Revista Hipatia

Estudios filosóficos y sociales sobre
la ciencia y la tecnología

Edición Nº 2 | Año 2020
ISSN 2683-7781
Universidad de Buenos Aires
Ciclo Básico Común

www.revistahipatia.com

Feedback quality according to the type of referees in the peer review process of scientific articles

Calidad de la retroalimentación según el tipo de evaluador en el proceso de revisión por pares de artículos científicos

por Varas Espinoza, German
Univ. Rennes, CNRS, CREM - UMR 6211, F-35000 Rennes, France
german.varas@univ-rennes1.fr

por Sabaj Meruane, Omar
Univ. de La Serena, Chile
omarsabaj@userena.cl

por Pina-Stranger, Alvaro
Univ. Rennes, CNRS, CREM - UMR 6211, F-35000 Rennes, France
alvaro.pina-stranger@univ-rennes1.fr

RESUMEN

El sistema de revisión por pares de artículos científicos es un proceso que determina la producción, difusión y consumo del conocimiento científico. Dada su confidencialidad, la investigación sobre los informes de arbitraje es muy escasa y no considera las características de los evaluadores que brindan retroalimentación a través de estos informes. En el siguiente trabajo hemos establecido que la calidad de la retroalimentación puede variar en función de la claridad de los comentarios y de la experiencia que tengan los árbitros en esta práctica. En este contexto, el propósito de este estudio ha sido determinar

niveles de calidad de la retroalimentación según el tipo de participación de los árbitros. Se analizaron 5,505 comentarios distribuidos en 118 informes enviados al editor de una revista chilena. Luego, se identificaron operativamente cuatro tipos de árbitros según su tipo de implicación en la revista: evaluadores que participan como árbitros una sola vez (EV), evaluadores que participan como árbitros en múltiples ocasiones (EV+), evaluadores que participan como autores una sola vez (A2F), y evaluadores que participan como árbitros y autores en múltiples ocasiones (A2F+). Para determinar los niveles de retroalimentación en los informes, identificamos propósitos comunicativos que agrupamos en tres niveles de calidad (Nivel I, II y III), donde el nivel superior es el más claro y útil para el autor. Los resultados mostraron que los árbitros A2F fueron los que brindaron una mejor retroalimentación, expresada en un bajo número de comentarios poco útiles, un bajo número de comentarios ambiguos y un alto número de comentarios directos y claros. Por el contrario, los árbitros EV proporcionaron una gran cantidad de comentarios poco útiles, mientras que los A2F+ proporcionaron una gran cantidad de comentarios ambiguos y una baja cantidad de comentarios claros y directos. Estos resultados pueden ser muy útiles para comprender cómo se construyen colectivamente el conocimiento y la legitimación científica en el proceso de revisión por pares.

Palabras clave: Calidad de la retroalimentación | Revisión por pares | Revistas científicas | Evaluadores | Informes de evaluación

ABSTRACT

The peer review system of scientific articles is a process that determines the production, diffusion and consumption of scientific knowledge. Given its confidentiality, research on referee's reports is very scarce and does not consider the features of the evaluators providing feedback through these reports. In the following work, it has been established that feedback quality may vary depending on the clarity of comments and the referees' experience. In this context, the purpose of this study was to determine feedback quality levels according to the referees' type of involvement. 5,505 comments issued in 118 reports submitted to a journal were analyzed. Four types of referees were operatively identified according to their type of involvement in the journal: evaluators who participate as referees only one time (EV), evaluators who participate as referees in multiple times (EV+), evaluators who participate as authors only one time (A2F), and evaluators who participate as referees and authors in multiple times (A2F+). To determine feedback levels, communicative purposes present in the corpus of referee reports were grouped into three levels of quality (Level I, II and III), where the higher level was the clearest and most useful for the author. Results showed that A2F referees were the ones

who provided better feedback, expressed in a low number of "useless" comments, a low number of ambiguous comments and a high number of direct and clear comments. On the contrary, EV referees provided a high number of "useless" comments while A2F+ provided a high number of ambiguous comments and a low number of clear and direct comments. These results can be useful to understand how scientific knowledge and legitimation are collectively constructed through peer review.

Key words: Feedback quality | Peer Review | Scientific journals | Referees | Review reports

1 Introduction

The Peer Review Process (PRP) is an evaluation system that has been used, with variations, since the late eighteenth century to legitimize scientific knowledge (Pontille & Torny, 2015). In this process, different actors exchange specific types of texts: the author writes the research article, the editor sends an evaluation request, referees produce an evaluation report, and so on (Sabaj, González & Pina-Stranger, 2016). The social interaction produced in this particular *event sequence* (Paltridge, 2001; Paltridge, 2017) triggers the generation of scientific knowledge, which generally adopts the concrete form of a research article.

Evaluation reports include the recommendation of referees and vary according to the editorial policies. In some journals, manuscripts may be accepted without amendments, accepted with major or minor amendments, or rejected, whereas in other journals manuscripts are never immediately accepted in its original version.

Given the private and confidential nature of the PRP, the access to data for exploring the PRP is limited, and the endeavor to explain how the referee system works is an almost impossible task without the cooperation of editors (Chubin & Hackett, 1990). Consequentially, the evaluation report, as a genre (Bolívar, 2008), has been scarcely studied due to its private and occluded character (Swales, 1996).

The PRP serves different purposes depending on the interests of the actors involved (i.e., authors, evaluators and editors). For submitting authors, the PRP is a mechanism by which they can improve the quality of their work in order to get published; for reviewers, it is an opportunity to access the community of their discipline (Bornmann, 2011); and for the journal editors, it is a way of discarding poor studies, ensuring that manuscripts have a fair and unbiased assessment (Bunner & Larson, 2012), and acquiring scientific legitimacy in the publishing industry.

