-rater reliabilities were generally low. Descriptively, the pattern was contrary to what we expected. That is, the reliabilities of paper scores based on ratings from same-discipline reviewers were descriptively lower than those based on ratings from different-discipline reviewers. These descriptive differences, however, did not reach significance.

Construct validity

We conducted several analyses to investigate the dimensionality and validity aspects of reviewers' evaluations, at the same time taking scientific discipline into account. First, we estimated the manifest correlations of the paper scores based on different rater subgroups. Second, we conducted an exploratory factor analysis, which is followed by a more complex confirmatory factor analysis in the CT-C(M-1) framework (Eid 2000; Eid et al. 2003). Then, we examined whether a multidimensional model or a unidimensional model would better fit the data. Finally, we assessed the criterion validity of the rating dimensions for predicting the citation rate of accepted papers.

Manifest correlations

In a first step we determined the correlations between the paper scores based on the same-discipline reviewers and the paper scores based on the different-discipline reviewers. Such correlations between different paper scores can reveal, for example, whether different rater groups applied the same criteria and came to the same conclusions. Large positive correlations would imply that papers with comparatively high (low) positions in the ranking order based on the ratings given by one rater group should also have high (low) positions in the ranking order based on the ratings given by the other rater group (e.g., Henss 1992). In

Table 5 Correlations between the paper scores based on the ratings of the same-discipline reviewers and the paper scores based on the ratings of the different-discipline reviewers (monotrait-heteromethod correlations)

Rating dimension	Covariance coverage	Estimated correlation	*SE*	Adjusted *p* (raw *p*)
Overall	93/145 = .64	.23	0.08	.017 (.004)
Relevance	85/145 = .59	.18	0.10	.221 (.074)
Novelty	73/145 = .50	.25	0.08	.008 (.002)
Significance	75/145 = .52	.07	0.09	.866 (.433)
Soundness	69/145 = .48	.05	0.11	.866 (.639)

Papers are the units of analysis. Correlations were estimated by MLR (maximum likelihood with robust standard errors) with the full information maximum likelihood (FIML) method on the base of $n = 145$ papers. A p value adjustment was conducted by the method proposed by Holm (1979) for $m = 5$ tests

the MTMM approach by Campbell and Fiske (1959, p. 82) such correlations are termed *monotrait-heteromethod* values and were used to examine the convergent validity.

As Table 5 reveals, however, only correlations for the overall evaluation and for novelty reached significance. For the other three rating dimensions there were only small and non-significant relationships. It is noteworthy that the significant correlations obtained were only moderately positive (e.g., of medium effect size, Cohen 1988). In other words, reviewer groups (same-discipline vs. different-discipline) did not agree (much) on paper quality. This is evidence against a corresponding convergent validity, where the same-discipline paper scores and the different-discipline paper scores would have been expected to measure the same construct.

Another picture appears if one looks at the correlations between the different rating dimensions within each reviewer group (*heterotrait-monomethod* values; see Campbell and Fiske 1959) and across all reviewers. Here, it is obvious that all correlations were (mostly) large, positive and significant (see Table 6). That is, a paper with a high (low) score on one rating dimension also has relatively high (low) scores on the other rating dimensions. This is evidence for rather low discriminant validity. Hence, reviewers (same-discipline and different-discipline alike) did not differentiate much in their assessments of the different paper attributes.

Exploratory factor analysis

In a next step, we conducted an exploratory factor analysis (EFA) to examine the ostensible factor structure of the paper scores that resulted from same-discipline and different-discipline ratings. Input variables were 10 variables, 5 each in these categories: (a) the same-discipline paper scores for each rating dimension and (b) the different-discipline paper scores for each rating dimension (see Table 7). Papers ($n = 145$) were the units of analysis. The EFA was estimated by MLR with the full information maximum likelihood (FIML) method. The Kaiser–Meyer–Olkin (KMO) measure of sampling adequacy of the correlation matrix, which should not be smaller than .50, reached a value of .77 (Dziuban

Table 6 Correlations between rating dimensions within both reviewer groups (heterotrait-monomethod correlations) and across all reviewers

Rating dimension	Class of reviewers	1	2	3	4	5
1 Overall evaluation	All reviewers	–	.65*	.64*	.65*	.63*
	Same-discipline	–	.53*	.54*	.58*	.67*
	Different-discipline	–	.67*	.66*	.69*	.67*
2 Relevance	All reviewers	.65*	–	.50*	.57*	.44*
	Same-discipline	.53*	–	.41*	.35*	.29*
	Different-discipline	.67*	–	.52*	.66*	.41*
3 Novelty	All reviewers	.64*	.50*	–	.56*	.41*
	Same-discipline	.54*	.41*	–	.57*	.44*
	Different-discipline	.66*	.52*	–	.68*	.45*
4 Significance	All reviewers	.65*	.57*	.56*	–	.40*
	Same-discipline	.58*	.35*	.57*	–	.49*
	Different-discipline	.69*	.66*	.68*	–	.41*
5 Soundness	All reviewers	.63*	.44*	.41*	.40*	–
	Same-discipline	.67*	.29*	.44*	.49*	–
	Different-discipline	.67*	.41*	.45*	.41*	–

Papers are the units of analysis. Correlations were estimated by MLR (maximum likelihood with robust standard errors) with the full information maximum likelihood (FIML) method on the base of (a) the ratings of all reviewers ($n = 145$ papers), (b) the ratings of reviewers from the same discipline as the paper (available for $n_{\text{same}} = 119$ papers), and (c) the ratings from reviewers from different disciplines (available for $n_{\text{diff}} = 119$ papers). A p value adjustment was conducted by the method proposed by Holm (1979) for $m = 10$ tests. This procedure was conducted three times (see the remarks above)

* $p_{\text{adjust}} < .05$, two-tailed, with a multiple (familywise) significance level of .05

and Shirkey 1974; Kaiser 1970; Kaiser and Rice 1974; see also Field 2009; Hutcheson and Sofroniou 1999). The measures of sampling adequacy (MSAs) for each single variable were also greater than .50 (see Table 7).

The number of factors were determined by the scree plot of the eigenvalues (based on the non-reduced correlation matrix: 3.90, 2.46, 0.77, 0.71, 0.57, 0.54, 0.37, 0.29, 0.20, 0.18) which is shown in Online Resource 5 (Cattell and Jaspers 1967; but see also Cattell 1966). Applying this criterion, the EFA suggests a two-factor structure. This decision was clearly supported by a more accurate parallel analysis (Horn 1965) and by the criterion that only those eigenvalues should be considered that lie above the 95th percentile of randomly generated eigenvalues (see Hayton et al. 2004; 1000 random data sets were generated; see Online Resource 5). After the extraction of two factors, the oblique Geomin rotation (Yates 1987) was applied. Both factors correlated significantly at .29 ($p = .003$).

As Table 7 shows, all of the same-discipline indicators only loaded significantly on factor F-I and all of the different-discipline indicators only loaded significantly on factor F-II. Therefore, it seems that this two-factor solution was an example of artificial method factors which, for instance, are typical of applications where positively and negatively formulated (reversed) items create their own factors (e.g., Brown 2015).

Again, these results provide evidence against the convergent validity of the paper scores from both reviewer groups, as well as against the discriminant validity of the five dimensions themselves. After all, the exploratory factor analysis did not reveal rating

Table 7 Measure of sampling adequacy (MSA), communality values (COM), and factor loadings for an exploratory factor analysis (EFA)

Rating dimension	Kind of reviewers	MSA	COM	Factor I	Factor II
Overall	Same-discipline	.72	.75	.87*	.00
Relevance	Same-discipline	.71	.32	.54*	.06
Novelty	Same-discipline	.83	.44	.66*	.03
Significance	Same-discipline	.79	.48	.69*	.00
Soundness	Same-discipline	.67	.54	.77*	−.17
Overall	Different-discipline	.78	.83	.01	.91*
Relevance	Different-discipline	.83	.56	.12	.71*
Novelty	Different-discipline	.86	.54	.01	.73*
Significance	Different-discipline	.75	.64	−.12	.83*
Soundness	Different-discipline	.74	.45	−.14	.70*

Papers are the units of analysis. Correlations and the EFA were estimated by MLR (maximum likelihood with robust standard errors) with the full information maximum likelihood (FIML) method on the base of $n = 145$ papers with Mplus (Muthén and Muthén 2012). The Kaiser–Meyer–Olkin (KMO) measure of sampling adequacy of the correlation matrix reached an adequate value of .77 (Dziuban and Shirkey 1974; Kaiser 1970; Kaiser and Rice 1974; see also Field 2009; Hutcheson and Sofroniou 1999). MSA and KMO values were estimated with the package psych (Revelle 2016) for R (R Core Team 2016). After factor extraction, the oblique geomin rotation method (Yates 1987) was performed. Both geomin factors are significantly correlated at .29 ($p = .003$). Thus, factor loadings (out of the pattern matrix) in the last two columns should be understood as partial regression coefficients (e.g., Brown 2015). A p value adjustment was conducted by the method proposed by Holm (1979) for $m = 20$ tests

* $p_{\text{adjust}} < .05$, two-tailed, with a multiple (familywise) significance level of .05

dimensions as distinct factors, as would have been expected if the rating dimensions had had a high convergent as well as discriminant construct validity. It is also notable that the overall evaluation indicators had the highest loadings in comparison to the other rating dimensions (see Table 7), which makes the overall evaluation a marker variable that is most essential for the interpretation of each factor.

