International Journal of Information Management 46 (2019) 286–293

Contents lists available at ScienceDirect

International Journal of Information Management

journal homepage: www.elsevier.com/locate/ijinfomgt

An overview of assessing the quality of peer review reports of scientific articles

Amanda Sizo[a,*], Adriano Lino[a], Luis Paulo Reis[b], Álvaro Rocha[a]

[a] Centre for Informatics and Systems, Department of Informatics Engineering, University of Coimbra, Coimbra, Portugal
[b] Department of Informatics Engineering, LIACC - Artificial Intelligence and Computer Science Laboratory, Faculty of Engineering of the University of Porto, Porto, Portugal

ARTICLE INFO

Keywords:
Systematic literature review
Peer review
Reviewers' report assessment
Reviewers' report quality

ABSTRACT

Assuring the quality control of publications in the scientific literature is one of the main challenges of the peer review process. Consequently, there has been an increasing demand for computing solutions that will help to maintain the quality of this process. Recently, the use of Artificial Intelligence techniques has been highlighted, applied in the detection of plagiarism, bias, among other functions. The assessment of the reviewer's review has also been considered as important in the process, but, little is known about it, for instance, which techniques have been applied in this assessment or which criteria have been assessed. Therefore, this systematic literature review aims to find evidence regarding the computational approaches that have been used to evaluate reviewers' reports. In order to achieve this, five online databases were selected, from which 72 articles were identified that met the inclusion criteria of this review, all of which have been published since 2000. The result returned 10 relevant studies meeting the evaluation requirements of scientific article reviews. The review revealed that mechanisms to rank review reports according to a score, as well as the word analysis, are the most common tools, and that there is no consensus on quality criteria. The systematic literature review has shown that reviewers' report assessment is a valid tool for maintaining quality throughout the process. However, it still needs to be further developed if it is to be used as a resource which surpass a single conference or journal, making the peer review process more rigorous and less based on random choice.

1. Introduction

The review of scientific papers is a mechanism used to evaluate and preserve the reliability of manuscripts reporting on scientific discoveries. Peer review is a qualitative evaluation system. It involves experts judging the quality of studies produced by their peers, and one of its main characteristics is the intrinsic subjectivity of the evaluation system. Although this system, which began its evolution in the mid-18th century (Davyt & Velho, 2000), has been criticized and questions have been raised over its deficiencies, it is the most widely used system by publishing groups and it is considered an integral part of the scientific communication system. In addition, no other procedures have so far been able to replace it.

Assuring the quality control of publications in the scientific literature is one of the main challenges of the peer review process. Consequently, there has been an increasing demand for computing solutions for maintaining the quality of this process. Last year, BioMed Central (Burley & Moylan, 2017) published a paper with their perspective of how the peer review process will look in 2030, forecasting that it will become a fully automated process through the use of Artificial Intelligence (AI) techniques. AI has been highlighted and suggested as a resource to automatize procedures within the process itself. Through tools which help to underpin the accomplishment of steps in the peer review process, such as assigning the article to a reviewer or assessing the review reports (Price & Flach, 2017).

The activity of classifying or assessing article reviews has often been reported as an activity that consumes time and effort by the editor. Editors grade each review according to its quality based on several criteria that have been found to be of value. However, these criteria have not been well disseminated (DeMaria, 2003). Without any form of formal feedback on review quality, it is difficult for reviewers to know if their reviews have been considered helpful and appropriate by the author and/or editor (Ward, Graber, & Mars, 2015). Some journals and conferences provide each reviewer with comments from other reviewers, which may give them an idea of whether their reviews are similar in terms of their positive and negative aspects. Although useful, this feedback should not be considered by the reviewer as a criterion measuring the quality of his/her review. Moreover, reviewers of a

* Corresponding author at: Centro de Informatica e Sistemas, Universidade de Coimbra DEI, Polo 2, Pinhal de Marrocos, 3030-290 Coimbra, Portugal.
E-mail addresses: sizo@dei.uc.pt (A. Sizo), adlino@dei.uc.pt (A. Lino), lpreis@fe.up.pt (L.P. Reis), amrocha@dei.uc.pt (Á. Rocha).

https://doi.org/10.1016/j.ijinfomgt.2018.07.002
Received 8 May 2018; Received in revised form 9 July 2018; Accepted 12 July 2018
Available online 20 July 2018
0268-4012/

single article differ in their opinions regarding the quality of the article (Lovejoy et al., 2011) which may be at the same time a confounding factor for reviewers but a valuable aspect for the editor's decision making.

A previous review (Jefferson et al., 2002) presented studies on rating methods used in studies of editorial peer review prior to the year 2000. The author concluded that most of published studies refer their own rating instrument. These tools included between 7 and 36 items rated using 2 to 10-point scales. The results showed the scales appeared to be invalidated or it was found to have low reliability.

Consequently, this literature review aims to establish an updated baseline on solutions for the classification and evaluation of peer review reports. The methodology is based on performing an analysis primarily on the rating method, quality criteria and the technologies that support the process computationally. The main purpose is to identify the level of automation of this activity and if it is aligned with the advances of the use of AI techniques for similar tasks.

More specifically, the research questions of this study are as follows:
RQ1. How are reviewers' reports assessed in the peer review process?
RQ2. What are the different tools and techniques for assessing review reports?
RQ3. What are the quality criteria used in the assessment?

This study uses a systematic literature review to identify solutions for assessing the quality of reviewers' reports. Therefore, a search was conducted in the main online databases on solutions that have been proposed over the past 17 years. Because the first review (Jefferson et al., 2002) presented studies that had been published prior to the year 2000, this current review focuses on the period covered since then. The search revealed 72 studies proposing review assessment mechanisms. Of these, 62 studies were excluded according to various inclusion and exclusion criteria. Therefore, 10 relevant studies were identified as containing elements satisfying the research questions.