Although specialized journals have validated the PRP as the most used mechanism for the evaluation of scientific articles, the system has been constantly criticized for at least four reasons:

1. Reliability of the process: Campanario (1998) describes two types of errors concerning reliability: Error Type-1, which is accepting an article that should not have been accepted; and Error Type-2, which is rejecting an article that should have been published. Several studies have shown that PRP' reliability is quite low (Campanario, 1998; Gans & Shepherd, 1994; Eckberg, 1991; Kostoff, 1995; Bailar, 1991).
2. Favoritism: Bias towards authors related to the journal, certain topics, disciplinary schools, geographical areas, specific institutions, and positive results has been found (Campanario, 1996; Ceci & Peters, 1982; Yotopoulos, 1961; Cole & Bowerg, 1973; Willis & McNamee, 1990). Ernst and Kienbacher (1991), for example, examined all works submitted in 1990 to four journals from Great Britain, Sweden, Unites States and Germany, and discovered that they were more likely to accept national works. In this sense, Crane (1967) called "Invisible college" to the small community of scientists who exchange information and raise the power positions within certain fields or disciplines.
3. Slowness of the process: the revision process may take on average about 6 months (Campanario, 2002). The length of the process varies according to the publication decision and the degree of agreement among the reviewers (Sabaj, Valderrama, González & Pina-Stranger, 2015a).
4. Coercion of new or risky ideas: the PRP is a natural obstacle to develop ideas that challenge the established knowledge in a discipline. Armstrong (1982) proposed six recommendations to get published: 1) do not choose an important problem; 2) do not challenge established beliefs; 3) do not get astonishing results; 4) do not use simple methods; 5) do not reveal everything; and 6) do not write clearly.

In addition to these criticisms, editors who manage the PRP in academic journals face a demanding mission. The editor's desk is always full of documents and the backlog of work is overwhelming. Editors have the difficult task of finding unbiased and competent reviewers that can provide authors (who are supposed to be expecting useful and insightfully comments) with a fair and timely revision. Good reviewers willing to accept the editor's invitation to participate in the PRP are a scarce resource since the time they spend on evaluating someone else's work means, for them, less time to produce their own work (Cabezas, Sabaj, Varas, & González, 2018). The scarcity of reviewers might be one of the reasons why editors usually contact former reviewers or authors to revise a submitted manuscript.

In this article, we explore two relevant dimensions of the PRP that have not been exhaustively addressed. The first dimension is the quality of evaluator's feedback provided in referee reports, an aspect which is directly related to one of

the purposes of the PRP, specifically, to improve articles submitted to a journal. The second dimension refers to the types of actors and roles played by the participants in the production and evaluation of scientific knowledge. Thus, the main objective of this research is to describe and analyze the discursive quality of feedback present in the comments issued in referee reports according to the type of participation of the referees.

The examination of the relation between discursive and social variables in the PRP, as we do in this research, is relevant for at least three reasons. First, from a purely scientific point of view, it is a window to understand how language varies according to the social trajectories of actors in the academy. The second and third reasons are of applied nature, i.e., data analyzed in this research can be useful for editors to accomplish the task of managing academic journals. On the one hand, thanks to the analysis of feedback quality, they will have linguistic criteria for discerning good from poor reviews. On the other hand, some of the information obtained from this study can help editors select adequate referees matching the characteristics of manuscripts submitted to their journals.

2 Actors and roles in the production and evaluation of science

One of the most recurrent abstractions frequently found among PRP researchers is the separation of the role of participants, i.e., authors, editors and referees. This abstraction ignores the fact that in the scientific world different roles can be fulfilled by the same actor. Only one work (Campanario, 1996), as far as we are aware, has accounted for this social complexity in relation to the multiple roles of actors in the scientific realm. Campanario (1996) identified two types of authors, i.e., those who have a relation with the journal ("journal related-authors"), specifically, as evaluators or editors; and those who have no relation with the journal ("other authors"). In his study, Campanario (1996) found that 58% of the authors who had published in 18 psychology journals had already participated in these journals as editors or evaluators.

Indeed, not all the actors have the same participation in a journal. In this study we have identified four types of referees assuming other roles according to their participation in the journal: the reviewer who participates in a journal as referee only once (EV); the evaluator who participates as referee in multiple times (EV+); the evaluator who participates as referee as well as author only once (A2F); and the evaluator who fulfills both functions in multiple times (A2F+). Different forms of participation of one actor in the PRP may symbolize social relevant

categories, which may account for the socialization of academic trajectories. For instance, an A2F+ is expected to be in a higher symbolic status in the academic career than an EV.

3 Peer review quality

The quality of the PRP has been analyzed from different perspectives. A first group of researchers has explored, from the perspective of authors and editors, the characteristics of the evaluators who produce good quality reviews. Evans, Mcnutt, Fletcher and Fletcher (1993) established that 'good reviewers' were the ones who were under 40 years old, had experience in methodology, and belonged to prestigious institutions. The assessment of quality was obtained from the editors' opinions via questionnaires regarding the referees' performance. Evans et al. (1993)'s questionnaire was modified and validated by van Rooyen, Black and Godlee (1998) and then applied by Black, van Rooyen, Godlee, Smith and Evans (1998) to a set of reviews in the field of general medicine. The conclusions of this last work were mainly two: a) the evaluators who were trained in epidemiology or statistics provided better reviews; and b) the reports written by members of editorial committees were poorer than those of the rest of the referees.

Another group of researchers has been focused on analyzing the effect of PRP on improving the quality of articles. Most of these studies are optimistic in this line, i.e., they assume that manuscripts improve, yet they differ on the degree. For example, some authors (such as Pierie, Walvoort & Overbeke, 1996, and Purcell, Donovan & Davidoff, 1998) have argued that the improvement is substantial especially in terms of formatting, whereas Roberts, Fletcher & Fletcher (1994) have shown that the PRP slightly increases the readability of articles. In an exhaustive review, Lu (2008) has concluded that it is not possible to empirically establish whether the PRP contributes to an increase in the quality of a manuscript submitted for evaluation. Lu's (2008) conclusion is based on the observation that research analyzing the quality of reviews does not start from a definition of what *improving the quality of an article* means (Evans et al., 1993; van Rooyen et al., 1999; Pierie et al., 1996; Purcell et al., 1998).

As can be observed, most of these studies have tended to measure the quality of a review by collecting the opinion of editors and authors through validated instruments or questionnaires. In the present work we assume a rather different approach by relating reviewers with their corresponding feedback quality from a discursive point of view.

3.1 Feedback quality

Quality is a complex construct whose various dimensions are not always related in research. As we have argued, the determination of quality depends on the interests of specific actors. In this investigation, we study the effect of the PRP not on the quality of manuscripts, but on the quality of feedback present in referee reports.

Feedback quality has received a lot of attention from the scientific community. Researchers interested in this topic have published both recommendations through editorials and scientific articles on this issue. As for editorials, Khonen (2017), for example, suggests that reviewers do not need to be specialists, however, "they must know the subject matter well" (p.1243). This author highlights some keys points when evaluating, such as a) the source of subjects, types of controls, comparability, description of treatments, methods and protocols, and fulfilment of ethical requirements; b) statistical analysis, regarding the level of significance and relation between data and conclusions; c) references, regarding the bias of citations; and d) readability, regarding grammar, style and word choice.