Confirmatory factor analysis

In a next step we tested the dimensionality and method specificity of the ratings in a more elaborated way with a special kind of confirmatory factor analysis (CFA): a *correlated trait–correlated* (*method minus one*) [CT-C(M-1)] model (Eid 2000; Eid et al. 2003). The corresponding model structure is illustrated in Fig. 2.

Here, each rating dimension was given an exclusive factor with two indicators: (a) the paper scores that were estimated by averaging the ratings of same-discipline reviewers and (b) the paper scores that were estimated by averaging the ratings of different-discipline reviewers. The usage of two aggregated indicators per factor is comparable to the approach of using item parcels (e.g., test halves) as manifest indicators (e.g., see Brown 2015).

Additionally, an asymmetric method factor (MF) was specified, as was proposed by Eid (2000). The MF was uncorrelated with the other factors and only indicators from one method loaded on the MF. The method without loadings had to be interpreted as the reference method. That is, one method took on the reference role and acted as a comparison standard, whereby the method specificity of the other method was caught by the

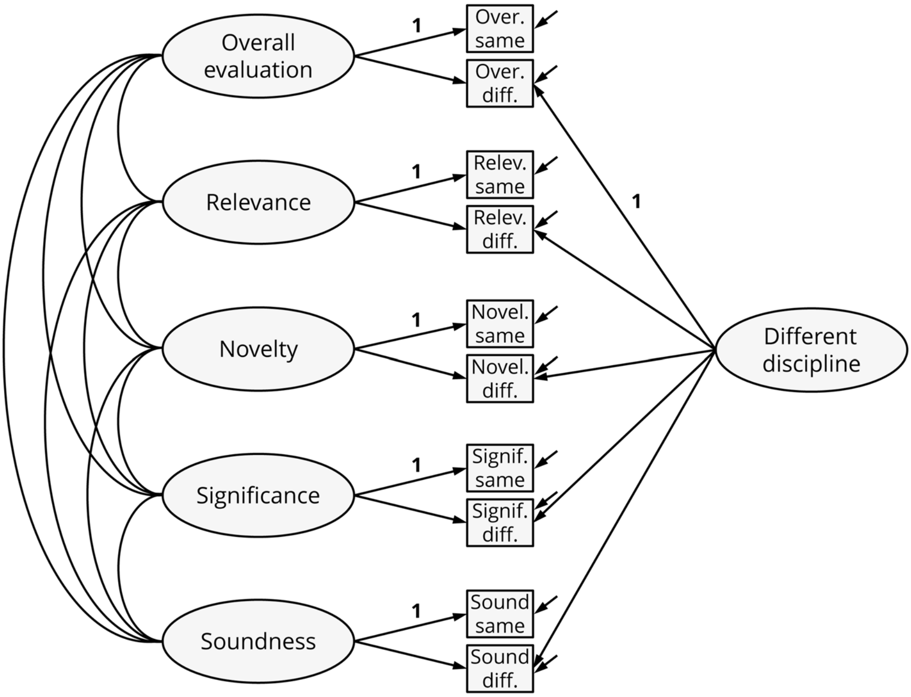

Fig. 2 A correlated trait–correlated (method minus one) [CT-C(M-1)] model (Eid 2000; Eid et al. 2003) in which the same-discipline paper scores serve as reference method

MF. Because such a model is therefore asymmetric and because there was no natural standard method, *two* models were estimated in which the roles of the same-discipline and the different-discipline scores were reversed. For scaling the latent variables and for identification purposes, one unstandardized loading per factor was fixed to the value of one (see Fig. 2).

However, such models can be prone to convergence problems (Eid 2000; Eid et al. 2003, p. 60). Therefore, and because of the relatively small sample size, the Bayes estimator was used. This estimator also has the advantage that implausible values (e.g., negative variances) are impossible (van de Schoot et al. 2014; Zyphur and Oswald 2015). Accordingly, the two models were estimated with the Bayes estimator by Mplus (Muthén and Muthén 2012).

The posterior predictive *p* value (PPP) was low for both models: .132 (same-discipline reviews as reference) and .191 (different-discipline as reference). PPPs should be around .50 and not very much smaller for models with an excellent fit (Muthén and Asparouhov 2011). However, the PPP should be regarded as a fit index and not as a statistical test. Moreover, the two models were already liberal and could be changed only with fit-reducing restrictions (e.g., equality constraints). Hence, the results of both models should not be ignored completely, but should be interpreted with great caution.

The estimated latent factor variances for both models are shown in Online Resource 6. It shows for each model that both variances for the rating dimension factors as well as the variance of the method factor were significant. The significant method factor variance provides support for considering the method factor in our model.

Table 8 shows the unstandardized factor loadings for both models (for the corresponding credibility intervals, see Online Resource 7). Interestingly, the indicators with free estimated loadings had significant loadings only on the method factor (if such paths were allowed). But in neither case were loadings on the rating dimension factors significant. Thus, the method factor seemed to be more important for an indicator variable than the corresponding rating dimension factor.

This impression was emphasized by examining the coefficients of the CT-C(M-1) framework (Eid 2000; Eid et al. 2003). The *reliability coefficient* of an indicator variable provides information about the proportion of variance that is not attributable to random measurement error. For indicators which have additional loadings on a method factor, the reliability coefficient is divisible into the *consistency coefficient* and the *method specificity coefficient* which both sum up to the reliability. The consistency coefficient provides information about the measurement-error-free proportion of variance that is determined by the comparison standard component. The method specificity estimates the measurement-error-free proportion of variance that is method-specific and that is therefore not shared with the comparison standard (for detailed formula, see Eid 2000; Eid et al. 2003). Both the

Table 8 Unstandardized factor loadings for a confirmatory factor analysis (CFA) correlated trait–correlated (method minus one) [CT-C(M-1)] model

Rating dimension	Kind of reviewers	F I	F II	F III	F IV	F V	MF
Overall	Same	1.00[a] (0.20)	– (–)	– (–)	– (–)	– (–)	– (1.00[a])
Overall	Different	0.13 (1.00[a])	– (–)	– (–)	– (–)	– (–)	1.00[a] (–)
Relevance	Same	– (–)	1.00[a] (0.13)	– (–)	– (–)	– (–)	– (0.27*)
Relevance	Different	– (–)	0.24 (1.00[a])	– (–)	– (–)	– (–)	0.42* (–)
Novelty	Same	– (–)	– (–)	1.00[a] (0.22)	– (–)	– (–)	– (0.37*)
Novelty	Different	– (–)	– (–)	0.15 (1.00[a])	– (–)	– (–)	0.41* (–)
Significance	Same	– (–)	– (–)	– (–)	1.00[a] (0.13)	– (–)	– (0.41*)
Significance	Different	– (–)	– (–)	– (–)	−0.04 (1.00[a])	– (–)	0.49* (–)
Soundness	Same	– (–)	– (–)	– (–)	– (–)	1.00[a] (0.09)	– (0.46*)
Soundness	Different	– (–)	– (–)	– (–)	– (–)	0.00 (1.00[a])	0.38* (–)

F = factor; MF = method factor (e.g., see Eid 2000). Two models were estimated with Bayes estimator by Mplus (Muthén and Muthén 2012). For identification purposes the loading of each same-discipline indicator on its corresponding factor was fixed to the value of one and the loading of the different-discipline indicator for overall evaluation on the MF was also fixed to the value of one (first model). The values in parentheses result if the loading of each different-discipline indicator on its corresponding factor and also the loading of the same-discipline indicator for overall evaluation on the MF were set to the value of one (second model)

* Bayesian 99% credibility interval does not contain the value of zero (significant)

[a] No credibility interval is estimated because the parameter was fixed to one

consistency coefficient and the method specificity coefficient can be defined (a) with the observed (manifest) variances as divisor or (b) with the measurement-error-free (true-score) variance as divisor. In the latter case, both coefficients add up to one.

