Scientometrics (2013) 97:317–356
DOI 10.1007/s11192-013-1002-z

# On peer review in computer science: analysis of its effectiveness and suggestions for improvement

**Azzurra Ragone · Katsiaryna Mirylenka · Fabio Casati · Maurizio Marchese**

Received: 17 August 2012 / Published online: 7 April 2013


**Abstract** In this paper we focus on the analysis of peer reviews and reviewers behaviour in a number of different review processes. More specifically, we report on the development, definition and rationale of a theoretical model for peer review processes to support the identification of appropriate metrics to assess the processes main characteristics in order to render peer review more transparent and understandable. Together with known metrics and techniques we introduce new ones to assess the overall quality (i.e. ,reliability, fairness, validity) and efficiency of peer review processes e.g. the robustness of the process, the degree of agreement/disagreement among reviewers, or positive/negative bias in the reviewers' decision making process. We also check the ability of peer review to assess the impact of papers in subsequent years. We apply the proposed model and analysis framework to a large reviews data set from ten different conferences in computer science for a total of ca. 9,000 reviews on ca. 2,800 submitted contributions. We discuss the implications of the results and their potential use toward improving the analysed peer review processes. A number of interesting results were found, in particular: (1) a low correlation between peer review outcome and impact in time of the accepted contributions; (2) the influence of the assessment scale on the way how reviewers gave marks; (3) the effect and impact of rating bias, i.e. reviewers who constantly give lower/higher marks w.r.t. all other reviewers; (4) the effectiveness of statistical approaches to optimize some process parameters (e.g. ,number of papers per reviewer) to improve the process overall quality while maintaining the overall effort under control. Based on the lessons learned, we

A. Ragone · K. Mirylenka · F. Casati · M. Marchese (✉)
Department of Information Engineering and Computer Science, University of Trento, Via Sommarive, 5, 38123 Trento, Italy
e-mail: maurizio.marchese@unitn.it

A. Ragone
e-mail: ragone@disi.unitn.it

K. Mirylenka
e-mail: kmirylenka@disi.unitn.it

F. Casati
e-mail: casati@disi.unitn.it

suggest ways to improve the overall quality of peer-review through procedures that can be easily implemented in current editorial management systems.

**Keywords** Peer review · Quality metrics · Reliability · Fairness · Validity · Efficiency

**Mathematics Subject Classification (2000)** 62-07 · 62P25 · 91C99

## Introduction

Over the last few centuries, peer review has been considered a fundamental part of the scientific research and dissemination process (Zuckerman and Merton 1971). The review process is used to ascertain quality of scientific contributions and project proposals and to provide credits assignment as well as career advancement to researchers. Indeed, nearly every scientific journal bases its selection on peer review, and scientists spend a significant amount of their work time in reviewing papers [in computer science (CS), for example, it is common for senior researchers to review more than a hundred papers per year].

Surprisingly, especially given that peer review is used by scientists and it is such a fundamental part of researchers' daily life and career, there have been very few studies aiming at obtaining scientific evidence that peer review is a good way (or even the *optimal* way) to assess the truthfulness, quality, and potential impact of a scientific contribution or project proposal. In most cases we just proceed on the intuition or belief that it works. Even fewer are the scientific studies aiming at identifying how the review process can be made more efficient in terms of the trade-off between the review effort by the community and the validity of the review result.

In this paper we (i) search for scientific evidence that peer review "works" (or that it doesn't), and (ii) search for ways to improve the peer review process so that it can "work better". We do this by defining a set of metrics that are indicative of the quality of peer review processes: that is, aim at measuring how peer review "works". The purpose of such metrics is to help us understand and improve the peer review process along the following main dimensions: *reliability, fairness, validity* and *efficiency*. A *reliable* peer review process is, in our view, a process that provides a good prediction of the entire committee consensus opinion: i.e.,how far the practical choice of involving a reduced set of reviewers (typically three) in the review of each contribution is from the ideal case where everybody in the review committee evaluate the contribution. *Fairness*, in our approach, is related to the monitoring of the contribution distribution process to the reviewers: the more fair the process is the less it depends on the particular set of reviewers within the program committee (PC) to which it is assigned. *Validity* is related to the final result: a review process is valid if the best contributions are chosen. *Efficiency* is related to the time spent in preparing and assessing the contributions and to the statistical accuracy of the review results: a process is efficient if the best proposals are accurately chosen with minimal time spent both by authors in preparing the contribution and by reviewers in performing the reviews.

In order to achieve our research objectives we designed and performed a large-scale analysis of review data, and we tried to present and explain the results in a way that allows readers to easily form an intuition for what they mean in practice.

We are certainly not the first to perform an analysis of review data. In the related work section we review papers that are more closely related to our research and also refer to surveys on this topic. With respect to the literature, however, our analysis on peer review differs from others for the following main contributions:

- we formally define—and measure—metrics for the *validity* of peer review, which we believe mirror what people expect today from peer review, namely:
  (i) that peer review identifies papers/proposals that are scientifically correct and that are likely to have a scientific impact in the future, *and/or*
  (ii) that peer review identifies papers/proposals that the scientific community is likely to be interested in reading and considers worth research directions. We explicitly do not consider in our current work another important aspect of peer review, namely that of providing feedbacks to authors. Here, we are only concerned with the selection process and with quantitative data.
- we analyze a large dataset including nearly 3,000 contributions and ca. 9,000 reviews in the domain of CS.
- we investigate the efficiency of the review process, to identify how to improve the effectiveness of the process while maintaining the same overall reviewing effort.
- we introduce intuitive ways of expressing the results of our analysis, providing measures that are understandable and "actionable".

The majority of the data we have been able to collect come from the engineering field, mostly from CS. As such, we cannot claim that the results have general validity—and the same applies for every study in a single domain. In particular CS is rather different from the domains in which research on peer review has been more active (such as Physics and Medicine), as it is characterised by an high number of papers per researcher, most of which are not oriented at trying to model or understand how the world or the human body behaves, but rather try to propose new models, algorithms, or software. In CS it is rare the case where the review points out that a paper is "wrong". The typical criticism is that a paper is not that novel, that the problem attacked is not useful or applicable in practice or that it lacks sufficient theoretical or empirical validation. Moreover, conference publications enjoy greater status in CS than in other disciplines (Chen and Konstan 2010; Freyne et al. 2010). This being said, the results we obtained are, we believe, relevant and point to the need for further scrutiny on peer review as well as alternative models of review. Here is a short summary of our main findings:

- in all our available data, there is only a low correlation between the rankings of the review process and the impact of the papers as measured by *citations*; this is also true in the similar study of *a posteriori review* of the same contributions at a later time;
- the influence of the assessment scale on the way how reviewers gave marks;
- the disagreement among reviewers is a useful metric to check and monitor during the review process. Having a high disagreement means, in some way, that the judgment of the involved peers is not sufficient to state the value of the contribution itself. This metric can be useful to improve the quality of the review process as it can support the decision whether more reviewers are needed in order to improve the process reliability.
- it has always been possible to identify groups of reviewers that consistently give higher (or lower) marks than the others independently from the quality of the specific contribution they have to assess. The information coming from such analysis and possibly un-biasing procedures (like the one we proposed in this article) could be useful to review processes chairs to improve the fairness of the review process;
- we have shown that it is possible to devise statistical approaches to tune review process parameters to improve quality while keeping the overall effort under control.

The paper is structured as follows. In "Related work" section we provide a brief description of related work. "Approach to peer review analysis" section introduces our generic framework to the analysis of peer review processes, in particular our proposed metrics and model for peer review used in the subsequent analysis, while "Data set description" section illustrates the data set we used for our analysis. In "Quality: preliminary study" section we introduce a preliminary analysis on the quality of peer review process, while in the subsequent "Quality: reliability, Quality: fairness, Quality: validity" sections we present in detail for each analyzed dimension—reliability, fairness and validity—the proposed metrics, related results and lessons learned. In "Analysis of the efficiency of the peer review process" section we investigate a new dimension: efficiency, in order to suggest possible improvements in the review process. Conclusion and discussion on the findings close the paper.

## Related work

Peer review is probably one of the most debatable topic among scientists and has been widely studied in the last years, although no one of these studies can be considered comprehensive or conclusive. In fact, while peer review has been analyzed and studied by several researchers, we notice that such analysis are not straight comparable, as they refer to review processes coming from different disciplines and different journals. Indeed, sometime even analysis done in the same field can lead to contradictory results (Jefferson et al. 2002a). Even if peer review has been used as method of evaluation since Greek time (Barnes 1981; Spier 2002; Zuckerman and Merton 1971), the first journals that were selective in the choice of their manuscripts were the Journal des Savants and the Philosophical Transaction of the Royal Society of London, both founded in 1665 (Spier 2002; Zuckerman and Merton 1971). The first journal that introduced officially the peer review process as we know it today has been the Medical Essays and Observations, first published in 1731 (Spier 2002; Benos et al. 2007).

Recently, many scientists started to study the effectiveness and, more in general, the qualities and properties of peer review. A significant number of papers report that peer review is a process whose effectiveness "is a matter of faith rather than evidence" (Smith 2006), that is "untested" and "uncertain" (Jefferson et al. 2002a), and on which we know very little because scientists are rarely given access to relevant data. Lock (1994) claims that peer review can at most help detect major errors and that the criteria for judging a paper is to look at how often its content is used and referred to several years after publication. Other experimental studies put in doubt the ability of peer review to even spot important errors in a paper (Godlee et al. 1998). In general, however, although crude and understudied, peer review is still considered a process to which no reasonable alternatives have been found (Kassirer and Campion 1994; Smith 2006).

The various studies on peer review differ in which metric they evaluate and in the kind and amount of available data. Indeed, having precise objectives for the analysis is one of the key and hardest challenges as it is often unclear and debatable to define what it means for peer review to be effective (Jefferson et al. 2002b). In general we can divide the metrics in two groups: those aiming at determining the effectiveness or validity or peer review, and those aiming at measuring what authors consider to be "good" *properties* of peer review but that per se do not imply that peer review "works".

Among the first group, studies aim at assessing the ability of peer review to detect errors and the ability to predict the future impact, measured in terms of citation count. For what

concerns the ability to detect errors, a study was conducted by Goodman et al. (1994) who tried to measure the quality of the papers submitted to the Annals of Internal Medicine between March 1992 and March 1993 before and after the peer review process. They did not find any substantial difference in the manuscripts before and after publication. Indeed, they state that peer review was able to detect only small flaws in the papers, such as figures, statistics and description of the results. An interesting study was carried out by Godlee et al. (1998): they introduced deliberate errors in papers already accepted by the British Medical Journal (BMJ).[1] Godlee et al. report that the mean number of major errors detected was two out of a total of eight, while there were 16 % of reviewers who did not find any mistake, and 33 % of reviewers went for acceptance despite the introduced mistakes.

Citation count was used as a metric in peer review processes analysis mostly in studies by Bornmann and Daniel. A first study reports on whether peer review committees are effective in selecting *people* that have higher citation statistics, and finds that there is indeed such a correlation (Bornmann and Daniel 2005b). Another interesting study concerns preliminary review of papers by staff editors of journals, before sending the papers through a peer review process. Besides emphasizing that the opinions of staff editors is often uncertain and different from that of the reviewers, the study observes that "three-quarters of the manuscripts that were rated negatively at the initial internal evaluation but accepted for publication after the peer review had—when published—far above-average citation counts" (Bornmann and Daniel 2010b).

Our work in terms of effectiveness or validity focuses on three metrics (citations, a posteriori review, and ability to predict consensus opinion of the reviewers' community, e.g., through (dis)agreement analysis). None of the prior art uses these metrics to analyze the result of peer review on scientific papers in the same way. Many works—such as the interesting work by Bornmann and Daniel (2005b)—do consider one of them (citations) as a validity metric, but do not consider the rankings that come out of the peer review process and compare them with citations, which is one of the main aspect we consider in this paper.

Research aiming at measuring properties of peer review has been mostly focused on identifying *biases* and understanding their impact in the review process. Indeed, reviewers' objectivity is often considered a fundamental quality of a review process: 'The ideal reviewer,' notes Ingelfinger (1974), 'should be totally objective, in other words, supernatural'. Among the large number of contributions that had concern in bias detection, there are works that have found *affiliation* bias (meaning that researchers from prominent institutions are favored in peer review) (Ceci and Peters 1982), bias in favor of US-based researchers (Link 1998), or *gender* bias against female researchers (Wenneras and Wold 1997; Bornmann 2007; Ceci and Williams 2011). Another source of bias in peer review is *conflict of interest* bias, particularly in health related domains (Davidoff et al. 2001). Yet application of the multiple logistic regression models in Reinhart (2009) for the Swiss National Science Foundation (SNSF)—funding organization for basic research in Switzerland for the natural and social sciences—reveals that all potential sources of bias (gender, age, nationality, and academic status of the applicant, requested amount of funding, and institutional surrounding) are non-significant predictors.

Multiple logistic regression models of detecting the potential sources of bias in the peer review process were also used in Bornmann and Daniel (2005b) for defining the most frequently examined potential sources of bias, that could appear in selection of research fellowship recipients, such as: the applicant's gender, nationality, major field of study and

[1] "With the authors consent, a paper already peer reviewed and accepted for publication by BMJ was altered to introduce eight weaknesses in design, analysis, or interpretation" (Godlee et al. 1998).

institutional affiliation. Generalized linear mixed model of the log-odds ratio was used again to detect gender bias in SNSF (Bornmann et al. 2008a): here the authors pointed out that—using the generalized linear mixed model to detect unequal odds ratios—indications of potential sources of bias (such as gender, nationality, social status) can be examined not only for grant peer review but also for journal peer review. The Kruskal–Wallis test, a nonparametric version of one way Analysis Of Variance [ANOVA (Kruskal and Wallis 1952)] was used to detect the bias of the application order, namely the bias that is due to the fact that application was considered as first (Bornmann and Daniel 2005a) and the authors detected that there is an evidence that being first increases the probability of being accepted. In the area of measuring properties of peer review, our work differs from the prior art for the large scale of the analysis and for the identification of metrics and approaches that can be more intuitively understood by the reader. In addition we compute the *rating bias* (reviewers consistently giving higher or lower marks), which is a kind of bias that appears rather often, that is easy to detect, and that can be corrected with quite simple procedures (see "Quality: fairness" section). We also examine its effect on other properties of peer review processes.

A common way to identify bias is also to compare single and double-blind review. Single-blind review provides anonymity to the reviewers and is used to protect the reviewers form the authors' requital. Nowadays, single-blind review became the commonly used practice. Double-blind review, where identities of both authors and reviewers are hidden is also used sometimes. The purpose of it is to help the reviewers to assess only scientific achievements of the paper, not taking into consideration other factors and therefore not to be somehow biased. For instance, ACM SIGMOD (a conference on data management) organizes conferences where double-blind review is adopted. Analysis of the merit of the double-blind review process are so far contradictory. In Madden and DeWitt (2006), a set of statistics had been provided with the conclusion that double-blind reviewing make no impact on ACM SIGMOD publications. But later opposite results where published by Tung (2006), where he made two studies which indicate that double-blind review in ACM SIGMOD do have impact on the performance of famous person compared to VLDB (another popular conference series on database technology, but where all conferences are not double-blind). Moreover, it is in general difficult to enforce the double-blind review policy, as authors always introduce (deliberately or by mistake) elements that help reviewers to identify them (Katz et al. 2002).