Tsang (2014), in his experience as "author, reviewer and Senior editor" (p. 191), recommends that "reviewers should adopt an improvement-focused approach when providing comments but an error-focused approach when checking compliance" (p.191). Interestingly, unlike other authors, Tsang (2014) identifies a common ambiguity problem and suggests that "authors should ask themselves the question: is the comment about an error or a suggestion for improving the manuscript? Note that a suggestion that is forced upon the receiver is not a suggestion but a command" (191). This distinction, as we will try to demonstrate in the present investigation, will be a key point in our research.

The research about *feedback* has been especially profuse in the field of education as well as in second language writing pedagogy, mainly, of Spanish and English (Tapia-Ladino, 2014). In the case of English as Second Language pedagogy, Ferris (1997) made a classification of the different comments made by teachers when providing feedback to students. The study gives evidence on the influence of each type of feedback in the texts produced by students, and concludes that information request, general request and summary comments on grammar lead to substantial modification of the texts produced by students. Ferris (1997) also showed that negative comments were longer and more specific than positive ones. The works by Hyland and Hyland (2006), and by Bitchener and Ferris (2012), are both the most comprehensive explorations of feedback in the context of Second Language Acquisition.

The available definitions of feedback are quite diverse. Gielen et al. (2010), for example, suggest that feedback corresponds to any information provided to the

student about the current state of their learning or performance whereas Kulhavy (1977) restricts it to that information that seeks reinforcing correct behavior, and thus increasing the possibility that a correct answer is repeated. Similarly, Nelson and Schunn (2009) define feedback as a motivational, reinforcing and informative comment. Wiggins (2012) and Sadler (1989), meanwhile, conceptualize feedback as the type of information that allows reaching a certain purpose, e.g., taking care of discrepancies or differences between a current and a reference level (Sadler, 1989).

The properties of good feedback and the classifications are also diverse. For Hatziapostolou and Paraskakis (2010), feedback should be timely, constructive, motivational, personal and affordable. Gielen (2010) argues that feedback should be directly related to the subject matter, be focused on performance and be sufficiently detailed and timely. Wiggings (2012) states that it should be objective, transparent, timely, affordable and consistent.

Chi (1996) classifies feedback in specific types: corrective feedback, which is oriented to the correction of an error; didactic feedback, whose purpose is to exhaustively explain a mistake; suggestive feedback, where the one who gives feedback suggests possible solutions; and ratifying feedback, which seeks to confirm what is right. Regarding suggestive feedback, Shashok (2008) distinguishes the ways through which content is addressed. These categories may correspond to those identified by Cho, Schunn and Charney (2010) as 'directive feedback', i.e., explicit requests for specific changes; and 'non-directive feedback', i.e., general observations that can be applied to any article.

Feedback provided in the context of second language teaching and feedback given in the peer evaluation process of research articles may differ widely. The first one has the dual purpose of being a contribution to the student's learning and his/her performance (Nelson & Schunn, 2009) while the second one aims at helping the author meet the journal editorial criteria and providing the editor with enough information so that he/she can decide whether the manuscript worth being published (Bornmann, 2011).

As we will describe in the following section 3.2, in this work we consider feedback quality as a construct which, based on discursive features, allows authors to obtain clear instructions to improve their manuscripts.

3.2 Discursive Quality of Feedback in Referee Reports

In this work we focus on quality as a construct that can be derived from the discursive analysis of referee reports. The logic of our proposal, which is specific to the PRP, is based on the idea, from the author's point of view, that clear and explicit instructions for modifications are much more useful than general and imprecise suggestions which in the end may confuse the author (Bakanic, McPhail & Simon, 1989; Paltridge, 2015; Tsang, 2014).

Our notion of discursive quality of feedback is based on the model of communicative purposes for referee reports elaborated by Astudillo (2015) and further developed by Sabaj et al. (2018). The communicative purposes express the functions that evaluators perform in their reports, for example, 'to highlight a positive aspect of an author's work' or 'to judge its form'. To determine quality levels, we assume that the 'referee report' genre, like any other, has a more or less stable core of purposes (Varas, 2015). This core of common purposes, which is characteristic of evaluative texts, is composed of functions that, given the nature of the genre, are expected in a referee report, i.e., 'describe general aspects of the article', 'describe an error', 'localize elements in the text', 'indicate what is not necessary' and 'describe what is missing'.

On the base of common communicative purposes, which are theoretically inherent to evaluative texts, three levels of feedback quality composed of *specific purposes* reflecting different degrees of clarity and usefulness for the author were identified. The relationship between common communicative purposes and the exclusive purposes distinguishing quality may be observed in the form of three pyramids that differ each other according to the level of feedback quality (see Figure 1 below).

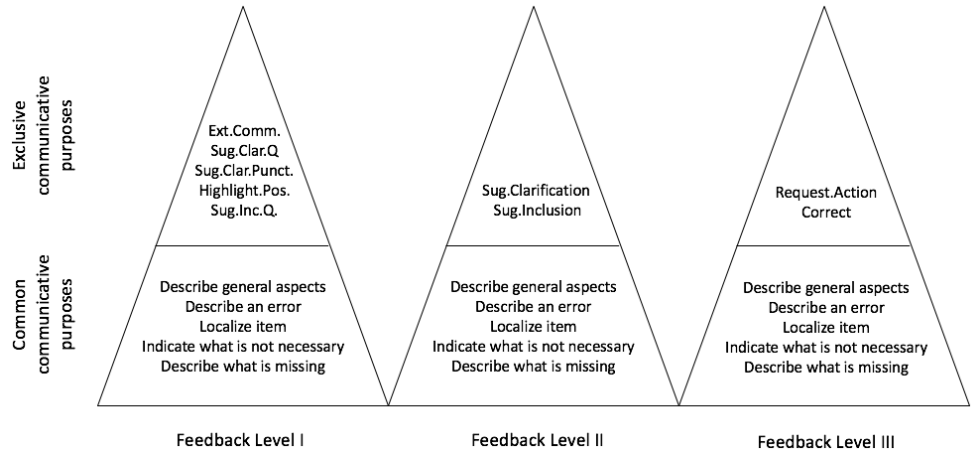

Figure 1. Exclusive communicative purposes (upper part of the pyramid) and Common communicative purposes (lower part of the pyramid) organized by the three levels of feedback.