All estimated coefficients for the model in which the same-discipline indicators act as a reference method as well as for the model in which the different-discipline indicators act as a reference method are listed in Table 9. As can be seen, particularly in the last two columns, it was nearly exclusively the method specificity which accounted for the measurement-error-free variance. As in the EFA, the relationships among the variables measured with the same method, that is, within each reviewer group, dominated the scenario. The fact that the method-specificity coefficients were much larger than the corresponding consistency coefficients indicates that the convergent validity was very low in all cases (Eid 2000; Eid et al. 2003). In other words, as found in the exploratory factor analysis, same-discipline reviewers and different-discipline reviewers did not agree in their evaluations.

Likewise, the discriminant validity of all constructs is also in doubt. Although the latent (measurement-error-free) correlations (see Table 10) hardly ever reached the critical value of |.80| (or |.85|), which implies poor discriminant validity (e.g., Brown 2015, p. 28), it is obvious that the discriminant validity, in general, should be regarded as rather small. This

Table 9 Coefficients for the confirmatory factor analysis (CFA) correlated trait–correlated method minus one [CT-C(M-1)] model

Rating dimension	Kind of reviewers	Divisor: observed variance			Divisor: true-score variance	
		Reliability	Consistency	Method specificity	Consistency	Method specificity
Overall	Same	.84 (.73)	.84 (.03)	– (.69)	1.00 (.05)	– (.95)
Overall	Different	.81 (.89)	.02 (.89)	.78 (–)	.02 (1.00)	.98 (–)
Relevance	Same	.60 (.30)	.60 (.02)	– (.27)	1.00 (.07)	– (.93)
Relevance	Different	.55 (.79)	.03 (.79)	.50 (–)	.05 (1.00)	.95 (–)
Novelty	Same	.80 (.46)	.80 (.05)	– (.41)	1.00 (.10)	– (.90)
Novelty	Different	.54 (.83)	.02 (.83)	.51 (–)	.04 (1.00)	.96 (–)
Significance	Same	.78 (.48)	.78 (.02)	– (.45)	1.00 (.04)	– (.96)
Significance	Different	.65 (.83)	.01 (.83)	.64 (–)	.01 (1.00)	.99 (–)
Soundness	Same	.80 (.54)	.80 (.01)	– (.52)	1.00 (.02)	– (.98)
Soundness	Different	.45 (.82)	.01 (.82)	.43 (–)	.01 (1.00)	.99 (–)

Two models were estimated with Bayes estimator by Mplus (Muthén and Muthén 2012). For identification purposes the loading of each same-discipline indicator on its corresponding factor was fixed to the value of one and the loading of the different-discipline indicator for overall evaluation on the MF was also fixed to the value of one (first model). Coefficients in parentheses result if the loading of each different-discipline indicator on its corresponding factor and also the loading of the same-discipline indicator for overall evaluation on the MF were set to the value of one (second model). Same-discipline reviewers constitute the reference method in the first model and different-discipline reviewers constitute the reference method in the second model. In each model the method specificity can be estimated only for the non-reference method (see Eid 2000). Consistency and method specificity were estimated in two forms: (a) by dividing the corresponding weighted factor variance components by the observed indicator variance and (b) by dividing the factor variance components by the (combined) true-score (measurement-error-free) variance (for more details, see Eid et al. 2003)

Table 10 Factor intercorrelations from the confirmatory factor analysis (CFA) correlated trait–correlated method minus one [CT-C(M-1)] model

Factor	I	II	III	IV	V	MF
I Overall evaluation	– (–)	.75* (.81*)	.67* (.77*)	.72* (.81*)	.81* (.78*)	–[a] (–[a])
II Relevance	.75* (.81*)	– (–)	.58* (.65*)	.54* (.82*)	.46* (.53*)	–[a] (–[a])
III Novelty	.67* (.77*)	.58* (.65*)	– (–)	.72* (.81*)	.56* (.54*)	–[a] (–[a])
IV Significance	.72* (.81*)	.54* (.82*)	.72* (.81*)	– (–)	.62* (.50*)	–[a] (–[a])
V Soundness	.81* (.78*)	.46* (.53*)	.56* (.54*)	.62* (.50*)	– (–)	–[a] (–[a])
MF method factor	–[a] (–[a])	–[a] (–[a])	–[a] (–[a])	–[a] (–[a])	–[a] (–[a])	– (–)

F = factor; MF = method factor (e.g., see Eid 2000). Two models were estimated with Bayes estimator by Mplus (Muthén and Muthén 2012). For identification purposes the loading of each same-discipline indicator on its corresponding factor was fixed to the value of one and the loading of the different-discipline indicator for overall evaluation on the MF was also fixed to the value of one (first model). Estimates in parentheses result if the loading of each different-discipline indicator on its corresponding factor and also the loading of the same-discipline indicator for overall evaluation on the MF were set to the value of one (second model). For each model the latent correlations resulted from the standardized solution

* Bayesian 99% credibility interval does not contain the value of zero (significant)

[a] MF has to be uncorrelated with the non-method factors (see Eid 2000)

corroborates the conclusion we reached through the inspection of the manifest correlations (see Table 6).

However, even though the latent inter-correlations were relatively high, it must be taken into account that they were based on measurement-error-free variables. Thus, given the fact that latent correlations were mostly below .80, the results indicated that the rating dimensions could be empirically separated. That is, the reviewers' evaluations of the different rating dimensions were quite similar but not identical. In other words, when several reviewers, for example, rated the novelty of a submission to be high (low), these same reviewers were also likely to rate the soundness of the same submission as rather high (low).

The conclusion about a low but serious discriminant validity was also supported by model comparisons between a unidimensional CT-C(M-1) model and a multidimensional CT-C(M-1) model for the dimensions relevance, novelty, significance, and soundness. Hence, a 1-factor model was compared to a 4-factor model. The *deviance information criterion* (DIC; Spiegelhalter et al. 2002) was used as a comparison criterion, whereby the model with a smaller DIC should be preferred (e.g., see Kaplan and Depaoli 2013). Models were estimated without considering overall evaluations. Table 11 shows the DIC values for both (a) models in which the same-discipline scores acted as reference method and (b) models in which the different-discipline scores act as reference method. In both cases, it seems clear that the DIC for the multidimensional model was smaller than the DIC for the unidimensional model. Thus, the supposed distinct dimensions relevance, novelty, significance, and soundness seem to be slightly better represented by a multidimensional model than by a unidimensional model.

Table 11 Relevance, novelty, significance, and soundness: comparison between unidimensional and multidimensional models

Reference method	DIC for the unidimensional model	DIC for the multidimensional model
Same-discipline reviews	1258.97	1224.13
Different-discipline reviews	1234.19	1231.26

DIC = deviance information criterion (Spiegelhalter et al. 2002)

Predictive criterion validity: Citation rate prediction

The distinctiveness of the rating dimensions is also relevant for the final analysis, namely the examination of criterion validity of the rating dimensions. For this purpose, we analyzed whether the accepted papers' citation rate was predicable by the rating dimensions.

From the $n = 145$ submissions which had been rated by at least two reviewers, a total of $n_\text{p} = 82$ accepted submissions were published in the conference proceedings (see footnote 2). Altogether, we analyzed citations in a window of about three years. The mean citation rate was $M = 1.24$ ($SD = 2.02$). The lowest citation rate was 0 ($n_{\text{c}0} = 40$; 48.8%). The highest citation rate was 11 ($n_{\text{c}11} = 1$; 1.2%).

The citation rate, as a genuine count variable, was regressed on the five rating dimensions by applying negative binomial regression models (Hilbe 2011). Parameters were estimated by MLR with Monte Carlo integration (Muthén and Muthén 2012). All analyses were conducted for paper scores (a) based on the ratings from all reviewers, (b) based on the ratings from same-discipline reviewers, and (c) based on the ratings from different-discipline reviewers. Table 12 summarizes the results of these negative binomial regressions.