The main contribution of this review is to provide readers with a comprehensive understanding of the review report assessment of scientific articles within the peer review process, as well as of the state of the art techniques currently being used in this scientific field. It is expected that this study will be able to promote methods for advancing the study area through computational support.

Beyond this introduction, the paper is organized into the following sections: 2 background, 3 methodology, 4 results, 5 discussion and 6 conclusion.

2. Background

Most of the studies to score the quality of peer reviews reports were based on the criteria proposed by Van Rooyen, that proposed a review quality instrument (Van Rooyen et al., 1999) with of 8 criteria. Each one reflects a different criterion of the review importance of the research question, originality of the research, method, presentation, constructiveness of comments, substantiation of comments, and interpretation of results. Although this instrument is considered reliable, it only classifies the cited criteria, which were defined by an editorial board of a single journal. On the other hand, Jefferson's review describes that the most often rated criteria of reviews were those relating to the methodological soundness, importance, originality and presentation of the reviewed study.

The different perceptions about the quality criteria are mainly due to the subjectivity inherent in the elaboration of the review report. The definitions about what should appear in a report vary according to the author and his/her own experience as a reviewer or editor. In addition, there are different types of research (systematic literature review, randomized controlled trial, among others), as well as different areas of knowledge (mathematics, medicine, among others) that considered different criteria for their evaluation. For this literature review, the studies analysed depict the researches related to the preparation of review reports of the original scientific article, that is, that follows the structure of standard scientific writing and that could be applied in any area of knowledge.

The literature search identifies a series of criteria that are used to elaborate review reports within the peer review process of scientific articles. These criteria can be divided into two main categories: (1) assessment of article and (2) assessment of review report. A series of studies have examined several points of view related to these criteria, which will be outlined below. It is important to note that these criteria are focused around recommendations based upon the assessment of original research articles; thus, some points may not be applicable to all types of articles.

2.1. Assessment of article

The assessment of article is relating with the evaluation of the content and format of article and includes judgements on the quality of the research narrative, according to sections of the article, assessing the impact of the research in its field, as well as concerns regarding ethics and stylerequirements.

2.1.1. Content quality

This is the most cited criterion in the literature, and is related to the research content itself. The main objective of this assessment criterion is to identify inconsistencies and ambiguities in the presentation of information (e.g., the literature review and framework, methods, analysis, interpretation and the discussion of the implications of findings) (Drotar, 2008). There are resources that guide the reviewer in their evaluation of an article's content (Allen, 2013), either in the form of checklists (Tandon, 2014) or questionnaires (Annesley, 2012). These features apply to all sections of the article by providing a list of considerations for each section, so that the reviewer can gather information when reading the manuscript.

2.1.2. Impact quality

Assessing review quality also involves an assessment of the impact of a research on its field. It is an explicit judgement which asks whether the research addresses a relevant and significant issue in its field of study, and whether it has the ability to advance or positively impact science (Drotar, 2008). This area of evaluation also considers the importance and meaning of the research question, the originality of the research and the identification of major methodological problems in the design, measurement and statistical analysis that might limit the scientific contribution of the article.

2.1.3. Ethics aspects

Several important ethical issues need to be assessed by reviewers and managed by editors (Drotar, 2008). The editor wants to ensure that the article to be published is beyond question from a scientific and ethical perspective, because this will encourage scientific community to read the journal's articles due to its credible reputation (Tandon, 2014). Moreover, this criterion includes ethical concerns related to the checking of references, conflicts of interest, the detection of plagiarism and redundancies (duplicate content), as well as other issues that depend on the type of study (Rosenfeld, 2010).

2.1.4. Writing quality

A peer-reviewed article is a means of communication; therefore, the writing is also an aspect that should be evaluated through the content of the article. The article should be easy to understand and should be clearly and succinctly organized. The reviewer should consider the writing style and lexical and grammatical errors (Allen, 2013) as well as determine whether the authors have ensured that the article flows appropriately, including a smooth transition between all its sections (Vintzileos & Ananth, 2010). Some journal boards often tell peer

reviewers that they need not concern themselves with spelling or grammar. Nevertheless, this is a much-cited criterion in the published guidance.

2.2. Assessment of review reports

The review report generally contains two parts. Usually, the first part "Comments to the author" describes all the considerations of evaluation of research. The second part, "Comments to the editor" contains the justification on the suggested decision as to whether or not to accept the article for publication. In addition to the aspects related to the research described above, the evaluation of the report considers two criteria:

2.2.1. Written in a positive tone

Reviewers should maintain a professional and respectful tone throughout their review and provide corrective feedback that improves the scientific merit of the manuscript. The use of pejorative and demeaning language undermines the fundamental purpose of peer reviews (Lovejoy et al., 2011). Moreover, reviews conveyed in a negative tone may not only offend the author, but may also damage the journal's credibility (Ward et al., 2015).

2.2.2. Well-presented and organized

A well-structured review report helps the author to identify points of correction and provides the editor with an adequate visualization for decision-making. Rosenfeld (Rosenfeld, 2010) recommends that the review organize their writing around "major points", which are those issues that are critical to the validity of the study, and "minor points", which are important for correcting the study but are not critical. Each concern should be a separate paragraph, to facilitate a point-by-point response from the authors. Consequently, the numbering of each specific point of criticism facilitates the authors' organization when responding to criticism (Drotar, 2008). The comments should refer to the page and line number, which in turn may help the editor and author to pinpoint precisely where an issue of concern is located in the manuscript.