Level I (low): This level groups communicative purposes such as 'provide External Comments' (Ext.Comm.); 'suggest Clarification through a Question' (Sug.Clar.Q); 'suggest Clarification through Punctuation' (Sug.Clar.Punct.); 'Highlight positive aspects' (Highlight.Pos); and 'Suggest Inclusion through a Question' (Sug.Inc.Q.). Level I comments are not useful enough for an author who wants to improve his/her article. External Comments, for example, do not evaluate the properties of a manuscript; rather, they are intended to contextualize the research and do not mean an immediate contribution to the author. Suggestions of inclusion and suggestions of clarification, both by means of questions, are interesting from

a dialogical point of view because, through this type of statements, the referee establishes a rhetorical interaction with the editor or the author. Highlighting positive aspects may be regarded as "motivational" in terms of Nelson and Schunn (2009) since they identify and reinforce the strengths of the author. However, these comments do not constitute a contribution to the author but a hedging or a politeness strategy, which is typical of this genre, usually, before statements of negative value (Bolívar, 2011, Samraj, 2016). Finally, the clarification through punctuation does not provide any direct feedback; rather, it casts doubt on some aspects of the article without uttering words. Thus, contextualization comments, rhetorical questions, positive remarks and the unconventional use of punctuation (e.g. adding questions marks after a sentence quoted from the author) are considered here as of "little usefulness".

Level II (medium): This level includes the communicative purposes 'Suggest Clarification' (Sug.Clarification) and 'Suggest Inclusion' (Sug.Inclusion). These comments are theoretically quite useful for the author of an article. However, the problem is that suggestions are ambiguous because most of the time they are not optional but mandatory requests conditioning the publication of the manuscript. For example, the statement 'we suggest revising punctuation' is an instruction that the investigator must follow so that the editor can publish his/her research.

Level III (high): This level includes communicative purposes such as 'Request an Action (Request.Action) and 'Correct' (Correct). These purposes are clearer than those of Level II and directly help authors improve their work. Unlike 'Suggest clarification/inclusion', 'Request an action' is more direct, and authors know with certainty that these are mandatory requests conditioning the publication. For example, 'the corpus of analysis must be specified' is clearer and more direct than 'it is suggested to specify the corpus of analysis', since modifying the corpus of analysis is not a suggestion, but an obligatory change. As for Correcting, this purpose is mainly carried out through statements related to aspects of form, such as spelling and punctuation. When few mistakes of this type are made, it is more useful correcting them than providing a general statement such as 'punctuation is suggested to be improved'. Table 1 shows some examples for each communicative purpose:

Table 1. Examples of communicative purposes for each level of feedback in the PRP of scientific articles. Translated by the authors from original examples in Spanish.

Levels	Communicative purposes	Examples
Level I (low)	Provide an external comment (Ext.Comm.)	1. After reading the manuscript, I have realized that the submitting author is actually an amateur. 2. The author should send me the original records to my email in order to verify his claim.
	Highlight positive aspects (Highlight.Pos)	3. The topic is very relevant, and the approach is adequate. 4. The manuscript is interesting and represents a contribution to the study of markers in the American…
	Suggest Clarification through a Question (Sug.Clar.Q)	5. How do the authors assure that the reaction mechanism is modified? 6. Marker in plural second person? That is confusing.
	Suggest Clarification through Punctuation (Sug.Clar.Punct.)	7. In the table, the average values are almost identical ??????? 8. Discursive markers are very important pragmatic categories. ??
	Suggest Inclusion through a Question (Sug.Inc.Q.)	9. There is no citation to the literature of the last 5 or 6 years. Does anyone work in this in the world? 10. There is no number [5]. The entire section has been forgotten?
Level II (medium)	Suggest Clarification (Sug.Clarification)	11. The author is suggested to review some details of writing, style and punctuation. 12. To obtain a panoramic view of the projection of the data, it would be more clarifying to include the number of hours of ….
	Suggest Inclusion (Sug.Inclusion)	13. Perhaps it would be convenient to include a statistical analysis of the results 14. First letter of names should be added.
Level III (high)	Request an action (Request.Action)	15. Clarify the selection criteria for doping percentages 16. Revise translation. It has several grammatical mistakes.
	Correct (Correct)	17. Change "is characterized" by "were characterized" 18. Change “Corpus” by “Methodology”.

Table 2 shows some examples of common communicative purposes:

Table 2. Examples of common communicative purposes in the feedback provided by referees in evaluation reports. Translated by the authors from original examples in Spanish.

Level	Communicative purposes	Examples
Common	Describe general aspects	1. The article mostly discusses how a local level selection test should be updated.
	Describe an error	2. The terminology used in the manuscript does not seem correct. "have reduced the morphological flexion" is understood as a change of morphological level in the forms of voseo, which is not part of the argumentation.
	Describe what is missing	3. The article does not mention some works carried out in Chile in the area of academic writing.
	Indicate what is not necessary	4. Any reference to the theoretical aspects that will be used as criteria for the classification of gastronomic expressions are not necessary in the introduction
	Localize elements in the text	5. In page 12, theoretical considerations about the proposals of Zorraquino and Portolés are provided.

Some of these common and exclusive communicative purposes match the "reviewer roles" identified by Starfield, Paltridge and McMurtrie (2014). Starfield et al. (2014) found, by identifying Halliday's *processes* and *participants* within PhD thesis, that revisors may assume different roles, for example: an *expert role* (e.g., direct mass balance measurements cannot a priori claim to exactly catch the end of the ablation season), a *reporter role* (e.g., this thesis reports the results of what is essentially a single study with boys with ADHD and a control group of normally developing boys), an *evaluator role* (e.g., the candidate shows a good knowledge of this field of research), a *commentator role* (e.g., I wonder if reciprocal blasts were used), an *editor role* (e.g., "to small" should be "too small"), a *mentor role* (e.g., the student should consider publishing the introduction, in a shortened version, as a review paper), and an *examiner role* (e.g., I recommend that the candidate should be awarded the degree on the basis of this thesis).

From our perspective, the comments associated with the *expert* and *reporter* roles correspond to common or core communicative purposes of the peer review report genre, i.e. reviewers always assume this type of role when evaluating a manuscript. The comments associated with the *evaluator* and the *commentator* roles

correspond to feedback Level I, since highlighting a positive aspect and asking for clarification with an indirect question are not useful enough for improving a manuscript. Finally, the comments associated with the *editor*, the *mentor* and the *examiner* role correspond to feedback level II since suggestions may be ambiguous.