Based on the adjusted p values, it can be stated that significant partial effects on citation rate appeared only for the two rating dimensions relevance and novelty, and then only if the scores were based on the ratings from same-discipline reviewers. In addition, a significant effect of relevance also appeared when the paper scores based on the ratings of all reviewers were considered. For a more detailed interpretation, the rate ratio coefficients (from a multiplicative model) in Table 12 should be inspected. These coefficients were obtained by exponentiation of the slopes (see Hilbe 2011).

The multiplicative rate ratio coefficient of 4.24 for relevance in Table 12 (same-discipline paper scores) means that, if all other predictors were held constant, one unit increase in relevance increased the citation rate by a factor of 4.24. In other words, one unit increase of relevance ratings made by same-discipline reviewers would result in 324% more citations. On the other hand, the rate ratio coefficient of 0.28 for novelty (same-discipline paper scores) means that, if all other predictors were held constant, one unit increase in novelty decreased the citation rate by a factor of 0.28. That is, a one unit increase of novelty ratings made by the same-discipline reviewers would result in 72% fewer citations. Table 12 also shows that several slopes in the regressions with only one predictor were significant. This different result pattern could be an effect of the high correlations among the predictors.

In sum, our results demonstrated the predictive power of reviewers' relevance and novelty ratings, provided that each reviewer belonged to the same discipline as the paper. These effects emerged even when the citation window was lengthened to a time span of seven years (see Online Resource 8).

Table 12 Negative binomial regressions (NBRs) of citation rate on the five rating dimensions (with a citation window of roughly three years)

Rating dimension	Kind of reviewers	Multiple NBRs		Simple NBRs	
		Slope b_i^{a}	$\exp(b_i)$	Slope b_1^{b}	$\exp(b_1)$
Overall	All reviewers	0.12	1.12	1.01*	2.74
	Same-discipline	0.19	1.21	0.44	1.55
	Different-discipline	0.15	1.16	0.68*	1.97
Relevance	All reviewers	1.64*	5.16	2.18*	8.84
	Same-discipline	1.44*	4.24	1.51*	4.51
	Different-discipline	0.87	2.38	1.57*	4.80
Novelty	All reviewers	−0.12	0.89	0.66	1.93
	Same-discipline	−1.26*	0.28	−0.59	0.55
	Different-discipline	0.06	1.06	0.72	2.06
Significance	All reviewers	0.39	1.47	1.08*	2.95
	Same-discipline	0.47	1.60	0.58	1.78
	Different-discipline	0.25	1.28	0.88*	2.41
Soundness	All reviewers	0.50	1.66	0.95*	2.57
	Same-discipline	−0.16	0.85	0.18	1.19
	Different-discipline	0.34	1.40	0.76	2.14

Only papers with at least one non-missing predictor value were considered for multiple NBRs (via Monte-Carlo numerical integration with 5000 integration points; see Muthén and Muthén 2012). The multiple NBRs were conducted with $n_{\text{both}} = 82$ papers (all reviewers), $n_{\text{same}} = 67$ papers (same-discipline), and $n_{\text{diff}} = 70$ papers (different-discipline). All predictors were mean-centered by using the means based on all reviewers. Estimated intercepts were −0.59 (all reviewers, $p = .017$), −0.31 (same-discipline, $p = .218$), and −0.35 (different-discipline, $p = .193$). Estimated negative binomial dispersion parameters were 0.72 ($p < .05$), 0.73 ($p = .012$), and 0.88 ($p = .010$) respectively. The multiple NBRs yielded smaller values for the Bayesian information criterion (BIC; Schwarz 1978) than Poisson regression models: 636.35, 692.20, and 726.38 versus 651.60, 704.31, and 746.09. Similar conclusions were reached by boundary likelihood ratio tests (Hilbe 2011). Sample sizes for simple NBRs ranged from 81 to 82 (all reviewers), from 61 to 67 (same-discipline), and from 66 to 70 (different-discipline)

* $p_{\text{adjust}} < .05$, two-tailed, with a multiple (familywise) significance level of .05 (rate ratios were not tested)

[a,b]For each class of reviewers a p value adjustment was conducted for $m = 5$ tests (see Holm 1979)

Discussion

The study set out to investigate two facets of the quality of reviews in an interdisciplinary context: inter-rater reliability and validity. Taken together, our findings draw a somewhat pessimistic and, to some extent, mixed picture. Not only did we find little agreement among reviewers, our findings also argue for little convergent as well as discriminant construct validity. However, with regard to predictive criterion validity, we found that the ratings for relevance and novelty were capable of predicting the citation rate of the accepted papers which were published in the conference proceedings. These effects were restricted to ratings made by reviewers from the same discipline as the papers. Let us address each aspect in detail.

Poor agreement among reviewers

Across all reviewers (same-discipline and different-discipline) the agreement among reviewers was above chance, yet rather poor (Cicchetti 1994). This is a common finding, as indicated by the most recent meta-analysis (Bornmann et al. 2010; see also Cicchetti 1991). It seems noteworthy, however, that the average agreement based on all reviewers and across all five rating dimensions (.22) was even below the mean of ICCs and r^2 coefficients obtained in the meta-analysis (.34).

It would seem plausible to suspect that the interdisciplinary context accounts for these findings, since most prior studies on inter-rater reliability were conducted within single scientific disciplines. So, one could expect that shared standards would enhance agreement. Our results do not confirm this argument, however. A differentiation between intra- and interdisciplinary reviews even indicates the opposite. Whereas agreement was poor and not even distinguishable from chance for intra-disciplinary reviews (average single-rater reliability: .16), it was higher and well above chance for interdisciplinary reviews (average single-rater reliability: .35). This is interesting in consideration of the fact that the discipline of an "outsider" may not only differ from the paper's discipline but also from other reviewers' disciplines (e.g., when a paper from the psychological-experimental category is rated by one reviewer with informational-technological background and by another reviewer with social-educational background). To be sure, "higher" agreement in this case meant "fair" instead of "poor" agreement, not "good" or even "excellent" agreement (Cicchetti 1994).

It must also be acknowledged that the differences between interdisciplinary and intra-disciplinary reviews were not statistically significant. Therefore, the descriptively higher agreement in interdisciplinary reviews should not be overemphasized. It does make clear, however, that the overall poor agreement obtained in our study cannot be attributed to the interdisciplinary context. At the same time, those results throw into question the implicit assumption that intra-disciplinary inter-rater agreement is, a priori, higher than interdisciplinary inter-rater agreement. It is still far too early to draw any conclusions, since more research is needed in this area. After all, our study was the first to examine agreement among reviewers in an interdisciplinary context, so now further studies are needed. And we strongly recommend appropriate statistics for future studies, as ill-structured designs seem to be the norm rather than the exception in peer-review data sets (e.g., when analyzing submissions to scientific journals or meetings).

Even more importantly, however, more agreement is needed. By now, several researchers have criticized peer ratings as they "fall short of acceptable standards of reliability" (e.g., Marsh et al. 2007, p. 37). Of course, agreement is relevant and desirable only if one assumes that manuscripts possess an inherent objective quality (Kirk and Franke 1997). From the standpoint of rejecting this idea (e.g., Luce 1993) the very notion of quality control becomes irrelevant, and consequently peer review would not be needed to serve a gatekeeping function. As long as decisions for or against the acceptance of a submitted manuscript or grant proposal are based on peer reviews, however, "appreciable levels of agreement and a principled, valid basis for agreement" are necessary (Whitehurst 1983, p. 78; Burdock et al. 1963). After all, these decisions have an impact on the career of researchers (e.g., van Dalen and Henkens 2012). Therefore, research should not only tackle the status quo of interrater agreement but also how to improve it.

Poor construct validity of the reviewers' evaluations

The findings of our various analyses regarding construct validity of the ratings draw a very coherent picture. On the one hand, they show evidence for poor convergent validity. First, manifest monotrait-heteromethod correlations of the same dimensions between different groups of raters were low. Second, the exploratory factor analysis yielded two factors instead of five. The pattern of the factor loadings clearly revealed that these two factors were not based on content but rather represented two method factors (one for the same-discipline reviews and one for the different-discipline reviews). Third, the confirmatory factor analysis also pointed to the differentiation between same-discipline and different-discipline reviews and revealed that the corresponding indicators did not significantly load on common rating dimension factors. Together with a large method factor variance and a high method-specificity coefficient, this indicates a poor convergent validity.