This literature review wants to identify what criteria are used to evaluate the quality of a review report and whether they are in accordance with the criteria aforementioned in literature.

3. Methodology

This study uses a systematic review guideline (Kitchenham & Charters, 2007), which is a recommended approach for revisions in the field of software engineering and technology and obtained relevant search results in different studies like, (Balaid et al., 2016), (Busalim & Hussin, 2016) and (Zahedi, Shahin, & Ali Babar, 2016) and (Rekik et al., 2018).

The reason for using this guideline is to execute a specific method to summarize the evidence, to identify any research gaps in the existing research through a deep investigation and to find out how computational approaches have been applied in the peer review process (Kitchenham & Charters, 2007).

The method of Kitchenham has six distinct stages, which are to: (i) formulate a review protocol; (ii) identify inclusion and exclusion criteria; (iii) describe the search strategy process; (iv) show the selection process; (v) bring about quality consideration; and (vi) use data extraction and synthesis. The details of each stage are reported in the following subsections.

3.1. Review protocol

The protocol identifies the review background, research questions, search strategy, study selection process, quality assessment, data extraction and synthesis of the extracted data. The research questions and the background of this review are found in Sections 1 and 2. Fig. 1 illustrates the plan outlining the conduct of this systematic literature review.

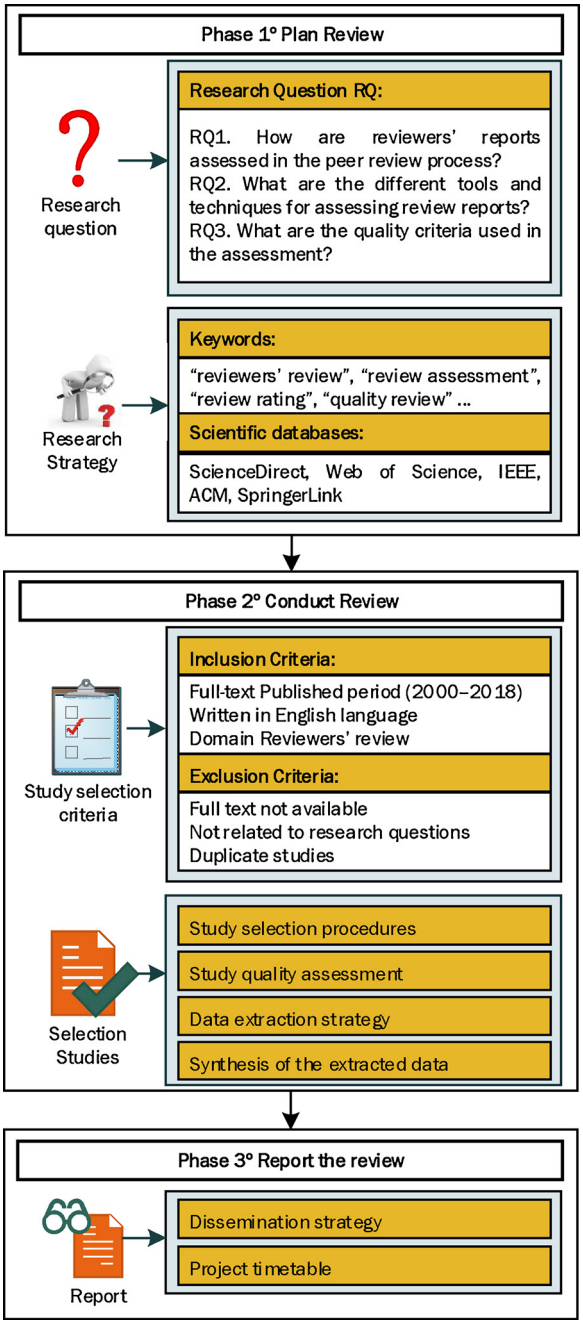

Fig. 1. Protocol review of the study.

3.2. Inclusion and exclusion criteria

Inclusion and exclusion criteria were defined to evaluate each potential primary study, to guarantee that the research was relevant and related to this review. We considered research articles (from journals and conferences) in the English language, published between January 2000 and February 2018, which could be found in digital databases. The reasons for choosing this period is because a previous review (Jefferson et al., 2002) presented studies on quality criteria and rating methods published prior to the year 2000; thus, this current research focuses on the period covered since then.

In this study, articles were included that dealt with the review

assessment of scientific articles within the peer review process. Articles considered to be outside the scope of this research were excluded, such as all unused and directed content in scientific articles, for example, reviews conducted in virtual learning environments and reviews of products or services. Fig. 1 shows the criteria used for this review. For a study to be considered eligible, it must have met all inclusion criteria and must not have met any exclusion criteria.

3.3. Search strategy

At this stage, a search strategy was defined to provide a comprehensive overview of how reviewers' reports on scientific articles are carried out within the peer review process. Five electronic databases were defined: ScienceDirect, Web of Science, IEEE, ACM, SpringerLink. These databases were chosen as they were considered the most relevant and provided the highest impact journals and conference proceedings in the field of information systems. The search was carried out using combinations of keywords such as: "reviewing review", "review assessment", "review rating", "quality review", "assessment of reviewers' reports", "reviewers' review" and "peer review". The search strategy contains the following steps:

1 Explore the titles, abstracts and keywords of the identified articles and select them based on the inclusion criteria.
2 Read those articles that were not eliminated in the previous steps to determine whether they should be excluded from the review according to the exclusion criteria.
3 Try to find new studies by scanning the bibliographies of the eligible articles. If found, these works must also meet the inclusion and exclusion criteria.