Research on open peer review (where the reviewer's name is known to the authors) is at present very limited. Initial studies showed that open reviews were of higher quality, were more courteous and reviewers spent typically more time to complete them (Walsh et al. 2000; Bornmann et al. 2012; van Rooyen et al. 1999). We did not come across any study that compares open versus blind reviews in terms of bias estimation.

Now, scientists and editors are exploring alternative approaches to tackle some of the pervasive problems with traditional peer review (Akst 2010). This include enabling authors to carry reviews from one journal to another, posting reviewer comments alongside the published paper, or running the traditional peer review process simultaneously with a public review. The ACM SIGMOD conference has also been experimenting variations of the classical peer review model where papers are evaluated in two phases, where the first phase filters out papers that are unlikely to be accepted allowing to focus the reviewers' effort on a more limited set of papers. In this paper we provide a model for multi-phase review that can improve the peer review process in the sense of reducing the review effort required to reach a decision on a set of submitted papers while keeping the same quality of results (see "Analysis of the efficiency of the peer review process" section).

As we have already mentioned, one of the main issue in peer review analysis is to have access to the data. We experienced the same problems in collecting our data. However, our work differs from the others we mentioned in this section for the scale of the analysis, in terms of number of papers and reviews taken into consideration. In other works, authors have been restricted to analyze only 1–2 conferences, grant applications processes or fellowships. Just to name a few: Reinhart (2009) analyzed 496 applications for project-base funding; Bornmann and Daniel (2005a) studied the selection process of 1,954 doctoral and 743 post-doctoral applications for fellowships; Bornmann et al. (2008b) analyzed 668 applications for funding; Godlee et al. (1998) involved in their experiments 420 reviewers from the journal's database; Goodman et al. (1994) analyzed 111 manuscripts accepted for publication. A very recent work (Cabanac and Preuss 2013) has been published where the authors have analysed 42 peer-reviews conference in Computer Science, but focusing only on the order effects in the bids for paper reviews.

In the present analysis, we succeeded in collecting review data from 10 conferences, for a total of 9,032 reviews, 2,797 submitted contributions and 2,295 reviewers.

## Approach to peer review analysis

This section presents our proposed framework for the analysis of peer review data. We first discuss the metrics we aim to measure from the available data and explain why we focus on them. Then we introduce the model and notation to describe peer review processes that we use in the later sections for our analysis.

### Metrics for peer review

We propose two classes of metrics for peer review: (1) metrics to study if peer review "works" and (2) metrics to identify good properties of peer review.

The first class of metrics is really at the heart of the problem of finding scientific evidence in support of peer review. Defining what we want out of peer review is considered a challenging topic in itself (Smith 2006; Jefferson et al. 2002a) and it is often a matter of opinions. Both from our experience and from the literature, e.g., Lock (1994), Godlee et al. (1998), Kassirer and Campion (1994) Smith (2006), peer review is considered to have one or more of the following goals:

1. Identify and select papers that are likely to have a relevance and *impact* in the future. In the case of projects, select proposals that are more likely to have an impact on science, business, or the society at large.
2. Identify papers that are likely to be of *interest* to the readers (for journals) or attendees (for conferences).
3. Spot *errors* in the paper and give feedbacks to authors so that they can realize a better paper.

In the following we do not focus on the third item, both because in our analysis we are concerned with understanding the ability of peer review to identify and select good papers and because this specific point have already been studied in the past (Goodman et al. 1994; Godlee et al. 1998).

Trying to measure impact is one of the highly debated topics in scientific dissemination and evaluation, also because it is one of the main factors considered in evaluating job applications in the academia and research labs (so it has a significant and direct impact on

people's life). To identify commonly accepted metrics for this, we looked both at how committees evaluate scientific impacts of candidates and at how 10-years award committees operate when they need to look back at papers published 10 years earlier. Selection and evaluation committees consider number of publications (possibly weighted by impact factor or other means) or, more recently, citation-based metrics (citation count, *h*-index, *g*-index, etc.) (Krapivin et al. 2010). In some cases, selected publications of short-listed applicants are reviewed by the committee members to further assess their quality. Our team has been involved in supporting 10-years awards committees for major conferences in CS, and the criteria there are not dissimilar: papers are screened by citation counts and then looked at by the committee.

In our work we take the same stand in finding metrics for impact: we consider citation count, and we consider an a-posteriori review of the papers (and even of extended version of the papers that report on further detail and on further elaboration of the work). These are the baselines over which we assess the validity of peer review and in the following sections we describe in detail how we measure them and compare with peer review data.[2]

In essence, we try to see how the ranking coming out of peer review is close or far with respect to those coming out of citation counts or additional reviews of the same work (sometimes more detailed versions of the same work).

As for the second goal above (interestingness), the way we measure it is by looking at the ability of a review process to predict the average opinion of the entire PC. In the domain of information engineering and CS, PCs of conferences are often very large, and in important conferences they typically range from 100 to 300 members, typical with an hierarchical organisation (e.g., general chair, regional chairs, meta-reviewers, reviewers). As such, the PC is a good approximation of the community of interest of the conference, and therefore estimating the opinion of the PC is a reasonable approximation of the interestingness of a paper for the target community.

Our approach to analyze review data is therefore driven by the needs of: (i) measuring or estimating the above metrics and their correlation with peer review; (ii) understanding and explaining the results, in addition to provide the numbers, also in a way that is intuitive and that give readers a feel for how well peer review works.

In our analysis we also compute other metrics (for example, metrics of *agreement* among reviewers (often called *reliability* in the literature) and *robustness* and this is because they all contribute to our main goal of establishing the validity of peer review.

## Peer review model

We present here a model for peer review that covers many types of submission and review procedures, including conferences submissions project proposals, and PhD thesis proposals assessment. The model focuses on *bulk* submissions where several proposals are sent by a deadline and are evaluated by a committee. This is to fit the analysis needs for the data we have which is mostly conference data. In this regard, we observe that in the area of information engineering and CS, where most of the data comes from, conferences are the primary outlet for publications and are often regarded as reputed or even more reputed than journals (Chen and Konstan 2010; Freyne et al. 2010). Conference data also makes it

[2] Notice that this pragmatic choice does not imply that the authors believe blindly in citation count as being the only measure of impact. Indeed prior art has shown that it has some flaws (Krapivin et al. 2010) and could be extended to other novel metrics like number of downloads (Li et al. 2012) or other alternatives metrics (Bollen et al. 2005). However, we adopt it as it is a commonly accepted and accessible metric.

possible to have information on rankings of papers by a committee, and as we will see this enables certain kinds of analyses that are useful for understanding peer review processes.

In bulk submissions, peer review procedures usually proceed along the following steps. Authors submit a set $\mathcal{C} = \{C_z\}, \quad z = 1, 2, \ldots, N$ of *contributions* for evaluation by a group $\mathcal{E}$ of *experts* (the peers, also called reviewers). Each contribution is assigned to a number of reviewers. Its flow through the process *may* be supervised by senior reviewers (a set $\mathcal{SR} \subset \mathcal{E}$ of distinguished experts that analyze reviews and help chairs take a final decision on the contribution). One typical setting for conferences is to have three reviewers and zero or one senior reviewer per paper. In the general case, each contribution may be assigned to a variable number of reviewers.

The review occurs in one or more *phases*. We denote with $N_{\mathrm{P}}$ the total number of phases. In each phase $p_k$, contributions are assigned, marks are given, and contributions that are allowed to proceed to the next phase are selected. The next phase may or may not require authors to send a revised version of the contribution. At the end of each phase there is a discussion over the reviews (possibly involving author feedback). Some processes require the discussion to end in a "consensus" result for the final mark (this is typically the case for example in PhD thesis proposals assessment, where the committee members must come to a consensus result). In all cases, the discussion results in a decision on whether each contribution is accepted or not. The entire process is supervised by a set $\mathcal{CH} \subset \mathcal{E}$ of *chairs*.

For example, a typical conference has a one-phase review, with discussion at the end leading to acceptance or rejection of each paper. Some conferences, such as ACM SIGMOD in the past, had a 2-phase review process where in the first phase each paper was assigned to two reviewers and only papers that have at least one accept mark go to phase two and are then assigned to a third reviewer. This is done to minimize the time spent in reviewing (or, seen differently, to focus the effort on papers that are not clear rejects). Regarding journal review processes, the editor-in-chief often acts as a first filter always in order to minimize the review workload.

Given the above, we model a *phase* $p_k$ of a peer review process as follows:

**Definition 1** A phase $p = (\mathcal{C}, \mathcal{E}, \mathcal{M}, \pi, \gamma, \sigma, \rho, \mathcal{A})$ of a peer review process consists of:

- a set $\mathcal{C} = \{C_z\}, \quad z = 1, 2, \ldots, N$ of *contributions* submitted for evaluation;
- a set $\mathcal{E}$ of *experts*, which includes:
  - a set $\mathcal{CH} \subset \mathcal{E}$ of *chairs* that supervise the review process
  - a set $\mathcal{SR} \subset \mathcal{E}$ of distinguished experts (sometimes called senior reviewers) that analyze reviews and help chairs take a final decision on the contribution
  - a set $\mathcal{R} \subseteq \mathcal{E}$ of experts that act as reviewers of the contributions

  and s.t. $\mathcal{CH} \cup \mathcal{SR} \cup \mathcal{R} = \mathcal{E}$
- a set $\mathcal{M}$ of mark sets, $\mathcal{M} = \{M_1, \ldots, M_q\}$, where $q$ is number of criteria and for each mark set a total order relation $\leq$ always exists. *Acceptance threshold*, denoted by $t^j$ may be defined for each mark set $M_j$, $j = 1, 2, \ldots, q$.
- an assignment function $\pi : \mathcal{C} \longrightarrow \mathcal{P}(\mathcal{R}) \times \mathcal{P}(\mathcal{SR})$ assigning each contribution to a subset of the reviewers and a subset of the senior reviewers (an element of the respective powersets $\mathcal{P}$).
- a scoring function $\gamma : \{c, r\} \mapsto M_1 \times \cdots \times M_q$ such that $c \in \mathcal{C}$ and $r \in \pi(c)$. This function models the marks assigned by each reviewer.

- a score aggregation function $\sigma : M_1 \times \cdots \times M_q \longrightarrow \mathbb{R}$. This models the way in which in some review processes one can derive an aggregate final mark based on the individual marks.
- a ranking function $\rho : \mathcal{C} \longrightarrow \mathbb{N}$.
- a subset $\mathcal{A} \subseteq \mathcal{C}$ that denotes the accepted contributions.

In the next sections we will introduce a number of novel metrics based on the above model and useful for our exploration and analysis of the different dimension of quality of the peer review process.

## Divergence

As the final result of the peer review process in conferences and candidate selection often includes a ranking (sometimes publicly published; other times only the list of accepted contributions is published), in our subsequent analysis, we often need to quantitatively assess the difference—in terms of concrete effects—between two rankings coming from different review processes or from a review process and a ranking determined from another quality metric. Examples are the difference between the ranking (of the same contributions) coming from the peer review process and the one, a posteriori, coming from citations; or the ranking coming from a preliminary evaluation of e.g. an extended abstract and the one from a subsequent and deeper evaluation, e.g. a full paper.

In the literature, the typical metric for measuring a difference between two rankings is the Kendall $\tau$ rank correlation coefficient (Kendall 1938). The Kendall $\tau$ coefficient is also used as a test statistic in a statistical hypothesis test to establish whether two rankings may be regarded as statistically dependent. This metric, however, computes the difference in the exact position of the elements between two ranks, while in the review process the main issue is not to be in 3rd or 10th position, whether to be accepted versus to be rejected.

To better capture this specific property and to give readers a more intuitive way to grasp the distance, we also use a measure called *divergence* to compute the *distance* between rankings. We next give the formal definition of divergence following Krapivin et al. (2010), that was adapted to our scenario.

**Definition 2** (*Divergence*) Let $\mathcal{C}$ be a set of submitted contributions, $n = |\mathcal{C}|$ the number of submissions, $\rho_i$ and $\rho_a$, respectively, the ideal ranking and the actual ranking, and $t$ the number of accepted contributions according to the actual ranking. We call divergence of the two rankings $Div_{\rho_i,\rho_a}(t, n, \mathcal{C})$ the number of elements ranked in the top $t$ by $\rho_i$ that are not among the top $t$ in $\rho_a$.

The normalized divergence $NDiv_{\rho_i,\rho_a}(t, n, \mathcal{C})$ is equal to $\frac{Div_{\rho_i,\rho_a}(t,n,\mathcal{C})}{t}$, and varies between 0 and 1.

Through this metric it is possible to assess how much the set of the contributions after one ranking procedure *diverges* (is different) from the set of contributions after another ranking procedure. Figure 1 schematically depicts three different divergence curves resulting from the fact that (i) the two rankings are *correlated*; (ii) they are *independent* (the analytical results for this case is the straight line in the figure);[3] (iii) they are *inversely correlated*.

[3] When the second ranking is random, the formula for the divergence can be expressed analytically as $NDiv_{\rho_i,\rho_a}(t, n, \mathcal{C}) = \sum_{i=0}^{t} p_t(i, n) w_i$, where $p_t(i, n) = \frac{C_t^i C_{t-i}^{n-t}}{C_t^n}$ and $w_i = \frac{t-i}{t}$.

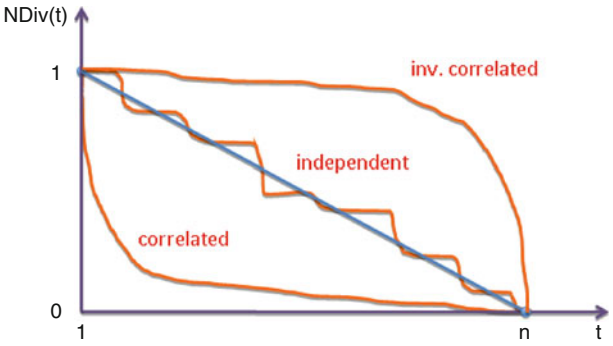


**Fig. 1** Examples of different divergence curves. The index $t$ in the X axis represents the top $t$ contributions that are considered by the divergence. The $y$-axis shows how many items among the top $t$ in the first ranking are also among the top $t$ in the second ranking. The divergence values ($y$-axis) are normalized by $t$

In the rest of the paper we will use this metric to assess the effect of variations in the peer review process. In particular, the effect of unbiasing algorithms ("Quality: fairness" section) and the overall validity of the peer review process ("Quality: validity" section).