At the same time, our findings also suggest a rather poor discriminant validity. First, manifest correlations between different rating dimensions within each reviewer group (heterotrait-monomethod correlations) and across all reviewers were rather high (even if below .80). Second, the exploratory factor analysis yielded high loadings within the factors of each reviewer group. Third, the confirmatory factor analysis yielded high latent correlations between factors. However, most latent correlations were not as high as would have been expected if reviewers' evaluations had been unidimensional (e.g., above .80). Similarly, our comparison between a unidimensional and a multidimensional model suggested that the different rating dimensions were closely related high but not close enough to suggest a unidimensional model. Rather, the multidimensional model yielded a better fit. In other words, despite the fact that we found only modest empirical support for the distinctiveness of the rating dimensions, it might still make sense to ask reviewers to take (those) different dimensions into account when evaluating a submission.

Obviously, the rating dimensions were similar, but not redundant. Consequently, they still provided more information than a single global rating would have, but they hardly reflected independently assessed dimensions. Here, our results are therefore similar to Marsh and Ball (1989), who concluded that "although there was support for the conceptual and empirical distinctiveness of four components, there was little support for their practical utility" (pp. 165–166).

Relevance and novelty as predictors for the citation rate

On the other hand, with regard to the prediction of citation rate, our findings point to a practical utility of distinct rating dimensions. That is, our findings demonstrate the predictive criterion validity of relevance and novelty ratings made by reviewers from the same discipline as the paper.

With regard to highly cited papers, it seems clear that "in order to get highly cited the content of the highly cited paper must be useful or of relevance for the research activity" (Aksnes 2003, p. 167). Accordingly, highly relevant papers, at least in the eyes of same-discipline reviewers, were cited more frequently than papers which received low relevance ratings. This finding is not trivial, as it indicates that same-discipline reviewers are in fact capable of evaluating the relevance of submissions for their scientific community. This does not mean, however, that attributed relevance automatically indicates objective quality which would later be reflected in the high resonance within the reviewers' scientific community. Rather, it means that reviewers seem to have a good sense for which submissions would generate that kind of resonance. And this resonance takes, among others, the form of countable citations, which is a kind of scientific currency. However, the implications, shortcomings, and benefits of the resonance metaphor cannot be discussed and deepened here (for a positively connoted socio-psychological concept of resonance, see Rosa 2016; for a comparison of theoretical approaches to citation behavior, see Bornmann and Daniel 2008c).

The observable success or impact of highly relevant papers in terms of citations does not necessarily imply that especially highly innovative papers generate a lot of resonance. Quite the contrary, because another important finding was that papers with high (low) novelty ratings from same-discipline reviewers were cited to a lesser (higher) degree. This is in line with the findings of Stephan et al. (2017) who concluded that "more-novel papers were more likely to be either a big hit or ignored compared with non-novel papers in the same field" and that "novelty needs time" (p. 412; for some remarks on the conservative bias, see, for example, Benda and Engels 2011; Lee et al. 2013). On the other side, it has also to be considered that novelty by itself does not guarantee scientific quality.

Limitations

One criticism might be that the ratings per paper should not be aggregated at all due to the rather low *k*-rater reliabilities. And the poor reliability for single reviewers and for both reviewer groups might explain the appearance of the method factors that we observed in the exploratory factor analyses. Thus, the agreement was simply too low for distinct content factors to emerge. In this case, however, the conclusion would be even more negative, as we could not speak of low construct validity but would have to say that the results of an analysis of validity were difficult to interpret. The implication would be even more obvious: we urgently need a better agreement in peer reviews. This resembles the insight from classical test theory that reliability (more precisely, the reliability index) restricts the possible upper limit for validity (e.g., Raykov and Marcoulides 2011, pp. 193–194). Here again, it is even more important to not only investigate the peer-review system, but to try to improve it.

Another limitation which could be discussed regards the nature of our data as well as the nature of data from review processes in general. Except for the estimations of the inter-rater reliability, almost all analyses were based on the paper scores, just as analyses in psychological research have usually been based on participant scores (reviewers and items become the measurement instruments in both scenarios). These paper scores were constructed by averaging the ratings (a) of all raters, (b) of same-discipline raters, and (c) of different-discipline raters. Thus, these variables could be regarded as quasi-continuous. With respect to the non-aggregated ratings, the question whether such single-item rating scales can be treated as continuous variables in the analysis process is still a heavily debated issue between "purists" and "pragmatics" (Bortz and Döring 2006, p. 181). The pragmatic strategy seems to be justified for new research questions and for cases in which important and consistent result patterns are obtained, which are later replicated with more sophisticated methods (Bortz and Döring 2006, p. 182; see, for example, also Hassebrauck 1993; Rhemtulla et al. 2012). For example, single-item rating scales have been successfully applied to collect ratings of physical attractiveness of target persons (e.g., Hassebrauck 1983; Hönekopp 2006), to measure self-rated political ideology (e.g., Cohrs et al. 2005), or to collect impressions about the persuasiveness of presented material (e.g., Lord et al. 1979).

A further limitation could be that we used only Web-of-Science databases for counting the citations rates. However, for the time span of our investigations (see footnote 2), Web-of-Science could be seen as an appropriate interdisciplinary citation index that ensured a comparatively high accuracy standard when searching for citations in scientific writings (e.g., de Winter et al. 2014; see also Baethge et al. 2013). A methodical advantage was the fact that all of the papers we analyzed were published at the same time. An offset correction for different time spans was therefore not necessary.

Implications and outlook

Despite the limitations of our study, we regard the results to be worrisome. While consequences of low inter-rater reliabilities may be limited with regard to conference submissions, they are more serious when it comes to journal submissions and even more severe in the case of grants, which ultimately shape careers. Thus, it seems to be time to not only examine the reliability and validity of the peer-review system, but rather to think of better ways to assess submissions. Several suggestions have been put forward in the recent past.

It is possible that critical appraisal tools (CAT; for an overview see Crowe and Sheppard 2011a) could be important in this regard. CATs are standardized instruments that can be used for the evaluation of scientific documents, and they were designed mainly as a tool for systematic reviews. They allow for a thorough evaluation of research articles and enable identification of the best articles on a given topic (e.g., Crowe and Sheppard 2011a). However, evidence is sparse for the reliability and validity of such tools as these (Crowe and Sheppard 2011a, b).

Another obvious starting point might be to improve inter-rater reliability by training reviewers (Oxman et al. 1991). Such training, however, requires a shared understanding of both the criteria for paper quality and of when and to what extent the criteria are met. Hence, journal editors, conference organizers, and grant suppliers would need to agree on these issues in order to be able to provide detailed instructions for reviewers. It might be, however, that formal training would be less efficient than one would hope (Callaham and Tercier 2007; see also Houry et al. 2012). Moreover, some reviewers might feel their academic freedom is threatened by training (e.g., Adams 1991).

A further possibility for enhancing agreement is to avoid reviewers who have been nominated by the author (Marsh et al. 2007). The numbers of reviewers could also be increased in order to obtain more reliable results (Wood et al. 2004). Since an increase in reviewers would increase substantially the effort, this particular solution might be appropriate mostly for cases with more serious consequences (e.g., grants) in order to decrease the impact of chance (Cole et al. 1981; but see also List 2017).

Another approach to increasing reliability might be the "reader system", which has been suggested by Jayasinghe et al. (2006). The most important aspects of this system are that the same three to four experts of a sub-discipline review all proposals and are asked to rank them. This proceeding is characterized by a shared frame of reference and eliminates rater-effects (leniency/harshness). The reader system has been developed with grant proposals in mind, all of which are supposed to be submitted at the same time. To apply this system to journals, which receive submissions on a continuous basis, however, would require making some adaptations. Applicability might not be realistic at all in cases where many submissions have to be evaluated.

Finally, open peer review has been suggested (Groves 2010). In this system, reviewers would not remain anonymous but would sign their reviews. Initial evidence suggests that open peer reviews are of higher quality (Walsh et al. 2000). We are not aware of any study which examines reliability or validity issues in open peer review, let alone one which provides a comparison to traditional (closed) peer review (but see DeCoursey 2006 and Khan 2010 for a discussion of advantages and disadvantages of open peer review).