The study result was added to the Mendeley application for reference management.

3.4. Study selection process

Searching with these keywords identified 72 articles. After the removal of duplicate studies using Mendeley, 69 remained. Next, the inclusion and exclusion criteria were applied to the abstract of each article. In this step, one article was excluded because full access was denied and 54 full-text articles were excluded because they were not related to the subject of this review. No other study was identified through a reference search, and only four studies were excluded by the quality assessment, which will be explained in the next subsection. A total of 10 articles met the criteria and were eligible for systematic review. A synthesis of this process is illustrated in Fig. 2.

3.5. Quality assessment (QA)

According to Kitchenham (Kitchenham & Charters, 2007), besides general inclusion and exclusion criteria, it is critical to evaluate the "quality" of primary studies. However, for this study, given the small number of studies found, a small checklist was applied for each study.

The identified studies must address the two questions regarding quality assessment:

QA1. Has the research presented, proposed or analysed any mechanism to evaluate or classify review reports?

QA2. Did the research follow a suitable methodology for reviewers' reviews?

Four studies were excluded after applying these two quality criteria. Three were excluded because the articles result of controlled trials and only did a statistical analysis on the reviewer's and editor's opinion on what criteria are used to review articles, but did not present any solutions for assessing the reviewers' reports. Furthermore, one study was excluded because its evaluation of the data was biased, as the corpus of the study was derived from an analysis of the reviewer's own reports.

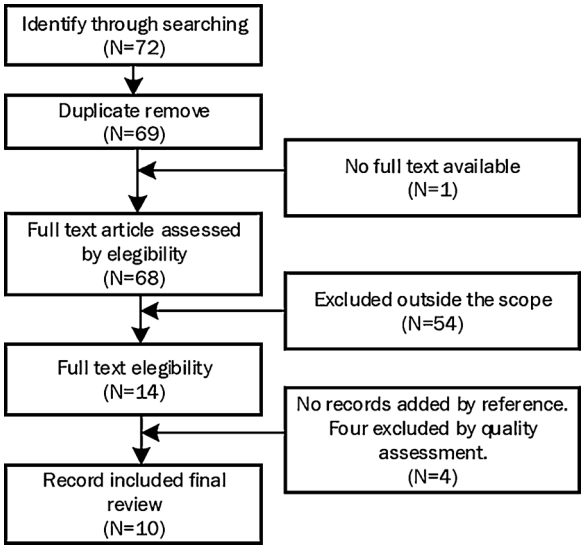

Fig. 2. Study selection process.

3.6. Data extraction and synthesis

Data extraction forms are designed to collect all the information needed to address the research questions. The data extraction strategy was performed by reading each of the ten studies and collecting the required information, which were managed in MS Excel spreadsheets. The content of the spreadsheets was laid out to help identify how the researchers have developed their reliability and eventual limitations. Thus, the following information was included: source, year, quality assessment criteria, rating method, result, application, limitations, computational support and techniques. A description of each item is shown in Table 1. The first author of this review conducted the data extraction process, and the other authors reviewed it.

Data synthesis involves collating and summarizing the results of the included primary studies. The synthesis of this review is descriptive (non-quantitative) due to the small number of studies. Consequently, by investigating the full text of each primary study, the necessary data were synthesized with the intention of answering the research questions. The synthesis results are described in the following sections.

4. Research questions results

4.1. How are reviewers' reports assessed in the peer review process? (RQ1)

Maintaining the quality control of publications in scientific literature is one of the main characteristics of the peer review process (Rowland, 2002), and the evaluation of review reports has been used as an instrument to evaluate this characteristic. In three of the studies, it becomes evident that the evaluation of reviewers by the editorial board

Table 1
Data extraction for each study.

Extracted data	Description
Bibliographic references	Authors, title, publication source and publication year
Quality criteria	Quality assessment criteria
Rating method	Quantitative, qualitative or mixed method
Result	Relevant results of applying the method
Validation	Description of the study domain
Limitations	Limitations of the study
Computational support	Does the method use computational support? (Yes or No) Which one?
Techniques	Does the method use artificial intelligence techniques? (Yes or No) Which one?

Table 2
Studies of the assessment of reviewers' reports.

Study, Year	Validation	Rating Method
(Falkenberg & Soranno, 2018)	Reviews of the Journal Limnology and Oceanography: Letters	Analysis of the characteristics of the reviewer's words, for example, entities and adjectives
(Willy et al., 2017)	Reviews of Binus Business Review	Automatic rating on a scale of 1 to 5
(Thompson et al., 2016)	Reviews of the Journal of Bone and Joint Surgery	Rating reviews between 80 and 100
(Ausloos et al., 2016)	Reviews of the Journal of the Serbian Chemical Society	Statistical analysis of reviews regarding the total number of words and different words
(Bornmann et al., 2012)	Reviews of the Journal of Atmospheric Chemistry and Physics	Textual analysis of reviews to describe their basic linguistic characteristics, like word count, words per sentence, long words as well as those connoting psychological processes
(Callaham & McCulloch, 2011)	Reviews of the Annals of Emergency Medicine	Rating on a scale of 1 to 5
(Henly & Dougherty, 2009)	Reviews of the Nursing Research Journal	Rating on a scale of 1 to 5
(Fortanet, 2008)	Reviews in the field of Linguistics and Business Organization	Classification of the review text as criticism, recommendation patterns or requests
(Shashok, 2008)	Reviews of the European Journal of Epidemiology	Classification of review comments as pertaining to content or writing
(Landkroon et al., 2006)	Reviews of the Dutch Journal of Medicine	Rating on a scale of 1 to 5

is standard practice.