4 Methods

Our research is descriptive in scope and is of transactional, nonexperimental qualitative type as it seeks to describe the characteristics of the referee reports provided by the reviewers during a specific period.

4.1 Referees Data and the Corpus of Review Reports

The data used in this research correspond to both sociological attributes of the referees and their discursive behavior crystalized in the comments of review reports. The research was revised and approved by the ethics committees of two institutions, namely, Universidad de La Serena and Pontificia Universidad Católica de Chile. The editor of the journal anonymized each referee report and participated in classifying each reviewer according to their participation. Table 3 shows data of the research:

Table 3. Referee reports according to the type of evaluator

Sociological data		Discursive data	
Type of referee	Referees	Review Reports	Comments
EV	53	53	2,253
EV+	15	37	1,371
A2F	13	13	1,151
A2F+	5	15	730
Total	86	118	5,505

These data were a purposive sample composed of 118 referee reports, issued by 86 evaluators between 2008 and 2012 for 59 manuscripts, which were submitted to Onomázein: Journal of Linguistics, Philology and Translation. All reports were written in Spanish. The journal is published biannually and is funded by the Faculty of Language and Literature of the Pontificia Universidad Católica de Chile, Santiago, Chile. Onomázein uses a double-blind process. Of the 118 referee reports analyzed, 14 (11.86%) recommended accepting the manuscript, 79 (66.94%) recommended conditioning the publication to major and minor modifications, and 25 (21.18%) recommended rejecting the publication of the manuscript.

The evaluation format of the journal contains several constructs and revision sections. Constructs (not analyzed in the present work) consist of propositions, such as "The article shows a clear domain of the work field" or "The methodology used in

the manuscript is coherent with the type of investigation proposed and has been rigorously applied", that the referee must check out through a Likert scale. Revision sections, on the other hand, consist of a space (without limits of words) in which evaluators can write comments accounting for the rejection or providing feedback that the author will have to take into consideration to improve the manuscript and get published.

Most of the reviewers (53) only participated once in the peer review process of Onomázein. 15 referees participated more than once, and 13 were also contributing authors once. 5 referees participated repeatedly both as reviewers and authors. The linguistic data analyzed correspond to a total of 5,505 textual fragments (i.e. statements with a predicative structure) contained in the118 reports which constituted the corpus of analysis.

4.2 Discursive Data Analysis

The tagging of the corpus was performed using ATLAS.ti by a group of 8 researchers. The discursive analysis model was constructed incrementally and was validated in two stages, i.e., calculating the Cronbach's alpha (0.94) and the Intercoder Agreement (0.65). Both measurements showed that the model was reliable since all different analysts agreed on the limits of the communicative purposes and on the assignment of a category (See Sabaj et al., in press).

Each textual segment was labeled with a communicative purpose, which allowed grouping them according to the three levels of feedback (low, medium and high). To associate the levels of feedback with the types of evaluator, referee reports were then assembled according to their corresponding type of referee. Using the same labeling software, feedback levels were thus associated with the types of evaluator. This association allowed obtaining frequencies and percentages.

The relation between referee types and the level of feedback allowed us to know the type of feedback that the different evaluators (i.e., EV, EV+, A2F, A2F+) tended to perform. Finally, to know the type of referee who provided the best feedback, we established that a good feedback is set by a low frequency of Level I, a low frequency of Level II, and a high frequency of Level III feedback (see section 5.2).

5 Results and discussion

5.1 Discursive feedback according to the type of referee

According to our classification of feedback, there are communicative purposes that are common or transversal at the three levels, i.e., describe general aspects, describe an error, localize item, indicate what is not necessary, and describe what is

missing. These common communicative purposes may be regarded, as suggested by Cho et al., (2010) and Shashok (2008), as 'non-directive feedback', i.e., general observations that can be made on any article and which do not imply a direct request for the author. In other words, this type of feedback would correspond to that information which, in terms of Gielen et al. (2010), is provided to generally account for the current state of the performance of the writer instead of asking for specific changes to improve the article.

Table 4. Levels of feedback according to the different types of referees (percentages).

Type of referee	Common communicative purpose	Exclusive communicative purposes		
	(Describe general aspects, describe an error, localize item, indicate what is not necessary, describe what is missing)	Level I (Ext.Comm., Sug.Clar.Q, Sug.Clar.Punct, Highlight.Pos, Sug.Inc.Q.)	Level II (Sug. Clarification, Sug. Inclusion)	Level III (Request Action, Correct)
EV	45.54	32.31	9.19	12.95
EV+	48.23	26.97	10.59	14.20
A2F	50.13	26.22	6.91	16.74
A2F+	46.57	27.86	14.00	11.57

The results confirm the idea of a core of common purposes, which, as can be seen in Table 4 above, varied very little among the different type of reviewers. Regarding the three levels of feedback, although the results are not very divergent, it is possible to establish some micro differences:

1. Level I: With a 5% difference with respect to their peers, the evaluator who participates in the journal only once (EV) is the actor who mostly provides lower quality comments, expressed by communicative purposes such as 'Provide an external comment', 'Highlight positive aspects', 'Suggest Clarification through a Question', 'Suggest Clarification through Punctuation', 'Suggest Inclusion through a Question'; whereas the rest of the evaluators showed a similar result around 27%. Unlike the reviewers with more experience in the journal (e.g., A2Fs), these referees providing low quality feedback are new to the "journal's culture", so they tend to suggest instead of requesting changes and they are more likely to offer positive comments about their peers.
2. Level II: The evaluator who participates in the journal as reviewer and author on several occasions (A2F+) is the actor that provides more "suggestions" (or indirect commands) for actions and inclusions (14%),

whereas the evaluators who participate only once as reviewers and authors (A2F) are the ones who provide fewer comments of this kind, with 6.91%. From an optimistic point of view, it could be established that A2F+s are the evaluators who provide the author with more information and this could be interpreted as a positive aspect. However, if the characteristic of feedback is considered, these are comments of a suggestive kind, i.e., they are characterized by their ambiguity. As previously indicated, these correspond to hedged messages which are typical of this discursive genre. This type of comments can make authors have certain difficulties when interpreting the recommendations since these are not suggestions, but mandatory requests for a change as noted by Tsang (2014). Thus, it can be established that the A2F is the evaluator who provides less ambiguous feedback.