Apart from thinking of variations of the traditional peer-review system, however, one might likewise think of alternatives to it. Low agreement is less of a problem, for instance, if peer review does not fulfill a gate-keeping function (for publication). Imagine a scenario where everything that was published had passed an initial screening (Wood et al. 2004). In this scenario, it would be essentially up to the entire scientific community to deal with the publication. How acceptable the publication proved to be would still be measurable by citations. Based on prior evidence that the impact in the scientific community is only loosely linked to reviewers' evaluations (Akerlof 2003; Gottfredson 1978; Harrison 2004) such a scenario might be particularly interesting. Moreover, it could be combined with novel elements that have become possible with Web 2.0 and social media, such as reader evaluations or post publication peer review (but see Anderson 2012). Of course, we do not know yet whether such a system would be superior to the traditional peer-review system (Smith 2003). As long as there are no empirical comparisons, however, traditional peer review may simply survive because of the lack of good alternatives. This is, however, weak justification for its implementation.

With regard to validity, our results indicated a rather poor convergent validity, although the model comparison did favor a multidimensional model with distinct rating dimensions. Therefore, it seems too premature to conclude that distinct rating dimensions are unnecessary. This is especially true with regard to the dimensions relevance and novelty, which were predictive of the citation rate provided the ratings were made by reviewers from the same discipline as the papers. Future studies should, therefore, investigate the predictive criterion validity of different rating dimensions with regard to a broad range of criterion variables. Criterion variables could be operationalized, for example, by social network analysis (SNS) methods. Such variables could indicate how central an article is in a citation network and to what degree an article has linked different disciplines (e.g., Halatchliyski and Cress 2014). In addition, a simple but perhaps effective method to predict citation and download rates could be to ask reviewers directly to assess the papers' potential for generating many citations and clicks. However, a legitimate question is whether the citation rate is actually an adequate proxy for scientific quality (e.g., Bornmann and Daniel 2008c; Lindsey 1989; Stephan et al. 2017; Tahamtan et al. 2016). But that is another story which is also worthy of more attention by future research.

References

Adams, K. M. (1991). Peer review: An unflattering picture. *Behavioral and Brain Sciences, 14*(1), 135–136.

Akerlof, G. A. (2003). *Writing the "The Market for 'Lemons'": A personal and interpretive essay*. https://www.nobelprize.org/nobel_prizes/economic-sciences/laureates/2001/akerlof-article.html. Accessed 4 Sept 2017.

Aksnes, D. W. (2003). Characteristics of highly cited papers. *Research Evaluation, 12*(3), 159–170.

Altman, D. G., & Bland, J. M. (2011). How to obtain the P value from a confidence interval. *BMJ, 343,* d2304.

Anderson, K. (2012). *The problems with calling comments "Post-Publication Peer-Review" [Web log message]*. Retrieved from http://scholarlykitchen.sspnet.org/2012/03/26/the-problems-with-calling-comments-post-publication-peer-review.

Asparouhov, T., & Muthén, B. (2010). *Bayesian analysis of latent variable models using Mplus*. http://www.statmodel.com/download/BayesAdvantages18.pdf. Accessed 30 Mar 2017.

Baethge, C., Franklin, J., & Mertens, S. (2013). Substantial agreement of referee recommendations at a general medical journal—A peer review evaluation at Deutsches Ärzteblatt International. *PLoS ONE, 8*(5), e61401.

Bailar, J. C., & Patterson, K. (1985). Journal peer review—The need for a research agenda. *The New England Journal of Medicine, 312*(10), 654–657.

Benda, W. G. G., & Engels, T. C. E. (2011). The predictive validity of peer review: A selective review of the judgmental forecasting qualities of peers, and implications for innovation in science. *International Journal of Forecasting, 27*(1), 166–182.

Beyer, J. M., Chanove, R. G., & Fox, W. B. (1995). Review process and the fates of manuscripts submitted to AMJ. *Academy of Management Journal, 38*(5), 1219–1260.

Blackburn, J. L., & Hakel, M. D. (2006). An examination of sources of peer-review bias. *Psychological Science, 17*(5), 378–382.

Bornmann, L., & Daniel, H.-D. (2005). Selection of research fellowship recipients by committee peer review. Reliability, fairness and predictive validity of Board of Trustees' decisions. *Scientomentrics, 63*(2), 297–320.

Bornmann, L., & Daniel, H.-D. (2008a). Selecting manuscripts for a high-impact journal through peer review: A citation analysis of communications that were accepted by Angewandte Chemie International Edition, or rejected but published elsewhere. *Journal of the American Society for Information Science and Technology, 59*(11), 1841–1852.

Bornmann, L., & Daniel, H.-D. (2008b). 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.

Bornmann, L., & Daniel, H.-D. (2008c). What do citations counts measure? A review of studies on citing behavior. *Journal of Documentation, 64*(1), 45–80.

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.

Bortz, J., & Döring, N. (2006). *Forschungsmethoden und evaluation für Human- und Sozialwissenschaftler [Research methods and evaluation for human and social scientists]* (4th ed.). Heidelberg, DE: Springer.

Brennan, R. L. (2001). *Generalizability theory*. New York, NY: Springer.

Brown, T. A. (2015). *Confirmatory factor analysis for applied research* (2nd ed.). New York, NY: Guilford Press.

Burdock, E. I., Fleiss, J. L., & Hardesty, A. S. (1963). A new view of inter-observer agreement. *Personnel Psychology, 16*(4), 373–384.

Callaham, M. L., & Tercier, J. (2007). The relationship of previous training and experience of journal peer reviewers to subsequent review quality. *PLoS Medicine, 4*(1), e40.

Campanario, J. M. (1998). Peer review for journals as it stands today—Part 1. *Science Communication, 19*(3), 181–211.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*(2), 81–105.

Campion, M. A. (1993). Article review checklist: A criterion checklist for reviewing research articles in applied psychology. *Personnel Psychology, 46*(3), 705–718.

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research, 1*(2), 245–276.

Cattell, R. B., & Jaspers, J. (1967). A general plasmode (No. 30-10-5-2) for factor analytic exercises and research. *Multivariate Behavioral Research Monographs, 67,* 1–212.

Chase, J. M. (1970). Normative criteria for scientific publication. *American Sociologist, 5*(3), 262–265.

Church, R. M., Crystal, J. D., & Collyer, C. E. (1996). Correction of errors in scientific research. *Behavior Research Methods, Instruments, & Computers, 28*(2), 305–310.

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.

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment, 6*(4), 284–290.

Cicchetti, D. V., & Conn, H. O. (1976). A statistical analysis of reviewer agreement and bias in evaluating medical abstracts. *Yale Journal of Biology and Medicine, 49*(4), 373–383.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37–46.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cohrs, J. C., Moschner, B., Maes, J., & Kielmann, S. (2005). The motivational bases of right-wing authoritarianism and social dominance orientation: Relations to values and attitudes in the aftermath of September 11, 2001. *Personality and Social Psychology Bulletin, 31*(10), 1425–1434.

Cole, S., Cole, J. R., & Simon, G. A. (1981). Chance and consensus in peer review. *Science, 214*(4523), 881–886.

Cornforth, J. W. (1974). Referees. *New Scientist, 62*(892), 39.

Crowe, M., & Sheppard, L. (2011a). A general critical appraisal tool: An evaluation of construct validity. *International Journal of Nursing Studies, 48*(12), 1505–1516.

Crowe, M., & Sheppard, L. (2011b). A review of critical appraisal tools show they lack rigor: Alternative tool structure is proposed. *Journal of Clinical Epidemiology, 64*(1), 79–89.

de Winter, J. C. F., Zadpoor, A. A., & Dodou, D. (2014). The expansion of Google Scholar versus Web of science: A longitudinal study. *Scientometrics, 98*(2), 1547–1565.

DeCoursey, T. (2006). The pros and cons of open peer review. *Nature*. Retrieved from http://www.nature.com/nature/peerreview/debate/nature04991.html.

Donner, A. (1986). A review of inference procedures for the intraclass correlation coefficient in the one-way random effects model. *International Statistical Review, 54*(1), 67–82.

Dziuban, C. D., & Shirkey, E. C. (1974). When is a correlation matrix appropriate for factor analysis? Some decision rules. *Psychological Bulletin, 81*(6), 358–361.

Eid, M. (2000). A multitrait-multimethod model with minimal assumptions. *Psychometrika, 65*(2), 241–261.

Eid, M., Lischetzke, T., Nussbeck, F. W., & Trierweiler, L. I. (2003). Separating trait effects from trait-specific method effects in multitrait-multimethod models: A multiple-indicator CT-C(M-1) model. *Psychological Methods, 8*(1), 38–60.