The results show that there are distinct ways to conduct the assessment of the reviewer's review. It is also possible to highlight that mechanisms to rank review reports according to a score, as well as the word analysis, are the most common tools.

Although no study defines the concept of quality, five studies present their own instrument for scoring the quality criteria of the review report according to a scale determined by the editorial board. Although this instrument is considered reliable, the entire task of classification is done manually by the editor. In only one of the cases studied, this classification was carried out automatically.

Thompson (Thompson et al., 2016) published a study that applied an instrument to evaluate the quality of reviews performed by their panel of reviewers. The main purpose was to figure out if members of a journal editorial board could consistently and reliably use a single numeric scoring system to evaluate the quality of peer reviews. A different strategy was proposed by Willy, Priatna, Manalu, Sundjaja, & Noerlina, (2017) in which they presented a recommendation system whereby the editor automatically received a rating. In (Landkroon, Euser, Veeken, Hart, & Overbeke, 2006) the study focused on analysing and assessing the adequacy and reliability of a simple five-point scale; the authors applied an internal validation process, that is, finding correlations between the classifications made by editors of the journal and those made by editors from other journals in the same research area. They also conducted a questionnaire with the authors in order to identify correlations between their classifications and those of the editors.

Furthermore, in Henly and Dougherty's study (Henly & Dougherty, 2009), a continuous improvement framework was used to develop a methodology for assessing the quality of the narrative portion of manuscript reviews submitted to Nursing Research. In two other studies, the main objective was to identify other evaluation characteristics. For example, Callaham and McCulloch's study (Callaham & McCulloch, 2011) characterized changes in review quality by individual peer reviewers over time, and found that 92% of the reviewers demonstrated a very slow but steady deterioration in their scores over the 14 years of the study.

On the other hand, the text analysis just identifies the standard language, the size of the report or make inferences through the quantity of certain words, this kind of evaluation does not have concern about a point of scale or the contemplation of certain quality criteria. Half of the results of this review shows these approaches.

Recently, Falkenberg & Soranno, (2018) proposed a strategy to identify the high level review reports through text mining technique, where the overall length of the report, the type of evaluated entities and also the evaluative adjectives are considered in the study. According to the authors, the reviews rated as highly relevant were typically longer, and had significantly more entities and adjectives than other reviews considered as reasonable ones. In (Ausloos et al., 2016) the study uses a quantifier based on Zipf's law, that allows the evaluation and quantification of deviation between diversity and redundancy of different texts in review reports. In (Bornmann et al., 2012) was used Linguistic Inquiry and Word Count (LIWC), a text analysis software program that counts words in meaningful categories (e.g., positive or negative emotions) using a standardized dictionary. In (Shashok, 2008) the author proposes the identification between content criteria and writing criteria in reviews reports through of a simple classification system, in order to avoid a misunderstood by authors. And in (Fortanet, 2008) proposed a way of assessing the language contained in reviewers' reports in terms of their syntactic and lexical features, in order to identify patterns in relation to particular types of reviews.

Five experiments were related to the fields of medicine and health. In nine of the studies, the experiment was internally validated through the assessment of reviews from a one single journal or scientific area. Only one study conducted an external validation, inviting editors from other journals to compare their responses with the internal assessment, as well as collecting assessments from authors via questionnaires. Furthermore, there was only one study in which the author mentioned that the assessment results should be reported to the reviewer.

A brief description of these ten studies is presented in Table 2.

4.2. What are the different tools and techniques for assessing review reports? (RQ2)

Of the five studies that did some text analysis, only two studies used text mining techniques to map words and make inferences about their value, the others used word counting software or program for language analysis. And differently from what has been expected, only one study used a natural language application.

Three studies used SPSS software (SPSS Inc., Chicago, IL) for statistical analysis, while eight used computer supports in the form of systems or tools to develop the study. The systems specifications are not described, if it is accessed online, for instance.

The systematic review identified three different approaches for assessing the quality of reviewers' reports on scientific articles.

a) Rating by score.

The first approach is focused on assignment the review quality by means of a score. This type of mechanism assigns to the report a score for each criterion preestablished by the editorial board. In some cases, a general score to the review report it is assigned.

The review quality level is represented by score number which

corresponds classification, typically in ranges of a scale from weak to excellent. Moreover, the score can reflect not only one but a set of criteria, where a certain score reflects the comprehensiveness of the report.

For instance, the score for each criterion, such as method, study impact, field contribution, and statistical analysis, can be replaced by a general score represents this set. This indicates that the reviewer appropriately evaluated the methodology, validated the statistical results, and confirmed the contribution to the field. In summary, the score reflects whether the report complies with the set of pre-established criteria.

- Semantic text Mining

This approach uses the linguistic analysis of the report narrative to evaluate the grammatical, lexical and semantic features of report, and also use quantifiers of deviation between diversity and redundancy of different texts. In addition, systems have been applied to identify the domain of each word, for the proper framing, for instance such as the identification of verbs, nouns and adjectives.

As an alternative for text analysis is to evaluate words according to patterns of entities or adjectives. Entities are representations of the key components of the article found in the reviewer's report, such as the description of the method. The "method" is the keyword that represents in the report particularly important content on the evaluation of the search method. Adjectives are the terms used by the reviewer to describe each entity in a way that qualifies it positively or negatively. For example, the description of an "innovative method". "Innovative" is an adjective that positively qualifies the entity "method".

- Assessment by AI with rating.

The third category was the only proposal to be highlighted for its use of computational approaches, with Natural Language Processing (NLP) techniques and unsupervised learning to offer automatic classification. This mechanism uses first a text mining to summarize the report and later an automatic score estimate is generated by using structural features of the review text.