3. Level III: Feedback of this level, which groups the communicative purposes 'Request an Action' and 'Correct', is clearer than that of Level II. Unlike comments of a suggestive type, direct requests for change are considered more useful for the author since they avoid confusion and misinterpretation. The results show a slight difference between the types of actors, where A2Fs provide 5% more comments of this Level III than A2F+s, the lowest pair (16.74% versus 11.57% respectively). This percentage is consistent with the previous result where A2Fs provided fewer comments of suggestive type to issue more direct comments. In the case of A2F+s, they provided more ambiguous comments and fewer direct comments (see Table 4 above).

As has been shown, although the differences are not marked among referees, it is possible to establish that there are certain differences between the levels of feedback according to the type of referee.

5.2 Quality of Feedback according to the Type of Evaluator

The quality of feedback was obtained by grouping the exclusive communicative purposes in three levels of feedback. Level I includes purposes that are not very useful for the author (i.e., Provide an external comment, Highlight positive aspects, Suggest clarification through a question, Suggest clarification through punctuation, and Suggest inclusion through a question). Level II includes comments of a suggestive nature, which run the risk of being misinterpreted or misunderstood (i.e., Suggest clarification or Suggest inclusion). Level III includes feedback through clear and direct communicative purposes (i.e., Request an action and Correct), which aim at directly helping the author.

Thus, it can be established that the best feedback is achieved when a configuration of low Level I, low Level II and a high Level III is set (Figure 2).

Figure 2. Configuration of levels of feedback according to high or low quality.

As seen in Figure 3, the A2F was the actor who provided the least frequency of Level I feedback (26.22% versus 32.31% of EVs), the one who provided the least frequency of Level II feedback (6.91% versus 14% of A2Fs) and the highest frequency of Level III feedback (16.74% versus 11.57 of A2Fs). Therefore, this pattern shows that the type of referee who provides the highest quality feedback to authors is the A2F.

The results also allow identifying who the evaluators that provide the lowest quality feedback are, i.e., those presenting a high frequency of Level I and Level II communicative purposes and a low frequency of Level III purposes. In this case, the referee who provides the highest frequency of the least useful comments for the author (i.e., Level I purposes) is the evaluator who participates in the journal only once (EV). Regarding the frequencies of Level II and Level III, the results showed that the referee who participates in the journal as evaluator and author in multiple times (A2F+) is the one who provides more ambiguous (14% versus 6.91% of A2Fs) and less direct comments (11.57% versus 16.74% of A2Fs), respectively. Therefore, this pattern shows that the type of referee who provides the lowest feedback quality to authors is the A2F+.

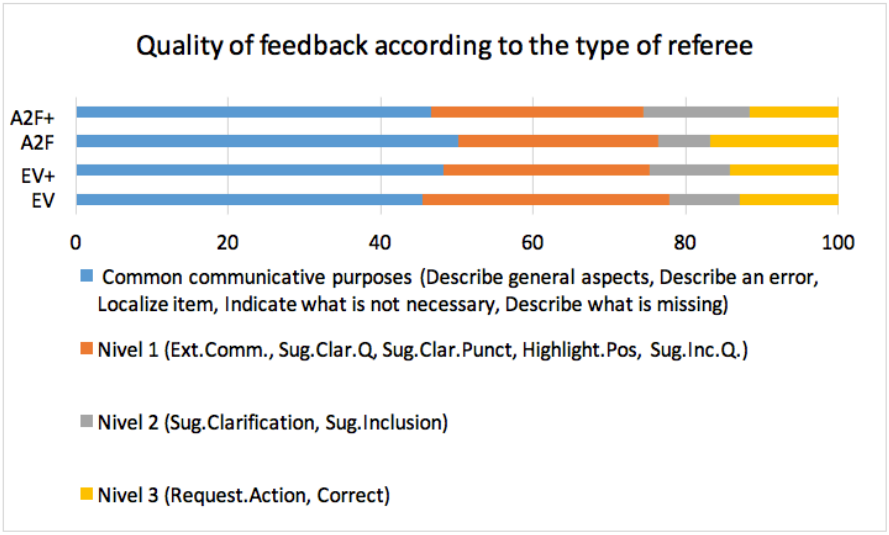

Figure 3. Quality of feedback according to the type of referee.

The A2F referee provides better feedback, probably, because he/she has already had the previous experience of participating in the journal once as an author and thus may be familiar with the allegedly "journal culture". This condition may redound in a greater commitment with the editor of the journal, who could have chosen him/her given the quality of his/her former published manuscript.

Regarding A2F+ low quality feedback, this result is consistent with the research by Stossel (1985) and Campanario (1998), who showed that consolidated researchers provide worse evaluations than the early referees who are just beginning their academic life. This low quality, according to Campanario (1998), may be explained by the lack of time of senior referees due to their various and time-consuming institutional commitments.

However, establishing that senior referees with vast experience in the journal are the worse referees is not totally fair. According to Lock (1985) and Honig (1982), although seniors spend less time on the article, they are quite better than younger referees in comparing results with previous literature and identifying the implications for the future (Evans et al., 1993). This capacity of seniors may be manifested in the high 14% of Level II feedback, made up by the communicative purposes Suggest Clarification and Suggest Inclusion. According to our results, senior referees are more likely to identify what is not clear enough and what is missing rather than correcting a manuscript. Since senior referees are often too busy to evaluate, to the point that some of them just skim the manuscript (Finke, 1990), tasks that take longer seem to be accomplished by those early reviewers who are participating in the journal for the first time as it is the case of EVs, with 32% of Level I feedback versus 27.86% of A2F+.

One last aspect which is interesting about the results regarding A2F+ is his/her tendency to be politer than the other referees. This politeness is expressed by the high frequency of hedged comments and suggestive acts (Level II) and by the low frequency of direct commands (Level III). Bakanic et al. (1989) showed how these suggestions can lead to misunderstanding between authors and evaluators, especially, when they not share the same culture or language. Suid, Sabaj and González-Vergara (2018) conducted a study to analyze the politeness strategies used by referees according to the academic degree of the referees and concluded that M.Sc.'s use more and more varied strategies than B.Sc.'s and Ph.D.'s. Our research, however, showed a clear tendency where the more experienced the referee, the more frequent the use of polite language.