Enders, C. K. (2001). The performance of the full information maximum likelihood estimator in multiple regression models with missing data. *Educational and Psychological Measurement, 61*(5), 713–740.

Enders, C. K. (2010). *Applied missing data analysis*. New York, NY: Guilford Press.

Feinstein, A. R., & Cicchetti, D. V. (1990). High agreement but low kappa: I. The problems of two paradoxes. *Journal of Clinical Epidemiology, 43*(6), 543–549.

Field, A. (2009). *Discovering statistics using SPSS* (3rd ed.). Thousand Oaks, CA: Sage.

Fisher, R. A. (1934). *Statistical methods for research workers* (5th ed.). Edinburgh: Oliver and Boyd.

Fiske, D. W., & Fogg, L. (1990). But the reviewers are making different criticisms of my paper! Diversity and uniqueness in reviewer comments. *American Psychologist, 45*(5), 591–598.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin, 76*(5), 378–382.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis* (3rd ed.). Boca Raton, FL: Chapman and Hall/CRC.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science, 7*(4), 457–472.

Gilliland, S. W., & Cortina, J. M. (1997). Reviewer and editor decision making in the journal review process. *Personnel Psychology, 50*(2), 427–452.

Gottfredson, S. D. (1978). Evaluating psychological research reports: Dimensions, reliability, and correlates of quality judgments. *American Psychologist, 33*(10), 920–934.

Groves, T. (2010). Is open peer the fairest system? Yes. *BMJ, 341,* c6424.

Gwet, K. L. (2008). Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology, 61,* 29–48.

Gwet, K. L. (2014). *The definitive guide to measuring the extent of agreement among raters* (4th ed.). Gaithersburg, MD: Advanced Analytics.

Halatchliyski, I., & Cress, U. (2014). How structure shapes dynamics: Knowledge development in Wikipedia—A network multilevel modeling approach. *PLoS ONE, 9*(11), e111958.

Hardwig, J. (1985). Epistemic dependence. *The Journal of Philosophy, 82*(7), 335–349.

Harrison, C. (2004). Peer review, politics and pluralism. *Environmental Science & Policy, 7*(5), 357–368.

Hassebrauck, M. (1983). Die Beurteilung der physischen Attraktivität: Konsens unter Urteilern? [Judging physical attractiveness: Consensus among judges?]. *Zeitschrift für Sozialpsychologie, 14*(2), 152–161.

Hassebrauck, M. (1993). Die Beurteilung der physischen Attraktivität [The assessment of physical attractiveness]. In M. Hassebrauck & R. Niketta (Eds.), *Physische Attraktivität [Physical attractiveness]* (1st ed., pp. 29–59). Göttingen, DE: Hogrefe.

Hayton, J. C., Allen, D. G., & Scarpello, V. (2004). Factor retention decisions in exploratory factor analysis: A tutorial on parallel analysis. *Organizational Research Methods, 7*(2), 191–205.

Hemlin, S., & Montgomery, H. (1990). Scientists' conceptions of scientific quality: An interview study. *Science Studies, 3*(1), 73–81.

Hemlin, S., & Rasmussen, S. B. (2006). The shift in academic quality control. *Science, Technology and Human Values, 31*(2), 173–198.

Henss, R. (1992). *"Spieglein, Spieglein an der Wand …": Geschlecht, Alter und physische Attraktiviät ["Mirror, mirror on the wall…": Sex, age, and physical attractiveness]*. Weinheim, DE: PVU.

Herzog, H. A., Podberscek, A. L., & Docherty, A. (2005). The reliability of peer review in anthrozoology. *Anthrozoos, 18*(2), 175–182.

Hilbe, J. M. (2011). *Negative binomial regression* (2nd ed.). Cambridge: Cambridge University Press.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics, 6*(2), 65–70.

Hönekopp, J. (2006). Once more: Is beauty in the eye of the beholder? Relative contributions of private and shared taste to judgments of facial attractiveness. *Journal of Experimental Psychology, 32*(2), 199–209.

Hönekopp, J., Becker, B. J., & Oswald, F. L. (2006). The meaning and suitability of various effect sizes for structured Rater x Ratee designs. *Psychological Methods, 11*(1), 72–86.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika, 30*(2), 179–185.

Houry, D., Green, S., & Callaham, M. (2012). Does mentoring new peer reviewers improve review quality? A randomized trial. *BMC Medical Education, 12*.

Howard, L., & Wilkinson, G. (1998). Peer review and editorial decision-making. *British Journal of Psychiatry, 173,* 110–113.

Hutcheson, G. D., & Sofroniou, N. (1999). *The multivariate social scientist*. Thousand Oaks, CA: Sage.

IBM Corp. (2011). *IBM SPSS Statistics for windows (version 20.0) [computer software]*. Armonk, NY: IBM Corp.

Jayasinghe, U. W., Marsh, H. W., & 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 A, 166*(3), 279–300.

Jayasinghe, U. W., Marsh, H. W., & Bond, N. (2006). A new reader trial approach to peer review in funding research grants: An Australian experiment. *Scientometrics, 69*(3), 591–606.

Kaiser, H. F. (1970). A second generation Little Jiffy. *Psychometrika, 35*(4), 401–415.

Kaiser, H. F., & Rice, J. (1974). Little Jiffy, Mark IV. *Educational and Psychological Measurement, 34*(1), 111–117.

Kaplan, D., & Depaoli, S. (2013). Bayesian statistical methods. In T. D. Little (Ed.), *The Oxford handbook of quantitative methods* (Vol. 1, pp. 407–437). New York, NY: Oxford University Press.

Kemper, K. J., McCarthy, P. L., & Cicchetti, D. V. (1996). Improving participation and interrater agreement in scoring ambulatory pediatric association abstracts: How well have we succeeded? *Archives of Pediatrics and Adolescent Medicine, 150*(4), 380–383.

Khan, K. (2010). Is open peer review the fairest system? No. *BMJ, 341,* c6425.

Kirk, S. A., & Franke, T. M. (1997). Agreeing to disagree: A study of the reliability of manuscript reviews. *Social Work Research, 21*(2), 121–126.

Kitcher, P. (1990). The division of cognitive labor. *The Journal of Philosophy, 87*(1), 5–22.

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.

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.

Li, D., & Agha, L. (2015). Big names or big ideas: Do peer-review panels select the best science proposals? *Science, 348,* 434–438.

Lindsey, D. (1988). Assessing precision in the manuscript review process: A little better than a dice roll. *Scientometrics, 14*(1–2), 75–82.

Lindsey, D. (1989). Using citation counts as a measure of quality in science measuring what's measurable rather than what's valid. *Scientometrics, 15*(3–4), 189–203.

List, B. (2017). Crowd-based peer review can be good and fast. *Nature, 546*(7656), 9.

Lord, C. G., Ross, L., & Lepper, M. R. (1979). Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of Personality and Social Psychology, 37*(11), 2098–2109.

Luce, R. D. (1993). Reliability is neither to be expected nor desired in peer review. *Behavioral and Brain Sciences, 16*(2), 399–400.

Marsh, H. W., & Ball, S. (1981). Interjudgmental reliability of reviews for the Journal of Educational Psychology. *Journal of Educational Psychology, 73*(6), 872–880.

Marsh, H. W., & Ball, S. (1989). The peer review process used to evaluate manuscripts submitted to academic journals: Interjudgmental reliability. *The Journal of Experimental Education, 57*(2), 151–169.

Marsh, H. W., Bond, N. W., & Jayasinghe, U. W. (2007). Peer review process: Assessments by applicant-nominated referees are biased, inflated, unreliable and invalid. *Australian Psychologist, 42*(1), 33–38.

Marsh, H. W., Jayasinghe, U. W., & Bond, N. W. (2008). Improving the peer-review process for grant applications: Reliability, validity, bias, and generalizability. *American Psychologist, 63*(3), 160–168.

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods, 1*(1), 30–46.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741–749.

Montgomery, A. A., Graham, A., Evans, P. H., & Fahey, T. (2002). Inter-rater agreement in the scoring of abstracts submitted to a primary care research conference. *BMC Health Services Research, 2*.

Muthén, B. (2010). *Bayesian analysis in Mplus: A brief introduction [manuscript]*. http://www.statmodel.com/download/IntroBayesVersion%203.pdf. Accessed March 30 2017.