Our findings demonstrated that the published evidence for the effectiveness of computer-based peer review is still in its early stages, and cannot be reliably used in the automatic assistance of reviewers.

4.3. What are the quality criteria used in the assessment? (RQ3)

The ten studies presented criteria defined by editorial boards, which were applied specifically to determine the quality of an individual journal. Table 3 shows the criteria evaluated in the selected studies. The criteria were identified and therefore extracted exactly as mentioned in the cited studies and they are structured according to the approach presented in Section 2.

According to the studies, in descending order of citation, the criterion of content evaluation was the most cited by the studies, i.e., the article's methodological solidity through the logical sequence of its sections. This was followed by criteria related to the impact of the study, which dealt with the importance and originality of the research. Finally, there were criteria related to the tone used in the writing of the report. No author considered any ethical aspects.

In five studies, the evaluated criteria concerned the linguistic characteristics of the comments. These studies made inferences about the reviewers' narratives. For example, in Falkenberg and Soranno's study (Falkenberg & Soranno, 2018), the author finds a relationship between the quality of reviews and the size and vocabulary of the review, considering reviews to be most helpful when they are longer in length, contain fully developed ideas, provide examples and use descriptive language.

Thompson (Thompson et al., 2016) published a study that considered the number of sentences and words in a review when evaluating it, and assigned the highest score to significantly long and detailed reviews. Word count was also used as a criterion in the study of Ausloos et al. (Ausloos et al., 2016) to quantitatively and objectively assess the quality of its linguistic and informational content using an algorithm (the so-called Zipf's law), allowing the evaluation and quantification of deviation between the diversity and redundancy of different texts.

Although the scientific literature addresses an impressive number of criteria, each journal summarizes those that it considers most relevant in the production of a quality report. There is no consensus on quality criteria between journals.

The criteria for reviewing scientific articles do not seem to be sufficiently thought out and, depending on the type of research, different weights are assigned to each one. This may be one of the reasons why this type of application has not been widely adopted or given visibility by the scientific community.

Therefore, more research is still needed to strengthen a deeper understanding among different criteria and how they are related to quality of review report, thus enabling it to be used as a global resource that will assist editors in decision making.

5. Discussion

The results of this study demonstrate that, over the last 17 years and notwithstanding some significant advances, findings in this area are still far from assuring a high-quality system in the peer review process. This is quite different from what may be observed in the evaluation of online reviews of products and services. In these areas, machine learning techniques have been used in different contexts to predict review quality, review sentiment and characteristics of the reviewer in electronic commerce systems of services and products (Lee et al., 2018). In addition, major websites such as Amazon.com and Yelp.com provide the "Most Helpful First" option in sorting and presenting customer reviews, and the characteristics of reviewers are already considered to predict the usefulness of the review (Ngo-Ye & Sinha, 2014).

Defending the different specificities of each context, the challenges that involves the peer review are bigger, mainly when we consider the limitations of funds expend on peer review (Gropp et al., 2017). In general terms, from the results of the literature review, we consider that the following main challenges need to be addressed and overcome:

- Limitations of the business domain: Although the peer review process started many years ago, there is little scientific evidence about its efficiency and quality. Review evaluation systems are specific to a particular journal and to a knowledge area. The diversity of systems and models restricts the applicability of the tools developed and creates a computational difficulty for data sharing. It is a great challenge to develop a global system that can preliminarily capture the basic requirements of all scientific article reviews and that is useful for most journals.
- Ethical problems: The increasing number of false or fraudulent searches creates a scenario where it seems irrelevant to evaluate a review report based on simple criteria such as originality, method and results, for example. In fact, the presented models would not be able to show if the reviewer observed whether the research had plagiarism or not. However, for this purpose there are already widely used plagiarism detection systems (Subroto & Selamat, 2014). The point is that the evaluation of the report should be seen as useful feedback for editor decision-making, increasing the likelihood of creating a pool of quality reviewers and decreasing the probability of problematic publications.
- Subjectivity: Evaluating a review is a complex activity and places a considerable burden on the editor. One of the main problems is the subjective nature of this task. Despite some journals providing guidelines for the review, the reviewer often does not know how to

Table 3
Quality criteria assessed.

Some quality criteria from results		
Scientific assessment	**Article section**	(Callaham & McCulloch, 2011) "Study design, methodology, author's interpretation of the data". (Henly & Dougherty, 2009) "problem statement, related literature/references, theoretical framework, design/methods, data presentation/analysis, discussion/interpretation of results". (Landkroon et al., 2006) "purpose of the study, study design, scientific validity and conclusions". (Willy et al., 2017) "title, abstract, introduction, methodology, results and discussion, references". (Thompson et al., 2016) "methodology".
	Impact	(Henly & Dougherty, 2009) "originality". (Landkroon et al., 2006) "something is new, important and useful to readers". (Thompson et al., 2016) "impact of the research, relevance".
	Writing style	(Callaham & McCulloch, 2011) "comments of the article as a written communication device". (Henly & Dougherty, 2009) "organization, writing style". (Shashok, 2008) "writing and style" (Bornmann et al., 2012) "writing"
	Ethics aspects	–
Report assessment	**Tone**	(Callaham & McCulloch, 2011) "constructive and professional comments". (Henly & Dougherty, 2009) "constructive, knowledgeable, balanced/fair, logical, clear, precise, useful to authors and useful to editor". (Landkroon et al., 2006) "comprehensive, objective, insightful, helpful comments".
	Structure	(Callaham & McCulloch, 2011) "accurate and productive comments". (Fortanet, 2008) "written communication". (Landkroon et al., 2006) "numbered questions, well-documented reasons for decisions made". (Thompson et al., 2016) "significant length".

elaborate a report, what to include in it or the scope of the report, which may lead the reviewer to decide by himself/herself on what is relevant to report. Besides, the guidelines, when available, vary according to journals and research areas. The absence of a reference model weakens the foundation of the definition of quality criteria for the assessment of review reports. The fact that these criteria are created by the editorial board creates the risk that they will be easily discarded or changed as a result of an internal order or through changes in the editorial board.