## Data set description

In this work we have analyzed data gathered from ten conferences that took place from 2003 to 2010, whose topics were related to the CS domain (Table 1). Among these, there are five conferences (C1, C3, C8–C10) which took place in the period from 2003 to 2006, therefore they are "old" enough for checking the *impact* of the accepted papers during the years ("Quality: validity" section).

As the data we used for the analysis are confidential we cannot disclose the name of the conferences. So we use an ID to identify the conference and we only report approximate numbers in Table 1 to guarantee the anonymity of the original data.

**Table 1** Description of the conference data

| Conf. ID<br>1 | No. of papers<br>2 | Rating scale<br>3 | No. of RPP<br>4 | No. of PPR<br>5 | Ac. rate (%)<br>6 |
|---|---|---|---|---|---|
| C1 | 900 | $1, 2, \ldots, 10$ | 3–4 | 1–4 | 21 |
| C2 | 250 | $1, 2, \ldots, 7$ | 3–4 | 1,2,9,10 | 16 |
| C3 | 700 | $0, 0.5, \ldots, 5$ | 3 | >3 | 27 |
| C4 | 200 | $0, 0.5, \ldots, 5$ | 3 | 1–2 | 26 |
| C5 | 200 | $-3, -2, \ldots, 3$ | 3–4 | 1,2,7 | 31 |
| C6 | 150 | $1, 2, \ldots, 5$ | 3–4 | 1,2,>5 | 33 |
| C7 | 120 | $-3, -2, \ldots, 3$ | 3–4 | 6–8 | 22 |
| C8 | 150 | $1, 2, \ldots, 7$ | 3 | 4,5 | 45 |
| C9 | 40 | $1, 2, \ldots, 4$ | 2–4 | 2,4,5,7 | 51 |
| C10 | 100 | $1, 2, \ldots, 7$ | 2–3 | 5–6 | 55 |

In Table 1 for each conference we show (i) the conference ID; (ii) the approximate number of papers submitted to the conference; (iii) the scale used by reviewers to assign marks to papers; (iv) the typical number of reviews per paper (RPP) and (v) of papers per reviewer (PPR) and (vi) the acceptance rate of each conference. The RPP and PPR reported in Table 1 are the most frequent values for each conference (specifically, those occurring for more than 10 % of the times for that conference). It is in fact quite normal that in one conference a paper is reviewed on average by three reviewers, but sometime, in particular for some disputed papers, there could be more than three reviewers. So we see from the table that while the typical number of RPP is constantly—in our data set—around 3–4, the number of papers assigned to reviewer (PPR) is more variable and some reviewers get an higher number of papers to review. All together our dataset consists of 9,032 reviews, 2,797 submitted contributions and 2,295 reviewers.

## Quality: preliminary study

Before starting the detailed description of our quality analysis, we first describe here a simple statistical analysis of our data set in order to put our results in the appropriate context. Moreover, we start to use the methodology to investigate the differences among different rankings described in "Divergence" section and apply it to a robustness analysis of the peer review process.

### Mark distribution analysis

A very simple analysis is to look at the distributions of the marks (following the experimental scientist's motto: "always look at your data"). Analyzing the distribution of marks in review processes with different mark scales, we notice that the way reviewers give marks can be influenced by the scale itself. In Fig. 2 we plot distribution of marks from processes from three conferences where marks range:[4]

(1) from 1 to 10 (no half-marks);
(2) from 1 to 7 (no half-marks);
(3) from 0 to 5 with half-marks.

In case (1) the distribution is slightly positively skewed and this finding is also confirmed in the recent study in Cabanac and Preuss (2013).

In case (2) reviewers tend not to give the central mark (4 in this case), but to give lower or higher marks (in this specific case the most frequent mark is 2). It seems that the use of the scale (2) "supports" the reviewer to take a decision, avoiding the central mark which corresponds to a *neutral* mark. Indeed in a (1,7) scale it is easy to identify mark 4 as borderline, and this is somehow reflected in the observed distribution.

In case (3) we notice that reviewers tend *not* to give half marks, indeed the curve has many oscillations with minima corresponding to such half marks; while if we consider case (1)—essentially the same scale, i.e a doubled scale with integer marks from zero to ten instead of half marks—the mark distribution appears concentrated around the middle of the ratings scale.

As a general remark, we were surprised by how much the mark distribution changes depending on the specific scale chosen.

[4] In Fig. 2 the different scales have been normalized in the $x$-axis.

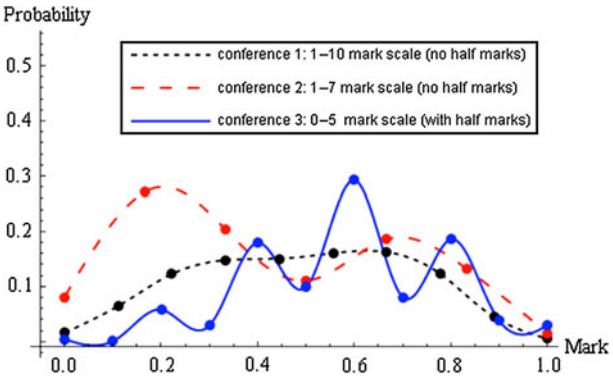


**Fig. 2** Examples of normalized mark distributions in conferences with different scales

## Robustness analysis

A mark variation sensitivity analysis is useful in order to assess if a slight modification of the value of marks could bring a change in the final decision about the acceptance or rejection of a contribution. The rationale behind this analysis is that we would like the review process to be *robust* to minor variations in one of the marks. When reviewers need to select within, let's say, a 1–10 score range of criteria, often they are in doubt and perhaps sometimes carelessly decide between, for example, a seven or an eight (not to mention the problem of different reviewers having different scoring standards, see "Quality: fairness" section).

With a robustness analysis we try to assess how much a slight difference in the mark value can affect the final (positive or negative) assessment of a contribution. To this end, we apply a stochastically positive/negative perturbation $\delta$ to each mark. This perturbation must be a multiple of a mark granularity $g$ of the process (e.g., $g = \pm 0.5$), $\delta = i * g$, where $i = 1, 2, 3$. The value of $\delta$ is therefore chosen according to the specific rating scale of the conference.We then rank the contributions with respect to these new marks and repeat the simulation a number of runs to collect proper statistical data (i.e. mean value and standard deviation) for every simulated case.

Intuitively, what we do with the mark variation is a naive way to transform a mark into a random variable with a certain variance, reflecting the indecision of a reviewer over a mark. Actually, we would like to know how this variation could change the fate of the papers (the papers are ranked according to their marks), i.e. how many papers that are above (below) the threshold, e.g., acceptance threshold, will appear below (above) the threshold after the variation.

Here we assume that the main criteria of the PC for accepting/rejecting a paper is acceptance rate and not paper quality only. Hence, we assume that the number of accepted papers must remain the same, regardless of paper marks.

To this end, we have developed an algorithm that computes the percentage of the papers which have changed their fate within a specific *window* after a small variation of the marks. The size of the window can be arbitrary chosen depending on the number of contributions for the specific conference. We are also interested in studying how the status (accepted/rejected) of the papers changes with respect to different "acceptance" thresholds, thus we used sliding windows centered on a variable "acceptance" threshold in order to compute

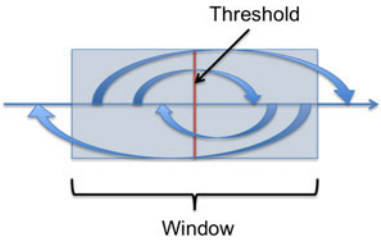


**Fig. 3** Schema for the robustness analysis for a fixed window of papers and corresponding threshold. The *rounded arrows* indicate the types of *relevant* cases for a paper within the window to change its fate due to a perturbation on its marks

the percentage of papers within the window that change their fate (a schema of the procedure is represented in Fig. 3). We assume that a paper changes its fate after perturbation if the comparison of the original paper ranking and the perturbed ranking shows that:

1. the paper has moved from one side of threshold to another side within the window, e.g. the two smaller rounded arrows in Fig. 3;
2. the paper has moved from one side of the threshold within the window to another side of the threshold outside the window, e.g. the two larger rounded arrows in Fig. 3.

We used the above procedure to compute the percentage of papers that change their fate for the two largest conferences in our dataset (C1 and C3) for a fixed sliding window of 100-papers. We have investigated the robustness of the two review processes considering a range of thresholds (i.e. centers of the sliding windows that represents different number of accepted contributions) in increments of 50 accepted papers. The results of this computation are shown in Fig. 4. We chose $g = 1$ and $g = 0.5$ for C1 and C3 correspondingly to the respective rating scales.

The analysis shows that the percentage of papers that changed their fate due to a perturbation of the marks is lower in the beginning and in the end of the ranking list. This reflects the obvious fact that papers at the top and at the bottom of the ranking have very clear marks (e.g., close to 10 at the top and close to 1 at the bottom if the range is between 1 and 10): in this case the applied perturbations have a reduced effect on their fate. Nevertheless, in both conferences and in these ranges (top and bottom) the percentage of affected paper for even the smaller perturbation is around 15–20 %. In the middle part of the ranking, already the minimal perturbation of $\delta = g$ can lead to a 30–35 % change in the fate of papers in a wide range of acceptance thresholds.

We believe that calculation of these kind of robustness curves may help PCs to make a more informed decision about the acceptance threshold, as they can estimate the influence of perturbation of different thresholds on the final ranking results. For instance, in the case of the analyzed conferences we found that even the smallest perturbation $\delta = g$ can change the fate of ca. 33 % of papers for C1 with the nominal acceptance rate of 21 %—ca. 180 accepted papers—and of ca. 38 % for C3 with the nominal acceptance rate of 27 %—ca. 210 accepted papers.[5]

If the review process chairs would have known these results (and these type of calculations could easily be a feature of current conference management systems) they could have decided to conduct additional reviews to make a better distinction between papers around the selected thresholds.

[5] See Table 1 for the nominal acceptance rate for all conferences.

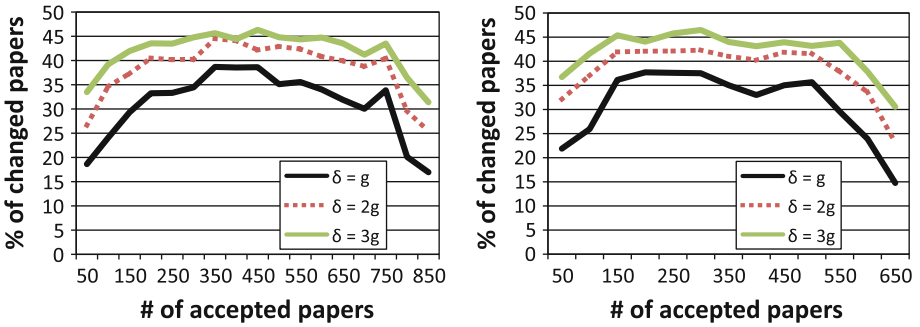


**Fig. 4** Robustness curves for conferences C1 (*left*) and C3 (*right*) with fixed sliding window = 100 observations. Granularity is $g = 1$ for C1 and $g = 0.5$ for C3. The $x$-axis shows the (variable) number of accepted paper used in the computation

Preliminary analysis: lessons learned

From the above analysis we can derive some useful insights and recommendations:

- Monitoring mark distribution is a simple analysis but very useful in order to understand "how" reviewers use the scale and if there are nonfunctional uses of the scale itself. It could also be convenient for the program chairs to design and adapt the scale for specific purposes.
- The mark distribution analysis can be coupled with a robustness analysis of the whole process in order to investigate how stable the process is w.r.t perturbations in the marks. Such analysis can provide review chairs indication if and for which papers conduct additional reviews to make a better decision on papers close to the acceptance/rejection threshold.

## Quality: reliability

Human decisions are classified as reliable when different persons come to the same or similar conclusions (Ebel 1951; Reinhart 2009). In traditional analysis of opinion reliability, the degree of agreement between people opinions is determined. In this section we first perform a classical analysis in this dimension, then we introduce and investigate some new metrics (namely, *disagreement* and *band agreement*) to assist us in the investigation of the reliability from different points of view.

The rationale behind these analyses and related metrics is that in a review process we expect some kind of agreement between reviewers. While it is natural that reviewers have different opinions on a given contribution, however, if the marks given by reviewers are comparable to marks given at random, then the results of the review process are also random, which defeats the purpose. The reasons for having reviewers (and specifically for having the typical number of 3 reviewers per contribution) is to evaluate based on consensus or majority opinion.

In the literature various methods are commonly used for the statistical measure of reliability: e.g. the Kappa coefficient proposed by Cohen (1960), its extension the weighted Kappa proposed in Fleiss (1971) and the *Intraclass Correlation Coefficient* (ICC)

(Bornmann and Daniel 2010a; Reinhart 2009; Cicchetti et al. 2008; Montgomery et al. 2002). In our work we have chosen to use ICC as its usefulness and applicability in the social sciences has been demonstrated in many applications (Ebel 1951). Intraclass Correlation Coefficient was first introduced by Fisher (1925). ICC returns the value 1 for complete agreement while the value 0 corresponds to the agreement level for a random process. The technique for computing ICC is based on the framework of the analysis of variance (ANOVA) and the estimation of a number of variance components.

There exist various forms of the ICC depending on the particular target process: one-way random effects model, two-way random effects model with or without interaction, two-way mixed model with or without interaction, and average score ICCs for one-way and two-way models. A detailed discussion on the distinction between the different coefficients could be found in McGraw and Wong (1996), Shrout and Fleiss (1979), Bartko (1966, 1974), Donner (1986). In our study we have used the average score ICC for one-way model (McGraw and Wong 1996) for assessing the inter-rater reliability in our 10 conferences (see Table 2). The average score ICC is in fact used in the case when the decision about the object in consideration is based only on the average mark (as in our case). The one-way model is used when the identity of the rater is not important: in our case we are interested only in the mark correlations not in reviewer feature correlations (e.g. identity, past behavior etc.).

In order to interpret the results collected in Table 2, we can recall that in the field of biostatistic analysis reliability measures below 0.4 are rated as poor, between 0.4 and 0.59 as fair and above 0.6 as high (Cicchetti and Sparrow 1981). The same classification has been used to access the reliability of reviewers recommendations for grant applications in biology and medicine in Reinhart (2009).

According to the above classification, in our case we have six conferences with ICC $> 0.6$, i.e. with significant correlation, 3 conferences with a fair correlation ($0.4 <$ ICC $< 0.59$) and one conference with poor correlation among raters (ICC $< 0.4$). In Table 2 we also report two other statistical parameters useful to evaluate the statistical significance of the results, namely: the 95 % confidence intervals (CI) for the computed ICC and the related probability values (i.e. $p$-value). According to the computed $p$-values reported in Table 2, all computed correlations are statistically significant ($p < 0.05$) regarding the following null hypothesis $H0$: ICC $= 0$ ($H1$: ICC $> 0$). In other words, the obtained ICC coefficients are not equal to zero (agreement of a random process) with a probability of 95 %.