Muthén, B., & Asparouhov, T. (2011). *Bayesian SEM: A more flexible representation of substantive theory [manuscript]*. http://www.statmodel.com/download/BSEMv4REVISED. Accessed March 30 2017.

Muthén, L. K., & Muthén, B. O. (2012). *Mplus user's guide* (7th ed.). Los Angeles, CA: Muthén & Muthén.

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), e48509.

O'Brien, R. M. (1991). The reliability of composites of referee assessments of manuscripts. *Social Science Research, 20*(3), 319–328.

O'Neill, T. A., Goffin, R. D., & Gellatly, I. R. (2012). The use of random coefficient modeling for understanding and predicting job performance ratings: An application with field data. *Organizational Research Methods, 15*(3), 436–462.

Opthof, T., Coronel, R., & Janse, M. J. (2002). The significance of the peer review process against the background of bias: Priority ratings of reviewers and editors and the prediction of citation, the role of geographical bias. *Cardiovascular Research, 56*(3), 339–346.

Oxman, A. D., Guyatt, G. H., Singer, J., Goldsmith, C. H., Hutchison, B. G., et al. (1991). Agreement among reviewers of review articles. *Journal of Clinical Epidemiology, 44*(1), 91–98.

Petty, R. E., Fleming, M. A., & Fabrigar, L. R. (1999). The review process at PSPB: Correlates of inter-reviewer agreement and manuscript acceptance. *Personality and Social Psychology Bulletin, 25*(2), 188–203.

Platt, J. R. (1964). Strong inference: Certain systematic methods of scientific thinking may produce much more rapid progress than others. *Science, New Series, 146*(3642), 347–353.

Popper, K. R. (1968). Epistemology without a knowing subject. *Studies in Logic and the Foundations of Mathematics, 52,* 333–373.

Pulakos, E. D., Schmitt, N., & Ostroff, C. (1986). A warning about the use of a standard deviation across dimensions within ratees to measure halo. *Journal of Applied Psychology, 71*(1), 29–32.

Putka, D. J. (2002). *The variance architecture approach to the study of constructs in organizational contexts* (Doctoral dissertation, Ohio University). http://etd.ohiolink.edu/. Accessed March 30 2017.

Putka, D. J., Lance, C. E., Le, H., & McCloy, R. A. (2011). A cautionary note on modeling multitrait–multirater data arising from ill-structured measurement designs. *Organizational Research Methods, 14*(3), 503–529.

Putka, D. J., Le, H., McCloy, R. A., & Diaz, T. (2008). Ill-structured measurement designs in organizational research: Implications for estimating interrater reliability. *Journal of Applied Psychology, 93*(5), 959–981.

Qiu, L. (1992). A study of interdisciplinary research collaboration. *Research Evaluation, 2*(3), 169–175.

R Core Team. (2016). *R: A language and environment for statistical computing (Version 3.3.1) [computer software]*. Vienna, AT: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org.

Ramasundarahettige, C. F., Donner, A., & Zou, G. Y. (2009). Confidence interval construction for a difference between two dependent intraclass correlation coefficients. *Statistics in Medicine, 28*(7), 1041–1053.

Raykov, T., & Marcoulides, G. A. (2011). *Introduction to psychometric theory*. New York, NY: Routledge.

Revelle, W. (2016). *Psych: Procedures for personality and psychological research (Version 1.6.9) [computer software]*. Evanston, IL: Northwestern University. http://cran.r-project.org/web/packages/psych/. Accessed March 30 2017.

Rhemtulla, M., Brosseau-Liard, P. É., & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological Methods, 17*(3), 354–373.

Rosa, H. (2016). *Resonanz - Eine Soziologie der Weltbeziehung [Resonance—A sociology of the relationship to the world]*. Berlin, DE: Suhrkamp.

Rubin, D. B. (1976). Inference and missing data. *Biometrika, 63*(3), 581–592.

Rubin, H. R., Redelmeier, D. A., Wu, A. W., & Steinberg, E. P. (1993). How reliable is peer review of scientific abstracts? Looking back at the 1991 annual meeting of the Society of General Internal Medicine. *Journal of General Internal Medicine, 8*(5), 255–258.

Satorra, A., & Bentler, P. M. (2010). Ensuring positiveness of the scaled Chi square test statistic. *Psychometrika, 75*(2), 243–248.

Scarr, S., & Weber, B. L. R. (1978). The reliability of reviews for the American Psychologist. *American Psychologist, 33*(10), 935.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6*(2), 461–464.

Scott, W. A. (1974). Interreferee agreement on some characteristics of manuscripts submitted to Journal of Personality and Social Psychology. *American Psychologist, 29*(9), 698–702.

Searle, S. R., Casella, G., & McCulloch, C. E. (1992). *Variance components*. New York, NY: Wiley.

Serlin, R. C. (1993). Confidence intervals and the scientific method: A case for Holm on the range. *Journal of Experimental Education, 61*(4), 350–360.

Shaffer, J. P. (1995). Multiple hypothesis testing. *Annual Review of Psychology, 46*(1), 561–584.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*(2), 420–428.

Smith, R. (2003). *The future of peer review*. http://pdfs.semanticscholar.org/7c06/8fcda6956132db6732e6c353ffe5fe6b6f62.pdf?_ga=1.116839174.1674370711.1490806067. Accessed March 29 2017.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B, 64*(4), 583–639.

Stephan, P., Veugelers, R., & Wang, J. (2017). Reviewers are blinkered by bibliometrics. *Nature, 544*(7651), 411–412.

Strauss, M. E., & Smith, G. T. (2009). Construct validity: Advances in theory and methodology. *Annual Review of Clinical Psychology, 5,* 1–25.

Tahamtan, I., Afshar, A. S., & Ahamdzadeh, K. (2016). Factors affecting number of citations: A comprehensive review of the literature. *Scientometrics, 107*(3), 1195–1225.

Thorndike, E. L. (1920). A constant error in psychological ratings. *Journal of Applied Psychology, 4*(1), 25–29.

Uebersax, J. S. (1982–1983). A design-independent method for measuring the reliability of psychiatric diagnosis. *Journal of Psychiatric Research, 17*(4), 335–342.

van Dalen, H. P., & Henkens, K. (2012). Intended and unintended consequences of a publish-or-perish culture: A worldwide survey. *Journal of the American Society for Information Science and Technology, 63*(7), 1282–1293.

Van de Schoot, R., Kaplan, D., Denissen, J., Asendorpf, J. B., Neyer, F. J., & van Aken, M. A. G. (2014). A gentle introduction to Bayesian analysis: Applications to developmental research. *Child Development, 85*(3), 842–860.

van Noorden, R. (2015). Interdisciplinary research by the numbers: An analysis reveals the extent and impact of research that bridges disciplines. *Nature, 525*(7569), 306–307.

Walsh, E., Rooney, M., Appleby, L., & Wilkinson, G. (2000). Open peer review: A randomised controlled trial. *The British Journal Of Psychiatry, 176*(1), 47–51.

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica, 48*(4), 817–838.

Whitehurst, G. J. (1983). Interrater agreement for reviews for Developmental Review. *Developmental Review, 3*(1), 73–78.

Wirtz, M., & Caspar, F. (2002). *Beurteilerübereinstimmung und Beurteilerreliabilität: Methoden zur Bestimmung und Verbesserung der Zuverlässigkeit von Einschätzungen mitttels Kategoriensystemen und Ratingskalen [Inter-rater agreement and inter-rater reliability: Methods on analysis and improvement of the reliability of assessments by categorical systems and rating scales]*. Göttingen, DE: Hogrefe.

Wood, M., Roberts, M., & Howell, B. (2004). The reliability of peer reviews of papers on information systems. *Journal of Information Science, 30*(1), 2–11.

Yates, A. (1987). *Multivariate exploratory data analysis: A perspective on exploratory factor analysis*. Albany, NY: State University of New York Press.

Yousfi, S. (2005). Mythen und Paradoxien der klassischen Testtheorie (I): Testlänge und Gütekriterien [Myths and paradoxes of classical test theory (I): About test length, reliability, and validity]. *Diagnostica, 51*(1), 1–11.

Yuan, K.-H., & Bentler, P. M. (2000). Three likelihood-based methods for mean and covariance structure analysis with nonnormal missing data. *Sociological Methodology, 30*(1), 165–200.

Zyphur, M. J., & Oswald, F. L. (2015). Bayesian estimation and inference: A user's guide. *Journal of Management, 41*(2), 390–420.