- Disclosure of the review rating: Only one study stated that the review rating results should be showed to the reviewers. The lack of transparency of the result and criteria prevents the reviewer from knowing their weaknesses and improving their performance in their evaluations. These systems should not create constraints or problems between editors and reviewers, but if the editor is sure about what he/she expects of the reviews it is easier for the reviewer to meet the criteria and provide a report that is useful for the editor decision-making.

Recent works have addressed the need for studies applying computational support in the peer review process (Price & Flach, 2017). However, before creating tools, it is recommended to firstly determine which criteria genuinely add quality to reviews, independent of the journal. A preliminary base model would provide a means of taking a key step towards a system that can be used by the entire scientific community. The intrinsic specificities of the area of knowledge and type of study would be added to this model as new research develops, involving the collaboration of the wider scientific community.

6. Conclusion

The establishment of a high-quality system in the peer review process is currently in high demand, and has been motivated by an increase in electronic publications. In 2002, when Jefferson (Jefferson et al., 2002) conducted a systematic review to identify the outcome measures used to assess editorial peer reviews, he surprisingly found that little was known of their aims or effects. The results of the current study demonstrate that, over the last 17 years and notwithstanding some significant advances, findings in this area are still slow.

One of the motivations behind this research is concerned with recent publications that apply technologies in the review phase of the peer review process. These include the RobotReviewer (Marshall et al., 2016), which is a machine learning system that automatically assesses bias in clinical trials and the StatReviewer (Papatriantafyllou, 2017) which is a tool that assists the review of manuscripts for the statistical area. This directly reflects how reviews will be presented between now and 10 or 20 years' time.

This review used the method defined by Kitchenham et al. (Kitchenham & Charters, 2007) and, after performing multiple processes, 10 studies were identified relating to this review. The main contribution of this literature review was to present mechanisms proposed in the assessment of reviews of scientific article published between 2000 and 2018 in five research databases. As a secondary contribution, the state-of-the-art techniques in the scientific field were outlined, with the expectation that this would facilitate the promotion of means of advancing this area through computational support.

This research highlights two key issues through the few publications found dealing with the topic: a) the difficulty of advancing this phase of the peer review process without a model establishing the quality criteria which can be used as a global resource by the most of journals and b) the opportunity for further computational research dealing with innovative approaches to AI.

Acknowledgement

We appreciate the financial support of AISTI (Iberian Association for Information Systems and Technologies).

References

Allen, T. W. (2013). Peer review guidance: How do you write a good review? *The Journal of the American Osteopathic Association, 113*(12), 918–920. https://doi.org/10.7556/jaoa.2013.070.

Annesley, T. M. (2012). Now you Be the judge. *Clinical Chemistry*. https://doi.org/10.1373/clinchem.2012.195529.

Ausloos, M., Nedic, O., Fronczak, A., & Fronczak, P. (2016). Quantifying the quality of peer reviewers through Zipf's law. *Scientometrics, 106*(1), 347–368. https://doi.org/10.1007/s11192-015-1704-5.

Balaid, A., Abd Rozan, M. Z., Hikmi, S. N., & Memon, J. (2016). Knowledge maps: A systematic literature review and directions for future research. *International Journal of Information Management, 36*(3), 451–475. https://doi.org/10.1016/J.IJINFOMGT.2016.02.005.

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. https://doi.org/10.1007/s11192-011-0569-5.

Burley, R., & Moylan, E. (2017). *What might peer review look like in 2030?* London, UK: BioMed Centralhttps://doi.org/10.6084/M9.FIGSHARE.4884878.V1.

Busalim, A. H., & Hussin, A. R. C. (2016). Understanding social commerce: A systematic literature review and directions for further research. *International Journal of*