**Table 2** Intraclass correlation coefficient, 95 % CI and related $p$-value for reviewers' ratings scores, sorted in decreasing order of ICC value

| Conference ID | ICC | 95 % CI | $p$-value | Reliability level |
|---|---|---|---|---|
| C6 | 0.76 | (0.68; 0.82) | 9.02E−27 | High |
| C9 | 0.72 | (0.46; 0.85) | 8.54E−05 | High |
| C7 | 0.63 | (0.5; 0.73) | 2.94E−11 | High |
| C5 | 0.61 | (0.49; 0.7) | 2.43E−13 | High |
| C1 | 0.61 | (0.56; 0.65) | 2.22E−63 | High |
| C8 | 0.60 | (0.46; 0.7) | 3.07E−10 | High |
| C2 | 0.57 | (0.47; 0.66) | 5.81E−15 | Fair |
| C4 | 0.52 | (0.39; 0.62) | 3.03E−10 | Fair |
| C10 | 0.45 | (0.16; 0.63) | 0.00254 | Fair |
| C3 | 0.39 | (0.3; 0.46) | 1.12E−14 | Poor |

However, as the classification of the intervals for ICC were defined in an arbitrary way, there is no clear evidence of the correctness of their applicability in our case. Therefore, in the following sections we have investigated additional metrics measuring the agreement among reviewers in order to better understand the difference in the various reviewing processes in our dataset.

Disagreement

In this section we look at the problem of measuring agreement among reviewers from a different perspective, i.e. measuring the disagreement among them as a refinement of inter-rater agreement coefficients specific for conference peer review processes. We underline here that disagreement per se is not necessarily a bad thing: novel and non-obvious ideas are often controversial (Grudin 2010; Birman and Schneider 2009), and different reviewers may give different importance to separate contributions in the paper. The problem of disagreement surfaces when papers are rejected merely because of averaging the scores of different reviewers.

Here, we compute first how much the marks of a reviewer $i$ differ from the marks of the other $r_z - 1$ reviewers for a specific criterion j and for a specific contribution $z$ (Definition 3). Then we compute the disagreement of a reviewer $i$ for a specific contribution $z$ for all the criteria (Definition 4) and average disagreement for each contribution through all its reviewers (Definition 5) and, finally, over all the contributions (Definition 6).

**Definition 3** (*Disagreement of a reviewer on a criterion and on a contribution*) Let $j$ be a criterion and $M^j_{i_z}$ be the mark set by the reviewer $i$ for the criterion $j$ assigned to a contribution $z$. Then, a disagreement $\phi^j_{i_z}$ among $r_z$ reviewers on a contribution $z$ is the euclidean distance between the mark given by the reviewer $i$, and the average $\mu^j_{i_z}$ of those given by the other $r_z - 1$ reviewers:

$$\phi^j_{i_z} = |M^j_{i_z} - \mu^j_{i_z}|$$

with:

$$\mu^j_{i_z} = \frac{1}{r_z - 1} \cdot \sum_{k=\{1,\ldots,r_z\}\{i_z\}} M^j_{k_z}. \tag{1}$$

**Definition 4** (*Disagreement of a reviewer on a contribution*) Let $q$ be the number of criteria in a review phase, then the disagreement of a reviewer $i$ on a contribution $z$ is:

$$\gamma_{i_z} = \frac{1}{q} \cdot \sum_{j=1}^{q} \phi^j_{i_z} \tag{2}$$

**Definition 5** (*Disagreement of a review phase on a contribution*) Let $r_z$ be the number of reviewers in a review phase on a contribution $z$, then the disagreement of a review phase on a contribution is:

$$\Gamma_z = \frac{1}{r_z} \cdot \sum_{i=1}^{r_z} \gamma_{i_z}. \tag{3}$$

**Table 3** Normalized average disagreement for all conferences sorted by decreasing order of the differences between computed and reshuffled disagreements (column 5). Each experiment consisted of 10 independent runs of the simulations. Average standard error is ca. 0.05 in all simulations

| Conf. ID | Computed | Reshuffled | Random | Difference between computed and reshuffled disagreement (%) | Difference between computed and random disagreement (%) |
|---|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 | 6 |
| C9 | 0.30 | 0.43 | 0.54 | 30.2 | 44.4 |
| C6 | 0.26 | 0.37 | 0.52 | 29.7 | 50.0 |
| C7 | 0.25 | 0.34 | 0.48 | 26.5 | 47.9 |
| C5 | 0.26 | 0.35 | 0.45 | 25.7 | 42.2 |
| C2 | 0.30 | 0.40 | 0.49 | 25.0 | 38.8 |
| C8 | 0.34 | 0.44 | 0.51 | 22.7 | 33.3 |
| C1 | 0.28 | 0.36 | 0.43 | 22.2 | 34.9 |
| C10 | 0.26 | 0.32 | 0.48 | 18.8 | 45.8 |
| C4 | 0.22 | 0.26 | 0.52 | 15.4 | 51.1 |
| C3 | 0.26 | 0.29 | 0.44 | 10.3 | 40.9 |

**Definition 6** (*Disagreement of a review phase*) Let $n$ be the number of papers in a review phase, then the disagreement over all the papers is:

$$\Psi = \frac{1}{n} \cdot \sum_{z=1}^{n} \Gamma_z. \tag{4}$$

In the second column of Table 3 we present the normalized computed average disagreement of a review phase (Definition 6) for all 10 conferences. We have normalized the disagreement value in order to allow direct comparisons among different conferences. To assist in the interpretation of the results, we also report in the same table, the average disagreement we have obtained in two simulations:[6]

1. *Reshuffle* experiment (third column): where we have randomly exchanged the actual marks given by the reviewers;
2. *Random* experiment (fourth column): where we have generated a new random (uniform) distribution of marks in the available range of marks unrelated with the actual marks distribution in our data set.

The reshuffle experiment mimics the case in which one reviewer is marking a certain number of contributions, but her/his marks are randomly given to other unrelated contributions, while her/his reviewed contributions get the marks of other randomly selected reviewers. So we are sampling from the actual mark distribution function, i.e. the one of the analyzed review process, but we randomize the association between marks and contributions.

[6] Also in these numerical experiments we repeated the simulations a number of runs (typically 10) to collect proper statistical data (i.e. mean value and standard deviation) for each experiment.

We would have expected these reshuffling disagreements to be much higher than the one computed with properly assigned marks since, again, we would have expected higher correlations between the opinions of a team of experts. For conferences C3, C4 and C10 the differences between original and reshuffled disagreements (fifth column in Table 3) are only 10.3, 15.4 and 18.8 % correspondingly, while for other conferences they vary from 23 to 30 %). On the other hand, the computed average disagreement is constantly lower than the random one, from 33 to 51 % (the sixth column in Table 3). This is expected since we would hope that a group of experts in a domain would tend to agree better than a completely random process. These results are consistent with the previous analysis where C3, C10 and C4 had the lowest ICC.

Moreover, we applied the Welch's test (Welch 1947) to verify if the differences between the computed disagreement value and the one based on reshuffled marks were statistically significant. As detailed in the following, the test shows that the differences are indeed significant. Welch's test was applied to compare the mean values (Definition 6) for two pairs of samples formed by disagreement on contribution (Definition 5). The first pair of samples was:

1. values of disagreement on a contribution (Definition 5) calculated for original marks;
2. values of the same metric calculated for reshuffled marks.

The second pair of samples used also the disagreement on contribution but compared original and random marks. The tests showed that for all the conferences the mean of the sample based on original marks is lower than the means of the samples based on reshuffled and random marks with confidence level $\alpha = 0.05$ [corresponding $p$-values varied from essentially 0 ($2.2 \times 10^{-16}$)–0.009], i.e. the differences are statistically significant.

### Band agreement

In order to further explore the reliability dimension, we introduced a new measure that we coined *band agreement*. Our goal here is to study the agreement in the decisions of reviewers about very good and very bad papers.

The approach is based on clustering review marks in "bands" and measuring the probability of giving a mark from a particular band in the condition that a mark from another band has already been given.[7]

To this end, all marks have been divided into non overlapping bands (see Table 4): (i) strong reject; (ii) weak reject; (iii) borderline; (iv) weak accept; (v) strong accept. Then, we have computed the overall probability of a paper to belong to each group.

We have analyzed the behavior of reviewers in three different conferences (C1–C3) with a high number of papers with marks from each "band" (Table 4) and different levels of ICC:

- C1 without threshold for marks for acceptance and with "high" ICC agreement;
- C2 without threshold for marks for acceptance and with "fair" ICC agreement;
- C3 with threshold for marks for acceptance and with "low" ICC agreement.

The results are shown respectively in Figs. 5, 6 and 7.

We note that for C1 and C2 (without threshold) when a reviewer gives a strong reject mark (i.e. from the strong reject band; dashed line in all figures) then the probability that

[7] Please note that in our reviews dataset the reviewers did not have access to other's reviewers marks, so they could not have been influenced by previous reviews.

**Table 4** Partition of marks into "bands" according to the conference rating scale (column 1) and approximate number of papers with at least one mark from a particular band (column 2)

| | 1 | | | 2 | | |
|---|---|---|---|---|---|---|
| | C1 | C2 | C3 | C1 | C2 | C3 |
| Band "strong reject" | 1–2 | 1 | 0–1 | 190 | 50 | 90 |
| Band "weak reject" | 3–4 | 2–3 | 1.5–2 | 550 | 200 | 280 |
| Band "borderline" | 5–6 | 4 | 2.5–3 | 600 | 80 | 460 |
| Band "weak accept" | 7–8 | 5–6 | 3.5–4 | 550 | 150 | 500 |
| Band "strong accept" | 9–10 | 7 | 4.5–5 | 130 | 20 | 220 |

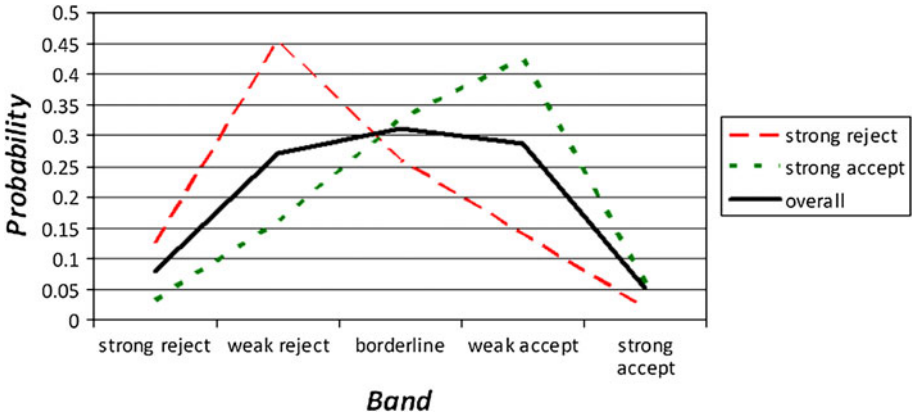


**Fig. 5** Band agreement for C1 ("high" ICC agreement)

other reviewers will give a mark from the weak or strong reject "band" is higher: these probabilities are significantly bigger than the overall probability that is shown in all figures with a black solid line. The same can be said about the "strong accept" band. So, in both cases we can say that reviewers seem to agree on very good and very bad papers.

For C3 (review process with a threshold mark for acceptance) the situation is different: overall probability is skewed in the direction of "weak accept" band. Here, we can suggest that when there is a mark threshold reviewers tend not to give very low marks since they know that even a mark from a "borderline" band and under the threshold will eventually "kill" a contribution (they tend to be polite!). A more detailed analysis shows that if somebody gives a mark from the "strong reject" band, this increases the probability of giving marks not only from strong and weak reject bands (by 14 and 63 % correspondingly) but also from borderline band (by 11 %). In the "strong accept" set the probability of others giving a "weak accept" mark is 20 % higher than the overall probability, but the probability of giving marks from other bands are almost the same as the overall probabilities. Therefore, we can say that we have marks skewed towards the "weak accept" and reviewers still agree on very bad contributions while disagree on very good.

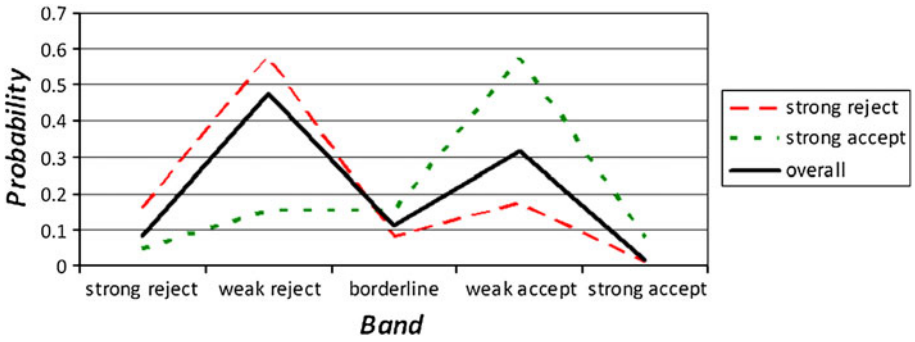


**Fig. 6** Band agreement for C2 ("fair" ICC agreement)

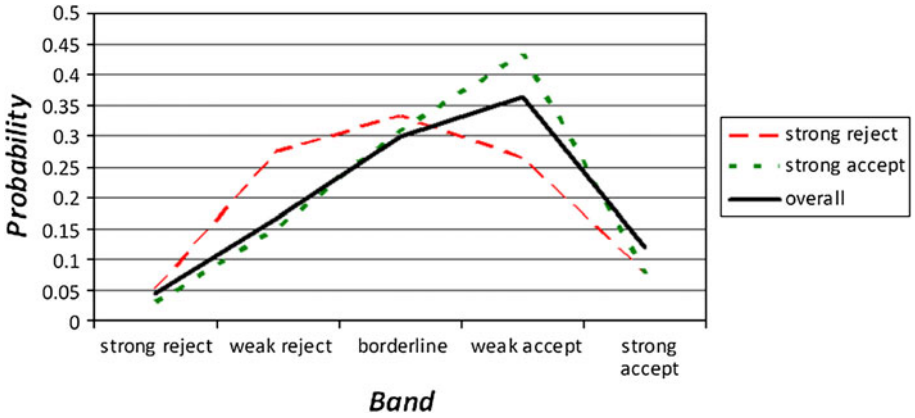


**Fig. 7** Band agreement for C3 ("low" ICC agreement)

Reliability: lessons learned

Through the measure of agreement/disagreement/band agreement among reviewers we derive the following findings:

- the measurement of the disagreement among reviewers is a useful metric to check and monitor the degree of process randomness. In particular, the disagreement is more useful than merely a statistical analysis of the degree of agreement since in this case we lack a clear reference process to compare with. Moreover, having a high disagreement value means, in some way, that the judgment of the involved peers is not sufficient to state the value of the contribution itself. So the monitoring of the disagreement metric could be useful to improve the quality of the review process as could help to decide, based on the disagreement value, if the used number of reviewers per contribution is enough to assess the contribution or if more reviewers are needed in order to ensure an higher quality process.
- from the *Band agreement* analysis we see that reviewers tend to agree on very good and very bad papers, except when the mark scale has a threshold. Moreover, band agreement results are consistent with the previous findings on the degree of agreement between reviewers.