6. Conclusion

Most research on the PRP, probably due to its private and confidential nature, consider the role of authors and evaluators separately (Campanario, 1996; Sabaj et al., 2016). In the case of the journal Onomázein, four types of reviewers were identified. Based on this, two main results were found: first, there are some patterns that relate the type of feedback with the type of referees (EV, EV+, A2F, A2F+) who participate in the same journal; secondly, the evaluators who provide better feedback to the authors are A2Fs whereas the ones who help the authors of the manuscripts the least are EVs.

The results of this research may be useful to understand how the use of language can micro vary according to the degree of socialization of the actors within the peer review process. When it comes to a reiterative contribution for the journal (A2F+), the tendency is to be ambiguous and polite. On the contrary, when it comes to playing both roles, i.e., being an author and a referee only once, the tendency is to find better and direct feedback.

Understanding how different referees may behave discursively may help editors in the task of managing the journal. Imagine an editor who is willing to accept a manuscript, but he/she knows it needs further revisions. By means of skimming through the text, the editor could determine which level of feedback the manuscript needs, and consequently select the most suitable referee. In this regard, our research has also suggested the inconvenience of repeatedly counting on the same actors to serve as referees for the journal since this repetition would imply a decline of the quality of the feedback that evaluators provide in the reports they write. Our data also suggest that asking for a review to a former author, i.e., A2F, is a good business since the editor would count with a report which is likely to include helpful comments.

The description of feedback quality, based on linguistic criteria, is a good source of knowledge for reviewers to write better evaluations. In our opinion, referees should be aware of the importance of the task they do when revising a manuscript for an academic journal.

Being aware of the characteristics of the ideal quality feedback, editors can elaborate or edit evaluation protocols to: a) avoid useless and inefficient comments belonging to Level I; b) encourage clarity when making suggestions (Level II); and c) stimulate direct commands (Level III). If referees knew what is considered to be a good review from the point of view of the editor, referee reports would hopefully contain more valuable comments. Better formats (Sabaj et al., 2015b) may even improve time responses to the editor (Sabaj et al. 2015a).

Editors and referees are not the only ones who can benefit from these findings. Authors, specifically, could have a more conscious idea of how helpful the comments contained in the referee reports are to improve their manuscripts. Also, they may distinguish more clearly the linguistic features of those ambiguous comments and direct commands provided by referees.

Finally, we think that our results can be useful to understand how scientific knowledge and scientific legitimation are collectively constructed through journals peer review.

Acknowledgements

This work was supported by The Chilean National Fund for Scientific and Technological Development (FONDECYT) under Grant Number 1130290.

References

- ARMSTRONG, J. (1982). Barriers to scientific contributions: The author's formula. *Behavioral and Brain Sciences*, 5, 197-199. doi10.1017/S0140525X00011201
- ASTUDILLO, C. (2015). Aplicación de un modelo discursivo para el análisis de los Informes de Evaluación de Artículos Rechazados (IEAR) en el Proceso de Evaluación por Pares (PEP) de tres revistas chilenas (Master dissertation). Universidad de La Serena, La Serena, Chile.
- BAILAR, J. (1991). Reliability, fairness, objectivity and other inappropriate goals in peer review. *Behavioral and Brain Sciences*, 14(1), 137-138. doi: 10.1017/S0140525X00065705
- BAKANIC, V., MCPHAIL, C. & SIMON, R. (1989). Mixed Messages: Referees' Comments on the Manuscripts They Review. *The Sociological Quarterly*, 30(4), 639-654.
- BITCHENER, J. & FERRIS, D. (2012). *Written corrective feedback in second language acquisition and writing*. New York: Routledge.
- BLACK, N., VAN ROOYEN, S., GODLEE, F., SMITH, R. & EVANS, S. (1998). What Makes a Good Reviewer and a Good Review for a General Medical Journal? *Journal of American Medical Association*, 280(3), 231-233.
- BOLÍVAR, A. (2008). El informe de arbitraje como género discursivo en la dinámica de la investigación. *Revista Latinoamericana de Estudios del Discurso*, ALED, 8(1) 41-64.
- BOLÍVAR, A. (2011). Funciones discursivas de la evaluación negativa en informes de arbitraje de artículos de investigación en educación. *Núcleo* 28, 59-89.
- BORNMANN, L. (2011). Scientific peer review. *Annual Review of Information Science and Technology*, 45(1), 197-245.
- BUNNER, C., & LARSON, E. (2012). Assessing the quality of the peer review process: author and editorial board member perspectives. *American journal of infection control*, 40(8), 701-704.
- CABEZAS, P., SABAJ, O., VARAS, G., & GONZÁLEZ, V. (2018). Peering into peer review: Good quality reviews of research articles require neither writing too much nor taking too long. *Transinformação*, 30(2), 209-218. doi: 10.1590/2318-08892018000200006
- CAMPANARIO, J. (1996). The competition for journal space among referees, editors, and other authors and its influence on journals' impact factors. *JASIS*, 47(3), 184-192.
- CAMPANARIO, J. (1998). Peer review for journals as it stands today - Part 1. *Science communication*, 19(3), 181-211.

- CAMPANARIO, J. (2002). El sistema de revisión por expertos (peer review): muchos problemas y pocas soluciones. *Revista española de documentación científica*, 25(3), 267-285.
- CECI, S. & PETERS, D. (1982). A naturalistic study of peer review in psychology: The fate of published articles, resubmitted. *Behavior and Brain Sciences*, 5, 187-252.
- CHI, M. (1996). Constructing self-explanations and scaffolded explanations in tutoring. *Applied cognitive psychology*, 10(7), 33-49.
- CHO, K., SCHUNN, C. & CHARNEY, D. (2006). Commenting on writing: Typology and perceived helpfulness of comments from novice peer reviewers and subject matter experts. *Written Communication*, 23(3), 260-294.
- CHUBIN, D. & HACKETT, E. (1990). *Peerless science: Peer review and US science policy*. New York: Suny Press.
- COLE, R. & BOWERS, T. (1973). Research article productivity of US journalism faculties. *Journalism Quarterly*, 50(2), 246-254.
- CRANE, D. (1967). The Gatekeepers of Science: Some Factors Affecting the Selection of Articles for Scientific Journals. *The American Sociologist*, 2(4), 195-201.
- ECKBERG, D. L. (1991). When nonreliability of reviews indicates solid science. *Behavioral and Brain Sciences*, 14(1), 145-146.
- ERNST, E., & KIENBACHER, T. (1991). Chauvinism. *Nature*, 352-560.
- EVANS, A., MCNUTT, R., FLETCHER, S. & FLETCHER, R. (1993). The characteristics of peer reviewers who produce good-quality reviews. *Journal of general internal medicine*, 8(8), 422-428.
- FERRIS, D. (1997). The influence of teacher commentary on student revision. *Tesol Quarterly*, 31(2), 315-339.
- FINKE, R. (1990). Recommendations for contemporary editorial practices. *American Psychologist*, 45(5), 669-670.
- GANS, J. & SHEPHERD, G. (1994). How are the mighty fallen: Rejected classic articles by leading economists. *Journal of Economic Perspectives*, 8(1), 165-179.
- GIELEN, S., PEETERS, E., DOCHY, F., ONGHENA, P., & STRUYVEN, K. (2010). Improving the effectiveness of peer feedback for learning. *Learning and instruction*, 20(4), 304-315.
- HATZIAPOSTOLOU, T. & PARASKAKIS, I. (2010). Enhancing the impact of formative feedback on student learning through an online feedback system. *Electronic Journal of e-Learning*, 8(2), 111-122.
- HONIG, W. M. (1982). Peer review in the physical sciences: an editor's view. *Behavioral and Brain Sciences*, 5(2), 216-217.
- HYLAND, K. & HYLAND, F. (Eds.). (2006). *Feedback in second language writing: Contexts and issues*. Cambridge University Press.