Information Management, 36(6), 1075–1088. https://doi.org/10.1016/J.IJINFOMGT.2016.06.005.
Callaham, M., & McCulloch, C. (2011). Longitudinal trends in the performance of scientific peer reviewers. *Annals of Emergency Medicine, 57*(2), 141–148. https://doi.org/10.1016/j.annemergmed.2010.07.027.
Davyt, A., & Velho, L. (2000). The evaluation of science and peer review: Past and present. What will the future be like? *História, Ciências, Saúde-Manguinhos, 7*(1), 93–116. https://doi.org/10.1590/S0104-59702000000200005.
DeMaria, A. N. (2003). What constitutes a great review? *Journal of the American College of Cardiology, 42*(7), 1314–1315. https://doi.org/10.1016/j.jacc.2003.08.020.
Drotar, D. (2008). Editorial: How to write effective reviews for the journal of pediatric psychology. *Journal of Pediatric Psychology, 34*(2), 113–117. https://doi.org/10.1093/jpepsy/jsn142.
Falkenberg, L. J., & Soranno, P. A. (2018). Reviewing reviews: An evaluation of peer reviews of journal article submissions. *Limnology and Oceanography Bulletin, 27*(1), 1–5. https://doi.org/10.1002/lob.10217.
Fortanet, I. (2008). Evaluative language in peer review referee reports. *Journal of English for Academic Purposes, 7*(1), 27–37. https://doi.org/10.1016/j.jeap.2008.02.004.
Gropp, R. E., Glisson, S., Gallo, S., & Thompson, L. (2017). Peer review: A system under stress. *BioScience, 67*(5), 407–410. https://doi.org/10.1093/biosci/bix034.
Henly, S. J., & Dougherty, M. C. (2009). Quality of manuscript reviews in nursing research. *Nursing Outlook, 57*(1), 18–26. https://doi.org/10.1016/j.outlook.2008.05.006.
Jefferson, T., Wager, E., & Davidoff, F. (2002). Measuring the quality of editorial Peer review. *JAMA, 287*(21), 2786. https://doi.org/10.1001/jama.287.21.2786.
Kitchenham, B., & Charters, S. (2007). Guidelines for performing systematic literature reviews in software engineering version 2.3. *Engineering (Beijing, China), 45*(4ve), 1051. https://doi.org/10.1145/1134285.1134500.
Landkroon, A. P., Euser, A. M., Veeken, H., Hart, W., & Overbeke, A. J. P. M. (2006). Quality assessment of reviewers' reports using a simple instrument. *Obstetrics and Gynecology, 108*(4), 979–985. https://doi.org/10.1097/01.AOG.0000231675.74957.48.
Lee, P.-J., Hu, Y.-H., & Lu, K.-T. (2018). Assessing the helpfulness of online hotel reviews: A classification-based approach. *Telematics and Informatics, 35*(2), 436–445. https://doi.org/10.1016/J.TELE.2018.01.001.
Lovejoy, T. I., Revenson, T. A., & France, C. R. (2011). reviewing manuscripts for peer-review journals: A primer for novice and seasoned reviewers. *Annals of Behavioral Medicine, 42*(1), 1–13. https://doi.org/10.1007/s12160-011-9269-x.
Marshall, I. J., Kuiper, J., & Wallace, B. C. (2016). RobotReviewer: Evaluation of a system for automatically assessing bias in clinical trials. *Journal of the American Medical Informatics Association, 23*(1), 193–201. https://doi.org/10.1093/jamia/ocv044.
Ngo-Ye, T. L., & Sinha, A. P. (2014). The influence of reviewer engagement characteristics on online review helpfulness: A text regression model. *Decision Support Systems, 61*, 47–58. https://doi.org/10.1016/J.DSS.2014.01.011.
Papatriantafyllou, M. (2017). Peer review - The future is here. *FEBS Letters, 591*(18), 2789–2792. https://doi.org/10.1002/1873-3468.12792.
Price, S., & Flach, P. A. (2017). Computational support for academic peer review: A perspective from artificial intelligence. *Communications of the ACM, 60*(3), 70–79. https://doi.org/10.1145/2979672.
Rekik, R., Kallel, I., Casillas, J., & Alimi, A. M. (2018). Assessing web sites quality: A systematic literature review by text and association rules mining. *International Journal of Information Management, 38*(1), 201–216. https://doi.org/10.1016/J.IJINFOMGT.2017.06.007.
Rosenfeld, R. M. (2010). How to review journal manuscripts. *Otolaryngology - Head and Neck Surgery, 142*(4), 472–486. https://doi.org/10.1016/j.otohns.2010.02.010.
Rowland, F. (2002). The peer-review process. *Learned Publishing, 15*(4), 247–258. https://doi.org/10.1087/095315102760319206.
Shashok, K. (2008). Content and communication: How can peer review provide helpful feedback about the writing? *BMC Medical Research Methodology, 8*(1), 3. https://doi.org/10.1186/1471-2288-8-3.
Subroto, I. M. I., & Selamat, A. (2014). Plagiarism detection through internet using hybrid artificial neural network and support vectors machine. *TELKOMNIKA (Telecommunication Computing Electronics and Control), 12*(1), 209–218. Retrieved from http://jogjapress.com/index.php/TELKOMNIKA/article/view/648.
Tandon, R. (2014). How to review a scientific paper. *Asian Journal of Psychiatry, 11*, 124–127. https://doi.org/10.1016/j.ajp.2014.08.007.
Thompson, S. R., Agel, J., & Losina, E. (2016). The JBJS Peer-review scoring scale: A valid, reliable instrument for measuring the quality of peer review reports. *Learned Publishing, 29*(1), 23–25. https://doi.org/10.1002/leap.1009.
Van Rooyen, S., Black, N., & Godlee, F. (1999). Development of the review quality instrument (RQI) for assessing peer reviews of manuscripts. *Journal of Clinical Epidemiology, 52*(7), 625–629. https://doi.org/10.1016/S0895-4356(99)00047-5.
Vintzileos, A. M., & Ananth, C. V. (2010). The art of Peer-reviewing an original research paper. *Journal of Ultrasound in Medicine, 29*(4), 513–518. https://doi.org/10.7863/jum.2010.29.4.513.
Ward, P., Graber, K. C., & Mars, H.vander (2015). Writing quality Peer reviews of research manuscripts. *Journal of Teaching in Physical Education, 34*(4), 700–715. https://doi.org/10.1123/jtpe.2014-0158.
Willy Priatna, W. S., Manalu, S. R., Sundjaja, A. M., & Noerlina (2017). Development of review rating and reporting in open journal system. *Procedia Computer Science, 116*, 645–651. https://doi.org/10.1016/j.procs.2017.10.035.
Zahedi, M., Shahin, M., & Ali Babar, M. (2016). A systematic review of knowledge sharing challenges and practices in global software development. *International Journal of Information Management, 36*(6), 995–1019. https://doi.org/10.1016/J.IJINFOMGT.2016.06.007.