### Quality: fairness

A review process is *fair* if and only if the contribution is judged solely on the basis of its scientific merit. Other data such as submission date (Cabanac and Preuss 2013), personal information, specific attributes of authors, such as their age, gender, nationality, academic post or number of previous publications and related impact should not influence the assessment.

There are numerous studies on the different kinds of biases of the peer review processes (see "Related work" section) Different sources of bias such as affiliation, topic, country, gender, clique bias have been also analyzed by using multiple logistic regression models in Hosmer and Lemeshow (2000). Unfortunately, in the collection of our datasets we were not provided with specific author information such as age, nationality, gender and other. In the framework of our present study, we therefore focused on the analysis of the *rating bias*, namely when reviewers are positively (negatively) biased i.e., they *consistently* give higher (lower) marks than their colleagues who are reviewing the same proposal. We briefly note that the procedure described below can be easily extended to other types of bias if the data (missing in our data set) would be available.

The way to compute the rating bias value is very similar to that described for the disagreement metric (see Eq. 3) i.e.:

$$\phi_i^j = M_i^j - \mu_i^j. \tag{5}$$

This time the sign of the equation is important in order to discover *positive* or *negative* biases. Indeed, if the value of $\phi_i^j$ is constantly positive, this means the reviewer tends to give always higher marks with respect to other reviewers *on the same set of contributions*; while if the value of $\phi_i^j$ is constantly negative then the reviewer tends to give always more negative marks than the other reviewers (always on the same set of contributions). Another type of rating bias is the *threshold bias*, which occurs when a reviewer gives marks that are always very close to the threshold for a given criteria (e.g. 3 in an evaluation scale form 1 to 5). This is computed by simply calculating the variance of the given mark for the specific criteria.

As for the disagreement metrics, there are several scopes to which we can apply the above bias metric. For example, we can measure the bias for a single reviewer and for a particular criterion, the bias over a review phase, and the bias over all the criteria.

Once biases are identified, a number of actions can be taken by the review chairs. One could be to compensate for the specific paper under review with additional reviews. Another action could be to apply automatic or semi-automatic unbiasing algorithms. A simple algorithm could be to modify the marks by adding or removing the bias values so that on average the overall bias of the most biased reviewers is reduced. In particular, if we take all reviewers $r$ that have a bias greater than $b$ and that have done a number of reviews higher than $n_r$, and subtract $b$ from all marks of $r$ (or from the top-$k$ biased reviewers), we can obtain a new debiased ranking. By comparing the obtained debiased ranking with the original ranking (for instance using the divergence metric—see "Divergence" section—that gives us the percentage of difference in rankings before and after unbiasing at acceptance threshold) we can assess the overall impact of the unbiasing procedure on the particular review process.

Applying the proposed rating bias metric, we were able to identify groups of potentially behavioral biased reviewers on actual review data in all 10 conferences in our dataset. These are all reviewers with an accepting or rejecting behavior with bias greater than $b$, a threshold value that depends on the rating scale granularity. So we could say that in our

dataset acceptance is a function of paper quality *but also* of chance of reviewer drawn. Table 5 reports for each analyzed conference:

1. the conference ID
2. the considered bias threshold $b$
3. the minimal number of reviews done by each reviewer (depending on the specific review process statistics)
4. the percentage of reviewers with accepting biased behavior
5. the percentage of reviewers with rejecting biased behavior
6. the divergence at acceptance threshold, which is used to measure the percentage of different papers between original and unbiased ranking.

The last column of Table 5 reports the percentage of papers affected by the proposed simple unbiased algorithm. The table shows that even with the simple metric we are proposing, it is relatively easy to detect rating biases. Moreover, following the simple unbiasing algorithm outlined above, it is also possible to measure quantitatively the effect of the bias on the review process. Depending on the specific conference, the accepting/rejecting bias impacts from 7 to 14 % of the overall contributions.

Fairness: lessons learned

From the above analysis we can derive some interesting points of interest and recommendations:

- the percentage of bias (e.g., accepting or rejecting behavior) is an important parameter to monitor by the review chairs and it is relatively easy to detect it through the use of simple metrics. If chairs will detect high number of biased reviewers they can decide to take some actions.
- it is also possible to devise simple and automatic unbiasing procedures; they do not need to be applied as black boxes, but together with the analysis of the divergence between the actual ranking and the unbiased one. Divergence provides quantitative data about the effect of unbiasing on the final review process: it indicates the percentage of papers in the accepted set whose fate is changed after applying the unbiasing

**Table 5** Percentage of reviewers with accepting/rejecting behavior (column 4–5) and percentage of affected papers for 10 different review processes (column 1) sorted by decreasing order of divergence values (column 6)

| Conference ID | $b$ | $n_r$ | Reviewers with accepting behavior (%) | Reviewers with rejecting behavior (%) | Divergence at acceptance threshold (%) |
|---|---|---|---|---|---|
| C5 | 1 | 3 | 4 | 3 | 14 |
| C9 | 0.5 | 3 | 11 | 7 | 14 |
| C2 | 1 | 3 | 7 | 7 | 12.5 |
| C8 | 1 | 3 | 17 | 16 | 11 |
| C1 | 1 | 3 | 5 | 4 | 10 |
| C3 | 0.5 | 3 | 8 | 5 | 9 |
| C4 | 0.5 | 2 | 3 | 3 | 9 |
| C7 | 1 | 4 | 6 | 3 | 8 |
| C6 | 0.5 | 3 | 1 | 2 | 7 |
| C10 | 1 | 3 | 5 | 13 | 7 |

procedure. This information can be used by the program chairs to better characterize and monitor the evaluation process: for example they can decide to unbias the score of particular reviewers.

Our future work in the dimension of fairness-related metrics includes studying of other types of biases related to affiliation, topic, country, gender, clique bias, different level of expertise and other aspects rather than limiting the analysis to accepting or rejecting biases. The challenge here is to collect and have access to the appropriate specific metadata.

### Quality: validity

*Validity* is related to the final result of the review process, i.e., the final ranking of the reviewed contributions. A review process is valid if it is able to select the best contributions. It may be claimed that this is the most important characteristic of the review process. However, not much research has been done on this topic, mainly because it is difficult to choose a measure for best-object detection. A well-known index such as citation count is a controversial measure of both quality and scientific impact of scholarly contributions (Bornmann et al. 2008b). Nevertheless, Lokker et al. (2008) succeeded in demonstrating for clinical papers that publications regarded—shortly after their appearance—as important by experts in the appropriate research field were cited much more frequently in subsequent years than publications that were less highly regarded. They used multiple regression model, checked the significance of 20 factors for 1,261 papers out of 105 most important clinical journals.

In our case, citation count was the main available measure, since other impact metrics (like the novel metric of number of downloads) are at present not available in a straightforward manner. By using citation count for the analysis of the conference peer review process, we tried to answer the question: how accurately did the selection process predict the longer-term impact of a contribution in the selected domain, i.e. CS?

Moreover, for one case, namely conference C3, we could perform a similar question but considering a two-phase review process. In fact, in C3 the reviewers were first asked to evaluate some extended abstracts; only a fixed number of proposed abstracts passes. Then, the selected authors provided the full contributions and this time the reviewers evaluated the full papers. So for this case we could explore—for the same accepted contributions—how accurately did the first review phase predicted the ranking outcome of the second review (full papers) phase.

In our analysis we applied both the *divergence* measure and the Kendall $\tau$-test—introduced in "Divergence" section—to compare the ranking output of the review process (using the obtained *final* marks of each contributions) and the ranking based on the a-posteriori estimated citation counts for each contribution. Therefore, we restricted the analysis to the set of accepted contributions $\mathcal{A}$ instead of the complete set of submitted contributions $\mathcal{C}$, as only for the accepted set we can have both reviewer's marks and citations. Moreover, we confined our analysis to the subset of relatively "old" conferences, namely before 2006, so we were able to compute citations received in the subsequent years using Google Scholar as the source for the estimate of the citation count.[8]

[8] Although Google Scholar has been criticized in the literature (e.g. Jacso 2010) mainly for the noise (spurious documents and citations) that it includes, it is however one of the few publicly available source of citations as well as with a high degree of coverage.

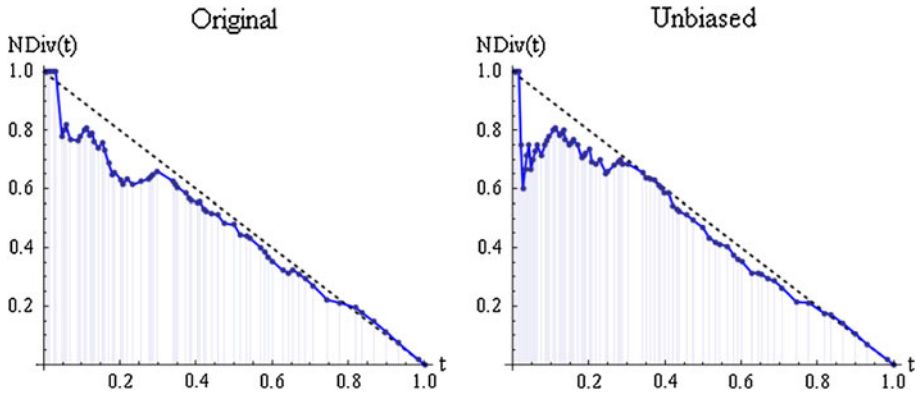


**Fig. 8** Normalized divergence between original and citations-based rankings for conference C1 for all accepted papers: **a** on the *left*, peer-review ranking versus citation-based ranking; **b** on the *right*, peer-review-unbiased ranking versus citation-based ranking

In Fig. 8a we report the divergence between the ranking of the conference C1 and the a-posteriori ranking based on citation counts. From Fig. 8a we can notice that the two rankings have a low correlation. Specifically, we can notice that the computed divergence curve is near to the diagonal, which—we recall—is the limiting case when two rankings are completely uncorrelated. We found similar divergence curves for all other "old" conferences in our dataset.

While in Fig. 8a we report the divergence value for the whole set of accepted papers, in Fig. 9a we report the divergence between the two rankings for only the first 50 % of accepted papers in C1. Also for these "top" papers, we can see again that the correlation remains low.

The results of the Kendall $\tau$ test—comparing citation and original peer review rankings for all "old" conferences[9]—are collected in Table 6. Furthermore, we present a Kendall $\tau$ test analysis applied to conference C3, where in place of citations we have used the second-phase review ranking of the accepted extended abstract to compare the rankings for the same contributions in the two phases.

Also for the $\tau$ test, we carried out the analysis for different sets of accepted papers for large conferences, such as C1 and C3. Namely, first for the whole set, and then for reduced sets of papers: 50, 33 and 10 % of the top accepted papers. We recall here that a value of Kendall $\tau = 0$ corresponds to independent rankings, $\tau = 1$ to correlated rankings and $\tau = -1$ to inversely correlated rankings.

As we can see in Table 6 for only two out of five conferences (C8, C10) there is some correlation between original and citation-based ranking, while for the other three investigated conferences the correlation is close to zero even for the "top" (50, 33 and 10 %) accepted papers. It is also interesting to note that the correlation between the first and second phase review (conference C3) is also close to zero in all cases. All results presented in Table 6 are statistically significant within a 95 % CI against the null hypothesis ($H_0 : \tau = 0$).

We have also conducted Kendall $\tau$ test for citation-based and unbiased ranking (i.e, ranking obtained applying the unbiasing procedure described in "Quality: fairness" section) In some cases (C9) and in some reduced sets of "top" accepted papers (C1 and

[9] Old conferences are the ones which took place in the period from 2003 to 2006, therefore "old" enough for checking the number of citations received during the subsequent years.

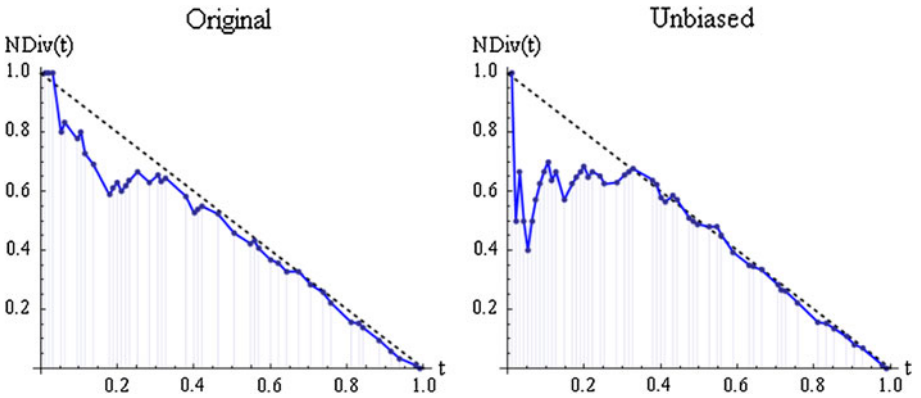


**Fig. 9** Normalized divergence between original and citations based rankings for conference C1 for the top 50 % accepted papers: **a** on the *left*, peer-review ranking versus citation-based ranking; **b** on the *right*, peer-review-unbiased ranking versus citation-based ranking

**Table 6** Results of Kendall $\tau$-test for five conferences

| Conf. ID | Top paper (%) | Citations-based versus original ranking Kendall $\tau$ | Citations-based versus unbiased ranking Kendall $\tau$ |
|---|---|---|---|
| C1 | 100 | 0.078 | 0.074 |
| | 50 | 0.097 | 0.066 |
| | 33 | 0.127 | 0.134 |
| | 10 | 0.067 | 0.152 |
| C8 | 100 | 0.392 | 0.346 |
| C9 | 100 | −0.026 | 0.178 |
| C10 | 100 | 0.310 | 0.269 |
| | | Second review versus first review | Second review versus unbiased first review |
| C3 | 100 | 0.054 | 0.078 |
| | 50 | −0.057 | −0.064 |
| | 33 | 0.053 | 0.034 |
| | 10 | 0.087 | 0.134 |

Different set of papers were used for the analysis of large conferences C1 and C3: 100, 50, 33 and 10 % of top accepted papers

C3) the correlation slightly improves. This is also visually confirmed in the divergence curves computed for C1 and presented in Fig. 8b for all accepted papers and in Fig. 9b for the top 50 % accepted papers. For the two conferences with better correlation before unbiasing (C10, C8), Kendall $\tau$ coefficient became *lower* after unbiasing but remained statistically significant. From these preliminary results, we cannot say whether or not the unbiasing procedure improves predictive validity of peer review process.