- KOSTOFF, R. N. (1995). Federal research impact assessment: Axioms, approaches, applications. *Scientometrics*, 34(2), 163-206.
- KOHNEN, T. (2017). How to write a good peer review. *Journal of Cataract & Refractive Surgery*, 43(10), 1243-1244.
- KULHAVY, R. (1977). Feedback in written instruction. *Review of Educational Research*, 47(2), 211-232.
- LOCK, S. (1985). *A Difficult Balance: Editorial Peer Review in Medicine*. Philadelphia: Nuffield Provincial Hospitals Trust.
- LU, Y. (2008). Peer review and its contribution to manuscript quality: an Australian perspective. *Learned Publishing*, 21, 307-318. doi: 10.1087/095315108X323884
- NELSON, M. & SCHUNN, C. (2009). The nature of feedback: How different types of peer feedback affect writing performance. *Instructional Science*, 37(4), 375-401.
- PALTRIDGE, B. (2001). *Genre and the language learning classroom*. Ann Arbor, MI: University of Michigan Press.
- PALTRIDGE, B. (2015). Referees' comments on submissions to peer-reviewed journals: when is a suggestion not a suggestion? *Studies in Higher Education*, 40(1), 106-122.
- PALTRIDGE, B. (2017). *The discourse of peer review. Reviewing submissions to academic journals*. London, England: Palgrave Macmillan.
- PIERIE, J., WALVOORT, H. & OVERBEKE, A. (1996). Readers' evaluation of effect of peer review and editing on quality of articles in the Nederlands Tijdschrift voor Geneeskunde. *The Lancet*, 348(9040), 1480-1483.
- PONTILLE, D., & TORNY, D. (2015). From manuscript evaluation to article valuation: The changing technologies of journal peer review. *Human Studies*, 38(1), 57-79.
- PURCELL, G., DONOVAN, S. & DAVIDOFF, F. (1998). Changes to manuscripts during the editorial process: characterizing the evolution of a clinical paper. *JAMA*, 280(3), 227-228.
- ROBERTS, J., FLETCHER, R. & FLETCHER, S. (1994). Effects of peer review and editing on the readability of articles published in Annals of Internal Medicine. *JAMA*, 272(2), 119-121.
- SABAJ, O., GONZÁLEZ, C., & PINA-STRANGER, Á. (2016). What We Still Don't Know About Peer Review. *Journal of Scholarly Publishing*, 47(2), 180-212.
- SABAJ, O., VALDERRAMA, J. O., GONZÁLEZ, C., & PINA-STRANGER, Á. (2015a). Relationship between the duration of peer-review, publication decision, and agreement among reviewers in three Chilean journals. *European Science Editing*, 41(4), 87-90.

- SABAJ, O., GONZÁLEZ, C., VARAS, G., & PINA-STRANGER, Á. (2015b). A new Form for the Evaluation of Scientific Articles under Peer Review. *Argos*, 32(62), 119-130.
- SABAJ, O.; GONZÁLEZ, C.; ASTUDILLO, C.; VARAS, G.; FUENTES, M.; CABEZAS, P.; SQUADRITO, K. & PINA-STRANGER, A. (2018). El informe de arbitraje: su variación según la recomendación de publicación y la productividad de los evaluadores. *Athenea Digital*, 18(2):1-26.
- SADLER, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science*, 18(2), 119-144.
- SAMRAJ, B. (2016). Discourse structure and variation in manuscript reviews: Implications for genre categorization. *English for Specific Purposes*, 42, 76-88.
- SHASHOK, K. (2008). Content and communication: How can peer review provide helpful feedback about the writing? *BMC Medical Research Methodology*, 8(1), 3.
- SUID, N.; SABAJ, O. & GONZÁLEZ, C. (2018). Ciencia, estatus y cortesía: atenuación en informes de arbitraje de artículos de investigación. *Tonos Digital* 34. Available at: http://www.tonosdigital.es/ojs/index.php/tonos/article/viewFile/1894/983
- STARFIELD, S., PALTRIDGE, B. & MCMURTRIE, R. (2014). *Evaluation and instruction in PhD examiners' reports: Roles and functions*. Paper presentation, AILA congress, Brisbane.
- SWALES, J. (1996). Occluded genres in the academy: The case of the submission letter. In E. Ventola & A. Mauranen (Eds.), *Academic writing: Intercultural and textual issues* (pp.45-58). Amsterdam: John Benjamins.
- TSANG, E. (2014). Ensuring manuscript quality and preserving authorial voice: The balancing act of editors. *Management and Organization Review*, 10(2), 191-197.
- VARAS, G. (2015), *El informe de arbitraje en el proceso de revisión por pares de artículos de investigación: Niveles de retroalimentación según el tipo de evaluador*. (Master dissertation). Universidad de La Serena, La Serena, Chile.
- 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.
- WIGGINS, G. (2012). Seven keys to effective feedback. *Feedback*, 70(1), 10-16.
- WILLIS, C. & MCNAMEE, S. (1990). Social networks of science and patterns of publication in leading sociology journals, 1960 to 1985. *Knowledge*, 11(4), 363-381.
- YOTOPOULOS, P. A. (1961). Institutional affiliation of the contributors to three professional journals. *The American Economic Review*, 51(4), 665-670.