While examining the above analyses, one could argue that the aim of peer review process *is not* the selection of high-impact papers, but is simply to filter *junk* papers and

accept only the ones above a certain quality threshold.[10] However in our view, it is important that the program chairs of a conference or a journal should decide their target parameter. The above analyses provide the procedures to check a-posteriori the validity of the review process w.r.t. a selected target measurable parameter.

To ensure the validity of the peer review process the chairs may also decide to control the accuracy of the papers' ideal mark estimation. This approach is described in the following subsection.

### Evaluation of the accuracy of a review

We focus now on the measure of the accuracy of papers marks obtained from the reviewers. By definition the "accuracy" of a measurement system is the degree of closeness of measurements of a quantity to its actual (true) value (see for instance the True Score Theory about measurement). Our working hypothesis is that the "true" mark is the one we would get in the ideal case we would be able to collect reviews (and related marks) from all the *experts* in the community (see our definition of the peer review model in "Peer review model" section).

We follow a standard statistical approach based on the assumption that for a large number of reviewers the mark ($x$) of a given paper is a random variable with Gaussian distribution $N(\mu, \sigma)$. The sample mean $\hat{\mu}(n) = \frac{1}{n}\sum_{i=1}^{n} x_i$ for each contribution is the estimation of the mathematical expectation value $\mu$ of the mark of the paper, and it converges to this value when the number of reviewers ($n$) tend to infinity. $\sigma^2$ represents the variance of the marks Gaussian distribution. We consider $\mu$ as the "true" mark for the contribution, i.e. the value that we want to estimate using a specific peer review process.

The goal of a real peer review process is to choose the number of reviewers $n$ so that the error of estimation would be less than $\varepsilon$ with probability $(1 - \alpha)$:

$$P\{|\mu - \hat{\mu}(n)| < \varepsilon\} = 1 - \alpha. \tag{6}$$

i.e. $\mu$ falls into confidence interval $(\hat{\mu}(n) - \varepsilon, \hat{\mu}(n) + \varepsilon)$ with confidence level $(1 - \alpha)$. If $\sigma$ is known, then confidence interval for unknown mathematical expectation $\mu$ with confidence level $(1 - \alpha)$ can be computed analytically as:

$$\hat{\mu}(n) - u_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} < \mu < \hat{\mu}(n) + u_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \tag{7}$$

where $u_{\frac{\alpha}{2}}$ is the quantile of the standardized normal distribution defined by the confidence probability $(1 - \alpha)$, and $\varepsilon = u_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$ is the accuracy (limiting error) point estimate of the mathematical expectation value $\mu$. An analysis of Eq. 7 shows that:

1. larger $n$ correlates with smaller confidence intervals, hence—as one would expect—the estimation is more accurate the higher the number of reviews;
2. increasing the probability confidence $(1 - \alpha)$ increases also the confidence interval length;
3. if we fix the accuracy $\varepsilon$ and the confidence probability $(1 - \alpha)$ then from the formula $\varepsilon = u_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$ we can obtain the required (optimal) amount of sampling (i.e. $n_{\min}$), that will provide the desired accuracy.

[10] This is the rationale behind some journals like PLoS ONE among others.

Unfortunately, in real cases $\sigma$ is not known and cannot be estimated a priori. However, we can use known point estimate of the true variance $\sigma$ using as an approximation the population standard deviation $s_n$ obtained either a posteriori (e.g. by crunching final data from the current or even past editions of a given conference) or dynamically using current marks for a single contribution. This approximation will not lead to analytically correct results ($\sigma$ is supposed to be known in the above method), but it allows to get an approximated estimate of the accuracy behavior depending on $n$.

We carried out a number of analyses with actual data from the largest conference C1 in our dataset. Figure 10 shows the results obtained using the computed (a posteriori) average value for the overall marks sample standard deviation $s_n = 1.51$ (absolute value, i.e. not normalized).

In this specific case, in order to have an accuracy around $\pm 1$ absolute marks with confidence level of 0.9 around the "true"' mark we would need around six RPP. However, the figure clearly shows that improving the accuracy is going to be hard since the accuracy curves level off (decrease very slowly) as the number of reviewers increases.

Another useful approach is to acknowledge that $\sigma$ is unknown, and use statistical approaches for obtaining the confidence interval of an unknown mathematical expectation $\mu$ from a random variable $X$ with unknown normal distribution $N(\mu, \sigma)$. Specifically we can write (following Brink 2008):

$$\hat{\mu}(n) - t_{\frac{\alpha}{2};n-1} \cdot \frac{s_n}{\sqrt{n}} < \mu < \hat{\mu}(n) + t_{\frac{\alpha}{2};n-1} \cdot \frac{s_n}{\sqrt{n}}, \tag{8}$$

where $t_{\frac{\alpha}{2};n-1}$ is the quantile of the Student's distribution defined by the confidence probability $(1 - \alpha)$ and by the number of degrees of freedom $n - 1$; $\hat{\mu}(n)$ and $s_n$ are the unbiased point estimates of the normal distribution parameters; $\varepsilon = t_{\frac{\alpha}{2};n-1} \frac{s_n}{\sqrt{n}}$ is the accuracy (limiting error) point estimate of the mathematical expectation value $\mu$.

Given a specific sample of actual marks, Eq. 8 can be used to compute the confidence interval for $\mu$ while it cannot be used to find directly the required amount of sampling. However, we suggest that it can be used to estimate (either in real-time or a posteriori)

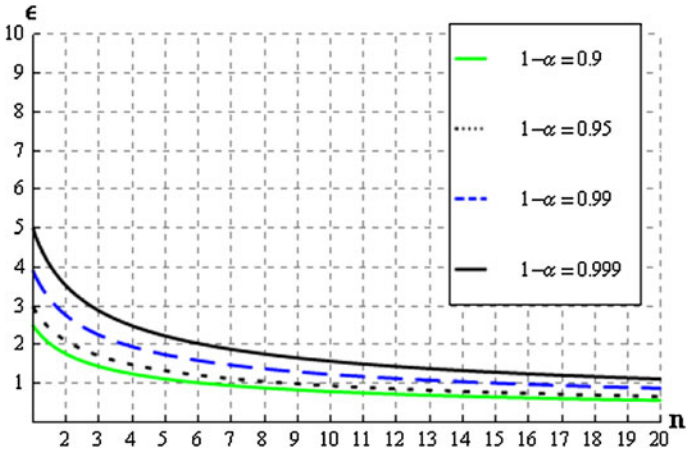


**Fig. 10** Accuracy versus amount of sampling (i.e. $n$ number of reviewers) for the most common values for the confidence probability. i.e. $(1 - \alpha) = 0.9, 0.95, 0.99, 0.999$ and using the computed approximation for the population standard deviation $s_n = 1.51$ in C1

whether the number of reviewers for a given paper is/was enough to determine $\mu$ with a defined accuracy, or if more reviews are/were needed. We note here, that similar but informal procedures are currently used in many review processes: for instance in the case when there is a relevant disagreement in the opinions among experts for a specific contribution, the review chairs can decide to include other reviewers in the evaluation. Our statistical approach provides a sound mathematical base for such procedures and adds an additional quantitative dimension—with a detailed estimate of accuracy of the process for a given confidence level—that could be implemented directly in state-of-the-art electronic editorial systems (Fig. 11).

As an example, in Fig. 12 we show how the suggested statistical approach could be used to estimate the accuracy "on-the-fly" during a review process for a particular contribution and adding more reviewers as a function of the desired target confidence level. The data for the specific example are based on a contribution from conference C1 with six reviews and corresponding marks [namely equal to (5, 8, 7, 5, 4, 4)] for the criterion used for the final ranking. In the analysis we sorted marks by review date and computed the accuracy of the estimation (depending on the confidence probability) for first $k$ reviews for $k$ in the range (3–6), as if we would dynamically add new reviewers (Fig. 12).

The accuracy curves show the increase of accuracy in the process as a function of the confidence level ($x$-axis) and of the number of reviews added (individual curves). For instance, for a confidence level of 0.90 the accuracy in the estimate of the mark values improves from ca. $\pm 2.5$ absolute marks with three reviews to $\pm 1.2$ when all six reviews are considered.

## Confidence level depending on the number of reviewers per paper

We also investigated, how the probability of having an average mark—given by a particular number of reviewers in a confidence interval with predefined accuracy $\pm\varepsilon$—changes depending on the number of reviewers. For this purposes we used the same model and the same assumptions that were described at the beginning of the previous section "Evaluation of the accuracy of a review". We use again standard statistical approaches for obtaining the confidence interval of an unknown mathematical expectation $\mu$ when $\sigma$ is known ($\mu$ and $\sigma$

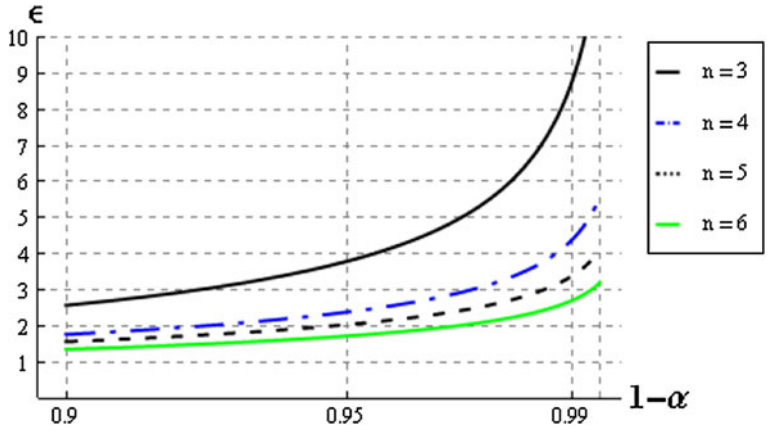


**Fig. 11** Accuracy of estimation versus confidence probability depending on the considered number of marks $n$

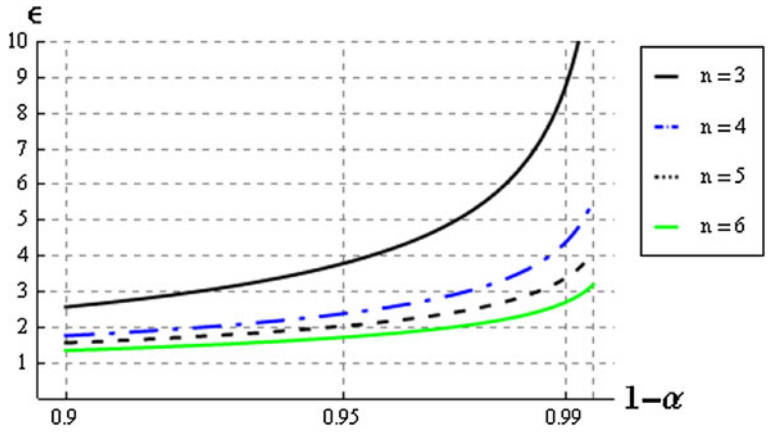


**Fig. 12** Accuracy of estimation versus confidence probability depending on the considered number of marks $n$

**Table 7** Standard deviation for all the conferences

| Conf. ID | $s_n$ |
|---|---|
| C1 | 0.16 |
| C2 | 0.19 |
| C3 | 0.14 |
| C4 | 0.12 |
| C5 | 0.17 |
| C6 | 0.11 |
| C7 | 0.16 |
| C8 | 0.20 |
| C9 | 0.19 |
| C10 | 0.17 |
| Average SD | 0.16 |

are the parameters of mark distribution $N(\mu, \sigma)$). In particular, from Eqs. 6 and 7 we can obtain:

$$1 - \alpha = 1 - 2 * F\left(\frac{\varepsilon\sqrt{n}}{\sigma}\right) \tag{9}$$

where the function $F()$ represents the cumulative density function of standardized normal distribution.

As we wanted to obtain a result for all conferences and not for a single conference or contribution, we calculated the *normalized*[11] average standard deviation (average within the papers marks) for each available conference (see Table 7) and than used their values to compute an estimate of the unknown $\sigma$.

Figure 13 shows the results from Eq. 9, obtained using three computed (a posteriori) average values of the sample standard deviation: minimum, maximum and average value. From these results we can conclude that 3 reviewers per paper—i.e. the number generally

[11] The marks before the computation were normalized to the scale [0,1].

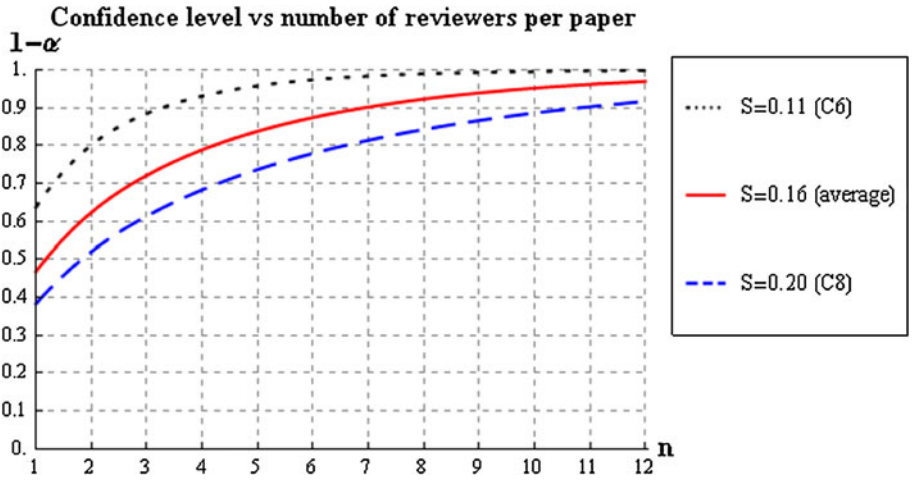


**Fig. 13** Confidence level versus number of RPP with error of the mark $\varepsilon = 0.1$

used in peer review in conferences—give highly confident results ($(1 - \alpha) > 0.9$) only for conferences with high agreement among reviewers (little standard deviation of the marks). In all other cases the accuracy of paper's marks estimation should be kept under control: for the conferences with low agreement, using only three reviewers produces results with a confidence level around 0.6, i.e. 40 % of the times the mark estimation can be wrong. The method that we described in the previous section "Evaluation of the accuracy of a review" could have been used here to improve the accuracy during the review process.

Validity: lessons learned

From the exploration of the validity dimension we can derive the following findings and recommendations:

- The divergence metric is a practical metric to compare the actual ranking of the conference against various target rankings.
- The application of the divergence metric has uncovered (for the available data sets) that there is low correlation between the ranking of contributions obtained in the analyzed review process and the actual impact (citation counts) of the same contributions in the community. This result is confirmed by the Kendall $\tau$-tests.
- We do not have enough evidence based on the available datasets that the unbiasing procedure proposed in "Quality: fairness" section improves the validity of peer review process when citations are used as target measurable parameter.
- The statistical procedures proposed in "Evaluation of the accuracy of a review" section can be used by program chairs to control the accuracy of review both on-the-fly or a posteriori and can be easily implemented in current editorial management systems.

## Analysis of the efficiency of the peer review process

In our view, the *efficiency* of a peer review process is linked to the effort spent in determining which contributions are accepted, and in particular to the trade-off between effort

and quality of the review process. It considers both the effort in writing contributions and in reviewing them. We add also the authors' effort because it can be affected by multiple phase reviewing process when, for example, in the first phase only short contribution with main ideas need to be prepared (e.g., extended abstract).

The basic working assumption of this section is that the quality-effort trade-off exists and that, in general, if a paper or proposal is long, and is reviewed very carefully by a large number of reviewers (all the reviewers and the chairs are considered to be experts), the selection is more informed than the case in which, say, one page proposal is briefly looked at by a couple of reviewers. Time is a precious resource, so the challenge is how to reduce the time spent while maintaining a "good" selection process that indeed selects the "best" contributions. A separate issue that we do not address (also as it is hard to measure) is the fact that a process is affected by the quality of the reviewers and the amount of discussion or the presence of a face to face discussion. For now we limit just to metrics that we can derive from raw review data (essentially marks data).

In the following we identify metrics that can help us understand if the review process is efficient. The *reviewing effort* of a review phase is the total number of reviews $N_R$ multiplied by the average time $\overline{t_r}$ (e.g., measured in person-hours) spent per review in that phase. Correspondingly, the *contribution preparation effort* is the number of submissions $N_C$ multiplied by the average time spent in preparing each submission $\overline{t_w}$. Reviews and submissions can span across $N_P$ phases. For simplicity, in the above definitions and in this section we use the average reviewing or writing time instead of considering the time spent by each reviewer or author and the fact that different phases may require different reviewing or writing efforts per contribution. We also assume that the set of experts is the same for all phases. The extension of the reasoning done here to remove these assumptions is straightforward.

In the ideal case—from a quality perspective—all reviewers are equally experts and read all contributions for as long as they need to take a decision; and contributions are as long as they need to be for the reviewers to fully grasp their value. With respect to the review time and contribution length, we assume in particular that as the review time and contribution length grow, the reviewer is able to narrow down the *uncertainty/error* on the review marks he or she wants to give. In other words, it will increase the confidence that the correct mark for the contribution is within a given interval.

Our hypothesis here is that beyond a certain review time threshold $t_{rx}$ and contribution length threshold $l_x$ the mark uncertainty remains constant. Reading a 10 pages paper for 4 h or 4 days will not likely make a difference (if we are in doubt between giving a six and a seven we will probably still be in doubt), but one minute versus four hours will.[12]

Essentially, in all real cases (conference, journal or project's proposal evaluation) the actual review process is far away from the above ideal case. It is therefore of interest to have some analysis and quantitative data and metrics to measure how far we are from the ideal case.

Informally, making the review process efficient requires reducing the effort and, at the same time, minimizing the quality degradation. In the following subsections we analyzed the process of stopping the reviews when the fate of a paper is clear and proposed an heuristic procedure to choose the "optimal" number of papers per reviewer.

[12] We recall again that in our work we focus only on the quantitative aspect of peer review (i.e. marks) and not on the other important dimension of providing constructive feedbacks to authors.

### Optimizing the number of reviews

A first line of investigation is around optimizing the number of reviews for submissions whose fate is clear. Assume that the review process is structured in as many phases as the maximum number of reviewers per paper (say, we plan to have at most five reviews for a paper, so at most five phases). In other words we are investigating the consequences of a sequential assignment of the reviewers. The analysis we want to make is to understand which is the earliest phase at which we can stop reviewing a given paper, because we have a sufficiently good approximation of the fate of the paper, which is the one we would get with all reviews (five in our running example). In particular, given the number $T$ of submissions we can accept (as long as they get marks above a minimal acceptance threshold), we want to estimate the earliest point (i.e. the minimum number of reviews) so that we can state whether a paper will or will not be in the top $T$. As an example, if a paper has two strong reject reviews and it is impossible for it to end up in the acceptance range, we can stop the review process for this paper just after two reviews. Stopping reviews for guaranteed acceptance is more complex as it also depends on the marks of other papers (being above a threshold is not enough as it is a competitive process). However, it is always possible to verify if there is a possible combination of marks for the missing reviews that can change the ranking to the point that the paper can end up below the acceptance threshold.

In Fig. 14 we show the results of such deterministic approach for a simulated case where the number of reviewers is $|R| = 5$, for each criteria $M_j$ the reviewer can assign a mark between $\{1, \ldots, 10\}$ with no half-marks and with a fixed acceptance threshold $T_a = 7.0$. The dark areas at the bottom of the diagram indicate the cases where the fate (rejection) of the contribution is already finalized and no further review will change it. The shaded areas at the upper part indicate the symmetric cases where the acceptance of the contribution is sure (in this case this is based on the existence of a minimal acceptance threshold).

In addition to the deterministic analysis mentioned above, which is conservative, we can also perform a statistical analysis relying on the fact that reviewers' marks exhibit some correlation (see our analysis in "Quality: reliability" section). In general, after each phase, we can estimate the probability of each paper ending up in the accept or reject bin, and to do so we can also leverage our previous band disagreement measures (see "Quality: reliability" section) to help estimate the confidence associated to the estimate. The results of such approach are also depicted in Fig. 14 as dashed areas with the corresponding probability estimate in the figure caption.

Notice that implementing the above process requires either a multi-phase review or to give to reviewers a priority on what they should review. This in order to increase the chances that the reviews that would have to be reviewed later may not be needed because the fate of the contribution has already been determined. A more formal analysis of such process is part of our current research work.

### Effort-invariant approaches

A second approach to the efficiency dimension, is around effort-invariant choices, that is, varying review process parameters to improve quality while keeping the effort constant. Here we investigate the efficiency of the review process from the view of an efficient ("optimal" number of PPR) review distribution among reviewers.

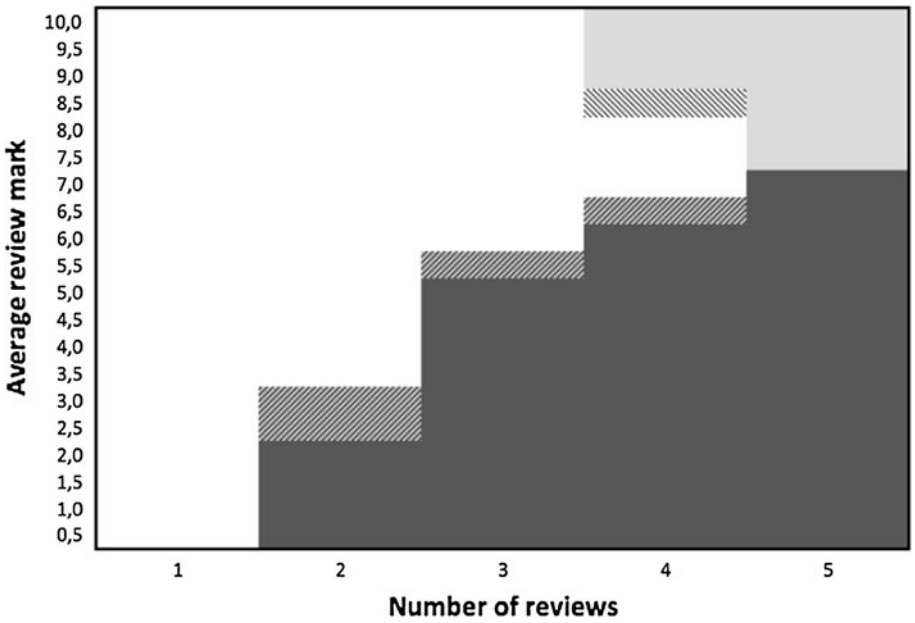


**Fig. 14** Deterministic and statistical acceptance and rejection analysis for a contribution. *Dark areas at the bottom*: 100 % rejected. *Shaded areas at the top* 100 % accepted. *Dashed areas at the top/bottom*: 95 % probability accepted/rejected

Our working hypothesis is that in all evaluation processes there are different groups of contributions to evaluate, typically: immature, average, good and eventually excellent papers. We presume that if a reviewer evaluates contributions only from one group, their evaluation scale will tend to expand, i.e. contributions from the same group could end up with very diverse marks. If a reviewer would have access to contributions belonging to different groups, the scale could be more realistic and probably more correct. Consequently, we would like to estimate how to distribute the papers among reviewers in a way such that each reviewer will have at least one paper from each group. The idea is to use statistical information about the distribution of the average marks for individual contributions (either an expected one or an historical one where available) in order to identify typical clusters of contributions for a given review process. Then, to use statistical approaches to compute the needed number of PPR in order to maximize the probability—with a specified confidence level—to have in the set of reviews at least one paper from each cluster.

In order to show a possible implementation of this idea, we first studied a posteriori the distribution of the average marks for individual papers and for one criterion (for example for the most significant one among the marks of the conference). Figure 15 shows the average marks distribution for one of the analyzed conference, namely C3. This information is used to evaluate the general behavior of the sample as we use it as an estimation of the mean values density function.

On the basis of this type of distribution, we then determine appropriate boundaries for papers clusters. As initial parameters in our statistical approach we have:

1. estimate of average mark distribution;

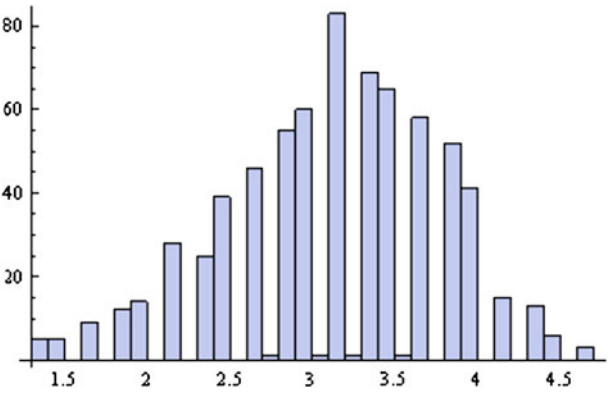


**Fig. 15** Distribution of average marks for individual papers for C3. On the *x*-axis we plot the average marks and on the *y*-axis the measured density in percentage

2. user's selection for cluster boundaries;
3. user's selection of desired confidence level $(1 - \alpha)$.

In the following analysis for conference C3, we choose three clusters—*immature*, *average* and *good/excellent* papers—with the following range [0, 2.7]; [2.7, 3.7]; [3.7, 5] correspondingly. The confidence level represents the probability that at least one paper from the group with minimal probability ($p_{\min}$) will be assigned to a reviewer (i.e. $\alpha$ is the probability that in the set of papers for each reviewer there will not be the paper from the minimal probability group). Then if we enforce that a reviewer reviews with the probability $(1 - \alpha)$ at least one paper from this group, the papers from the other groups will appear with higher probabilities.

If $n$ is the desired value for the number of PPR, then—assuming that we have a large number of observations—we can estimate $n$ as: $\alpha = (1 - p_{\min})^n$ hence

$$n = \log_{1-p_{\min}}(\alpha). \tag{10}$$

If the number of observations ($N$) is not very large (i.e. the group probability changes significantly if we pull one paper out) then we can approximate the solution with the expansion:

$$\alpha = (1 - p_{\min})\left(1 - p_{\min}\frac{N}{N-1}\right)\cdots\left(1 - p_{\min}\frac{N}{N-n+1}\right) \tag{11}$$

In this case, we cannot obtain an analytical expression for $n$, but we can estimate it using the following computation procedure:

1. Set initial parameters: average mark distribution, cluster boundaries, confidence level $(1 - \alpha)$.
2. Calculate the cluster distribution $\{p_1, p_2, \ldots, p_k\}$, where $k$ is the number of paper clusters, $p_i = \frac{N_i}{N}, \quad i = 1, \ldots, k, N$—total number of papers, $N_i$—number of papers in the $i$th group.
3. Find minimal $p_i, i = 1, \ldots, k$. Define it as $p_{\min}$.
4. Obtain $n$ from Eq. 11.

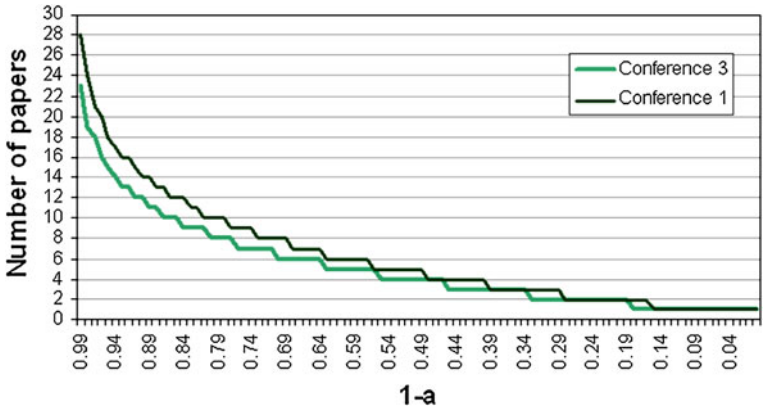


**Fig. 16** Number of PPR for different values of $1 - \alpha$ ($0.99 \leq 1 - \alpha \leq 0.01$)

This approach can be used to estimate the quality of the peer review process dynamically (collecting and analyzing marks distribution from reviews as they are coming in during the evaluation process) or a posteriori (to check within which confidence level the initial assumption—each reviewer have had at least one paper from each cluster—has been met).

Results from an a posteriori analysis are reported in Fig. 16 with real data from two conferences: C1 and C3. Review chairs could have seen from the graphs that reviewers with a small number of papers have had a small probability of reviewing the papers from all the groups. In particular for these conferences, if the reviewers have received on average only four papers to review the probability of reviewing a paper from the "immature" cluster[13] ranges from 45 to 51 % for conference C3 and from 47 to 38 % for conference C1. In order to have a confidence level around 80–90 % that each reviewer has seen a contribution from every cluster, each reviewer should have been assigned around 9–12 contributions for conference C3 and 10–14 for conference C1.

Efficiency: lessons learned

Our preliminary investigations along the efficiency dimension led us to the following results:

- definition of a framework and a number of metrics for investigating how to analyze the efficiency of peer review processes.
- definition of both a deterministic and statistical procedures to support the chairs of a review process to optimize (reduce) the number of reviews for the papers with clear fate.
- definition and application of an heuristic procedure for calculating "optimal" number of PPR in order to ensure that each reviewer has a good chance (within a defined confidence interval set by the chairs) to access contributions of all qualities (immature, average, good/excellent) in order to be more consistent in their evaluation scale.

It would be interesting to apply our proposed efficiency metrics also to analyze and optimize the effort spent by both authors and reviewers during their work, but unfortunately the data about time spent in writing or reviewing are not easily available.

[13] Both in C1 and C3 the cluster with minimal probability was the "immature" cluster.

## Conclusion

In this paper we have presented and discussed the results of the analysis of peer review data from 10 conferences whose topics were related to the CS field for a total of ca. 9,000 reviews on ca. 2,800 contributions. We have conducted the analysis along four different dimensions: *reliability*, *fairness*, *validity* and *efficiency*. Together with the traditional metrics and analysis found in literature, we have also performed additional analysis studying the infuence of the mark scale on the rating process, the robustness of the peer review process and introduced different measures to compare rankings, disagreement/agreement among reviewers, rating bias, and accuracy of the papers marks obtained from reviewers.

In regard to reliability of peer review processes, we found evidence in our data set that there is an overall agreement among the reviewers according to Intraclass Correlation Coefficient analyses. However, disagreement among reviewers exists as well and it is relatively high, although different from random processes. Moreover, according to our proposed band agreement analysis, we found quantitative evidence that reviewers tend to agree more on very bad or very good papers. Thus, we can claim that the analyzed peer review processes can be considered reliable mainly for very bad or very good papers since the analyzed processes tend to produce there much higher agreement than random or semi-random processes.

In regard to fairness of peer review processes, we have analyzed a specific source of bias, namely *rating* bias, to find out if there are reviewers that constantly give higher (positive bias) or lower (negative bias) marks than their colleagues while reviewing the same proposal. We found out that in every conference in our dataset, it is possible to identify a set of reviewers with a positive/negative bias, that is reviewers with, respectively, accepting/rejecting bias and that this behavior impacts from 7 to 14 % of the overall submitted contributions in our dataset. However, once the bias has been detected, program chairs may take some actions to compensate it, as giving the paper to additional reviewers, or adding/removing the bias values to obtain a new unbiased final ranking for the contributions. Therefore, though in our data set rating bias is always present in the marks given by reviews, there are ways to identify it and, luckily, there are also ways to deal with it to compensate its impact.

From the analyses of the validity of peer review processes, we have found no evidence of correlation between the rankings outcome of the investigated review processes and the impact of the selected contributions measured by citations; the low correlation is also confirmed in a similar study of a posteriori review of the same contributions at a later time. Although it might be that the selected target parameter (i.e., citations) or the citation source (i.e., Google Scholar) for evaluating the validity of the review process could not be the ideal ones, our proposed analysis provides a straightforward procedure to check a-posteriori a review process validity with respect to any specific (and measurable) target parameter selected by the review process chairs. Moreover, chairs can also decide to monitor the accuracy of the mark of a paper in terms of the size of estimation error within a probabilistic confidence level. Our analysis shows that the standard number of reviewers per paper (typically 3) is often not enough to reach a satisfactory accuracy (see "Evaluation of the accuracy of a review", "Confidence level depending on the number of reviewers per paper" sections). To achieve a small error of estimation or, in other words, more accurate results with high confidence level, a dynamic control over mark estimation approach could be used. A possible approach could be to add reviewers until a predefined accuracy level is achieved.

Finally, we presented some investigations on the efficiency of peer review processes and reported some preliminary results related to the possibility to devise statistical approaches to tune review process parameters to improve quality while keeping the overall effort under control.

We had two main goals in our work: (i) search for scientific evidence that peer review works (or that it doesn't), and (ii) search for ways to improve the peer review process so that it can work better. With respect to the first goal, the analyzed datasets did not provide us with a definite answer. In the analyzed dataset (10 conference in CS) we have found that: there is a significant degree of randomness in the analyzed review processes, more marked than we expected; the disagreement among reviewers is high and there is a low correlation between the rankings of the review process and the impact of the selected papers as measured by the most used indicator of impact, i.e. citations. This is also true in the similar study of a posteriori review of the same contributions at a later time. If these trends would be confirmed for more and diverse (i.e. from different domains) review processes then we could affirm that current peer review processes do not work very effectively.

On the second goal we can be more specific: the proposed analysis model and framework can be used as the basis to develop a support system in state-of-the-art editorial management systems to support review process chairs both during the review and as a posteriori check on the overall quality of the process. Using the various methods proposed in this paper (e.g. robustness analysis, disagreement analysis, band agreement analysis, bias analysis, un-biasing procedures, a-posteriori validity analysis with respect to specific target parameter(s), a-posteriori or on-the-fly marks accuracy evaluation, as well as statistical approaches to tune review process parameters) the chairs of a peer review process could arrive at a deeper understanding of their specific selection process and pursue a number of appropriate ways to improve it.

We do not claim that our results are general and final, but we think that they indicate an applicable quantitative methodology to tackle the analysis of peer review and provide important suggestions to improve current peer review process.

In the future we want to extend the analysis to more conferences and journals peer review processes also from fields different from CS and analyze from a more theoretical approach ways to improve the efficiency of current peer review processes. Finally, we will continue to aim at providing all stakeholders in peer review processes with an intuitive understanding of what the various metrics imply, in an effort to explain the numbers, so that all involved stakeholders will more easily assess "how well" peer review works.

**Acknowledgements** This paper is an extended version of the 12 pages paper titled "A Quantitative Analysis of Peer Review" presented at the 13th Conference of the International Society for Scientometrics and Informetrics, Durban (South Africa), 4–7 July 2011 (Ragone et al. 2011). We acknowledge the financial support of the Future and Emerging Technologies (FET) programme within the Seventh Framework Programme for Research of the European Commission for the LIQUIDPUB project under FET-Open grant number: 213360. We also want to acknowledge the anonymous reviewers of our manuscript. Their comments have really helped us to improve our work, underlying something that we knew already (and mention in our work): peer review is not only focused on filtering and selecting manuscripts to publish but also to provide constructive feedbacks to authors.

## References

Akst, J. (2010). I hate your paper. *The Scientist, 24*(8), 36–41.

Barnes, J. (1981). Proof and the syllogism. In E. Berti (Ed.), *Aristotle on science: The posterior analytics* (pp. 17–59). Padua: Antenore.

Bartko, J. J. (1966). The intraclass correlation coefficient as a measure of reliability. *Psychological Reports, 19*, 2–11.
Bartko, J. J. (1974). Corrective note to "the intraclass correlation coefficient as a measure of reliability". *Psychological Reports, 34*, 418.
Benos, D. J., Bashari, E., Chaves, J. M., Gaggar, A., et al. (2007). The ups and downs of peer review. *Advances in Physiology Education, 31*(2), 145–152.
Birman, K., & Schneider, F. (2009). Program committee overload in systems. *Communications of the ACM, 52*(5), 34–37.
Bollen, J., Van de Sompel, H., Smith, J., & Luce, R. (2005). Toward alternative metrics of journal impact: A comparison of download and citation data. *Information Processing & Management, 41*(6), 1419–1440.
Bornmann, L. (2007). Bias cut: Women, it seems, often get a raw deal in science—So how can discrimination be tackled?. *Nature, 445*, 566.
Bornmann, L., & Daniel, H. D. (2005a). Committee peer review at an international research foundation: Predictive validity and fairness of selection decisions on post-graduate fellowship applications. *Research Evaluation, 14*(1), 15–20.
Bornmann, L., & Daniel, H. D. (2005b). Selection of research fellowship recipients by committee peer review. Reliability, fairness and predictive validity of board of trustees' decisions. *Scientometrics, 63*(2), 297–320.
Bornmann, L., & Daniel, H. D. (2010a). Reliability of reviewers' ratings when using public peer review: A case study. *Learned Publishing, 23*(2), 124–131.
Bornmann, L., & Daniel, H. D. (2010b). The validity of staff editors initial evaluations of manuscripts: A case study of angewandte chemie international edition. *Scientometrics, 85*(3), 681–687.
Bornmann, L., Mutz, R., & Daniel, H. D. D. (2008a). How to detect indications of potential sources of bias in peer review: A generalized latent variable modeling approach exemplified by a gender study. *Journal of Informetrics, 2*(4), 280–287.
Bornmann, L., Wallon, G., & Ledin, A. (2008b). Does the committee peer review select the best applicants for funding? An investigation of the selection process for two European Molecular Biology Organization Programmes. *PLoS ONE, 3*. doi:10.1371/journal.pone.0003480.
Bornmann, L., Wolf, M., & Daniel, H. D. (2012). Closed versus open reviewing of journal manuscripts: How far do comments differ in language use? *Scientometrics, 91*(3), 843–856. doi.10.1007/s11192-011-0569-5. http://www.akademiai.com/content/0436287611KJ2063.
Brink, D. (2008). *Statistics*. Fredriksberg: Ventus Publishing ApS.
Cabanac, G., & Preuss, T. (2013). Capitalizing on order effects in the bids of peer-reviewed conferences to secure reviews by expert referees. *Journal of the American Society for Information Science and Technology*. doi:10.1002/asi.22747.
Ceci, S., & Williams, W. (2011). Understanding current causes of women's underrepresentation in science. *Proceedings of the National Academy of Sciences, 108*(8), 3157–3162.
Ceci, S. J., & Peters, D. P. (1982). Peer review: A study of reliability. *Climate Change, 14*(6), 44–48.
Chen, J., & Konstan, J. A. (2010). Conference paper selectivity and impact. *Communications of the ACM, 53*(6), 79–83. doi:10.1145/1743546.1743569.
Cicchetti, D., & Sparrow, S. (1981). Developing criteria for establishing interrater reliability of specific items: Applications to assessment of adaptive behavior. *American Journal of Mental Deficiency, 86*, 127–137.
Cicchetti, D. V., Lord, C., Koenig, K., Klin, A., & Volkmar, F. R. (2008). Reliability of the autism diagnostic interview: Multiple examiners evaluate a single case. *Journal of Autism and Developmental Disorders, 36*(4), 764–770.
Cohen, J. (1960). A coefficient of agreement for nominal scales. *Education and Psychological Measurement, XX*(1), 37–46.
Davidoff, F., DeAngelis, C., Drazen, J., et al. (2001). Sponsorship, authorship, and accountability. *JAMA, 286*(10), 1232–1234. doi:10.1001/jama.286.10.1232/data/Journals/JAMA/4799/JED10056.pdf.
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.
Ebel, R. L. (1951). Estimation of the reliability of ratings. *Psychometrika, 16*(4), 407–424.
Fisher, R. A. (1925). *Statistical methods for research workers*. Edinburgh: Oliver and Boyd.
Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin, 76*(5), 378–382.
Freyne, J., Coyle, L., Smyth, B., & Cunningham, P. (2010). Relative status of journal and conference publications in computer science. *Communications of the ACM, 53*(11), 124–132. doi:10.1145/1839676.1839701.

Godlee, F., Gale, C. R., & Martyn, C. N. (1998). Effect on the quality of peer review of blinding reviewers and asking them to sign their reports a randomized controlled trial. *JAMA, 280*(3), 237–240.
Goodman, S. N., Berlin, J., Fletcher, S. W., & Fletcher, R. H. (1994). Manuscript quality before and after peer review and editing at annals of internal medicine. *Annals of Internal Medicine, 121*(1), 11–21.
Grudin, J. (2010). Conferences, community, and technology: Avoiding a crisis. In *iConference 2010*.
Hosmer, D. W., & Lemeshow, S. (2000). *Applied logistic regression*. Chichester: Wiley.
Ingelfinger, F. J. (1974). Peer review in biomedical publication. *American Journal of Medicine, 56*(5), 686–692.
Jacso, P. (2010). Metadata mega mess in Google Scholar. *Online Information Review, 34*(1), 175–191.
Jefferson, T., Alderson, P., Wager, E., & Davidoff, F. (2002a). Effects of editorial peer review: A systematic review. *JAMA, 287*(21), 2784–2786.
Jefferson, T., Wager, E., & Davidoff, F. (2002b). Measuring the quality of editorial peer review. *JAMA, 287*(21), 2786–2790.
Kassirer, J. P., & Campion, E. W. (1994). Peer review: Crude and understudied, but indispensable. *Journal of American Medical Association, 272*(2), 96–97.
Katz, D. S., Proto, A. V., & Olmsted, W. W. (2002). Incidence and nature of unblinding by authors: Our experience at two radiology journals with double-blinded peer review policies. *The American Journal of Roentgenology, 179*, 1415–1417.
Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika, 30*(1–2), 81–93.
Krapivin, M., Marchese, M., & Casati, F. (2010). Exploring and understanding citation-based scientific metrics. *Advances in Complex Systems, 13*(1), 59–81.
Kruskal, W. H., & Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association, 47*(260), 583–621.
Li, X., Thelwall, M., & Giustini, D. (2012). Validating online reference managers for scholarly impact measurement. *Scientometrics 91*(2), 461–471. doi:10.1007/s11192-011-0580-x. http://www.akademiai.com/content/35146TH23T1J1284.
Link, A. M. (1998). US and non-US submissions an analysis of reviewer bias. *JAMA, 280*(3), 246–247.
Lock, S. (1994). Does editorial peer review work?. *Annals of Internal Medicine, 121*(1), 60–61.
Lokker, C., McKibbon, K. A., McKinlay, R. J., Wilczynski, N. L., & Haynes, R. B. (2008). Prediction of citation counts for clinical articles at two years using data available within three weeks of publication: Retrospective cohort study. *British Medical Journal, 336*(76450), 655–657.
Madden, S., & DeWitt, D. (2006). Impact of double-blind reviewing on sigmod publication rates. *ACM SIGMOD Record, 35*(2), 29–32.
McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods, 1*(1), 30–46.
Montgomery, A., Graham, A., Evans. P., & Fahey, T. (2002). Inter-rater agreement in the scoring of abstracts submitted to a primary care research conference. *BMC Health Services Research, 2*(1), 8.
Ragone, A., Mirylenka, K., Casati, F., & Marchese, M. (2011). A quantitative analysis of peer review. In E. Noyons & P. Ngulube (Eds.), *Proceedings of ISSI 2011—The 13th IIternational conference on scientometrics and Iiformetrics*, South Africa, Durban, July 4–7, pp. 724–746.
Reinhart, M. (2009). Peer review of grant applications in biology and medicine. Reliability, fairness, and validity. *Scientometrics, 81*(3), 789–809.
Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *International Statistical Review, 86*(2), 420–428.
Smith, R. (2006). Peer review: A flawed process at the heart of science and journals. *JRSM, 99*(4), 178.
Spier, R. (2002). The history of the peer-review process. *Trends in Biotechnology, 20*(8), 357–358.
Tung, A. K. H. (2006). Impact of double blind reviewing on sigmod publication: A more detail analysis. *SIGMOD Record, 35*(3), 6–7.
van Rooyen, S., Godlee, F., Evans, S., Black, N., & Smith, R. (1999). Effect of open peer review on quality of reviews and on reviewers' recommendations: A randomised trial. *British Medical Journal, 318*, 23–27.
Walsh, E., Rooney, M., Appleby, L., & Wilkinson, G. (2000). Open peer review: A randomised controlled trial. *The British Journal of Psychiatry, 176*, 47–51.
Welch, B. L. (1947). The generalization of student's problem when several different population variances are involved. *Biometrika, 34*(1/2), 28–35.
Wenneras, C., & Wold, A. (1997). Nepotism and sexism in peer-review. *Nature, 387*, 341–343.
Zuckerman, H., & Merton, R. (1971). Patterns of evaluation in science: Institutionalisation, structure and functions of the referee system. *Minerva, 9*, 66–100. doi:10.1007/BF01553188.