Fallout and Miss in journal peer review

Peer-reviewed author version

EGGHE, Leo & Bornmann, L. (2013) Fallout and Miss in journal peer review. In: JOURNAL OF DOCUMENTATION, 69 (3), p. 411-416.

DOI: 10.1108/JD-12-2011-0053
Handle: http://hdl.handle.net/1942/14333

Fallout and Miss in journal peer review

by

L. Egghe[1,2]
and
L. Bornmann[3]

[1] Universiteit Hasselt (UHasselt), Campus Diepenbeek, Agoralaan, B-3590 Diepenbeek, Belgium[1]
[2] Universiteit Antwerpen (UA), IBW, Stadscampus, Venusstraat 35, B-2000 Antwerpen, Belgium
[3] Max Planck Society, Administrative Headquarters, Hofgartenstr. 8, 80539 Munich, Germany

bornmann@gv.mpg.de
leo.egghe@uhasselt.be

ABSTRACT

Purpose: We further exploit the analogy between journal peer review and information retrieval. In this way we want to quantify some imperfections of journal peer review.
Design/methodology/approach: We define fallout rate and missing rate in order to describe quantitatively the weak papers that where accepted and the strong papers that we missed, respectively. To assess the quality of manuscripts we use bibliometric measures.
Findings: Fallout rate and missing rate are put in relation with the hitting rate and success rate. Conclusions are drawn on what fraction of weak papers will be accepted in order to have a certain fraction of strong accepted papers. These curves are even new in peer review research when interpreted in the information retrieval terminology.

---

[1] Permanent address

I. Introduction

In Bornmann and Egghe (2012), we described already some imperfections of journal peer review (based on the assumption that bibliometric measures can be used to assess the quality of manuscripts). This may be caused by the fact that peer review is a human activity where different referees might have different perceptions on a paper, leading to different advices on the acceptance or rejection of a paper – see e.g. Bornmann and Daniel (2009), Egghe (2010) and Schultz (2010). As a consequence of this, the final decision on acceptance or rejection by the editor of the journal is not always perfect. Most of the strong (good) papers will be accepted and most of the weak (bad) papers will be rejected but it can well be that a strong paper is rejected and that a weak paper is accepted.

Let us fix a set of submitted papers to a journal (in a certain time period). It is clear that the accepted papers are known, but how to determine the strong (good) papers? This is done (see Bornmann and Daniel (2010)) by normalised citation counts of the accepted papers as well as the rejected ones but which are published elsewhere. Normalised citation counts for a single paper are its citation counts divided by the average number of citations per paper in the field under consideration. The papers above a certain threshold of this indicator are defined as qualified (Q). Throughout this paper the term "qualified" will be used instead of the adjective "strong" or "good" paper.

Let us define Ω as the set of submitted papers to a journal that are accepted for publication in this journal or rejected by this journal and published elsewhere. So this is the set we can study on the qualifiedness or non-qualifiedness of the papers (in the definition given above). Ideally, all accepted papers are qualified and all rejected papers (but published elsewhere) are non-qualified but in reality we have four sets of papers:

1. The set of accepted and qualified papers,
2. The set of accepted and non-qualified papers,
3. The set of rejected (but published elsewhere) and qualified papers,
4. The set of rejected (but published elsewhere) and non-qualified papers.

Hence these four sets constitute a partition of the set Ω. Sets 1 and 4 represent correct editorial decisions while sets 2 and 3 represent wrong editorial decisions. As in Bornmann and Egghe (2012) we can compare the above situation with an information retrieval process: interpret the accepted papers as "retrieved" and qualified papers as "relevant" (the papers that we want) in the "database" Ω.

Bornmann and Daniel (2010) give a concrete example of the above situation: 615 papers submitted to the journal *Angewandte Chemie-International Edition* (AC-IE) in the year 2000. These papers have review scores: 12,11,…,1,0 (12 is the best judgement). So we can arrange these papers in decreasing order of these scores and, if review scores are the same, we order the papers randomly by manuscript identifier. For each of these papers we can determine if they are qualified (Q) or not (N), based on normalised citation scores and a threshold (cut-off) value (as described above).

For each $n = 1, 2, 3, ..., 615$ we only consider the first n papers. We consider these papers as "accepted" (A) and, as said above, for each paper we can determine if it is qualified or not. The papers on ranks $n+1, n+2, ...$ are considered as "not accepted" (rejected).

In Bornmann and Egghe (2012) we applied a classical IR evaluation technique by calculating the analogue of precision and recall: precision is the fraction of retrieved papers that are relevant and recall is the fraction of relevant papers that are retrieved. Denote by AQ the number of accepted and qualified papers, by AN the number of accepted and non-qualified papers and by RQ the number of rejected and qualified papers. Then the Success rate S, defined in (1), is the analogue of precision

$$S = \frac{AQ}{AQ + AN} \tag{1}$$

and the hitting rate H, defined in (2), is the analogue of recall

$$H = \frac{AQ}{AQ + RQ} \tag{2}$$

So, for each *n*, we have a point (S,H): this forms the S-H curve which is the equivalent of the precision-recall curve in IR (see also Salton and McGill (1987)). This curve shows which "price" we have to pay in Hitting rate when we want the Success rate to be higher and vice-versa (since the S-H curve in general is decreasing).

In IR, two other evaluation measures exist: fallout (F) and miss (M), see Egghe (2007,2008). In IR terminology, F is the fraction of non relevant papers that has been retrieved and M is the fraction of not retrieved papers that are relevant. In the next section we will interpret these measures in our framework. In the same way as described above we will also produce the curves linking any two of the measures S, H, F and M and give interpretations in terms of the journal decision process.

Fallout and Miss rate and their relations to Success and Hitting rate

The analogue of F in our framework of accepted (A), rejected (R), qualified (Q) and non-qualified (N) papers is: F is the fraction of non qualified papers that is accepted. The analogue of M in our framework is: M is the fraction of rejected papers that are qualified. Denote by AN the number of accepted and non-qualified papers, by RN the number of rejected and non-qualified papers and by RQ the number of rejected and qualified papers, then we have the following formulae for the calculation of F and M:

$$F = \frac{AN}{AN + RN} \tag{3}$$

$$M = \frac{RQ}{RQ + RN} \tag{4}$$

F and M are "negative" measures since, the lower they are, the better our editorial process is (contrary to S en H which are "positive" measures: the higher they are, the better is our editorial process).

We can again use the data from Bornmann and Daniel (2010): 615 papers submitted to AC-IE, arranged in decreasing order of review scores and, if review scores are the same, we order the papers randomly by manuscript identifier. For each of these papers we know as well if they are qualified (Q) or not (N) as described above. For each $n = 1, 2, 3, ..., 615$ we only consider the first *n* papers as "accepted" and the other ones as "rejected". So, for each $n = 1, 2, 3, ..., 615$ we can calculate the values of S, H, F and M and hence the $\binom{4}{2} = 6$ points (H,S), (F,M), (F,H), (M,H), (S,F) and (S,M) yielding (because of the *n*-dependence) six scatterplots. The one (H,S) (the "recall-precision" curve in IR) appeared already in Bornmann and Egghe (2012). We add it here the five other curves in Fig. 1 for the sake of completeness.

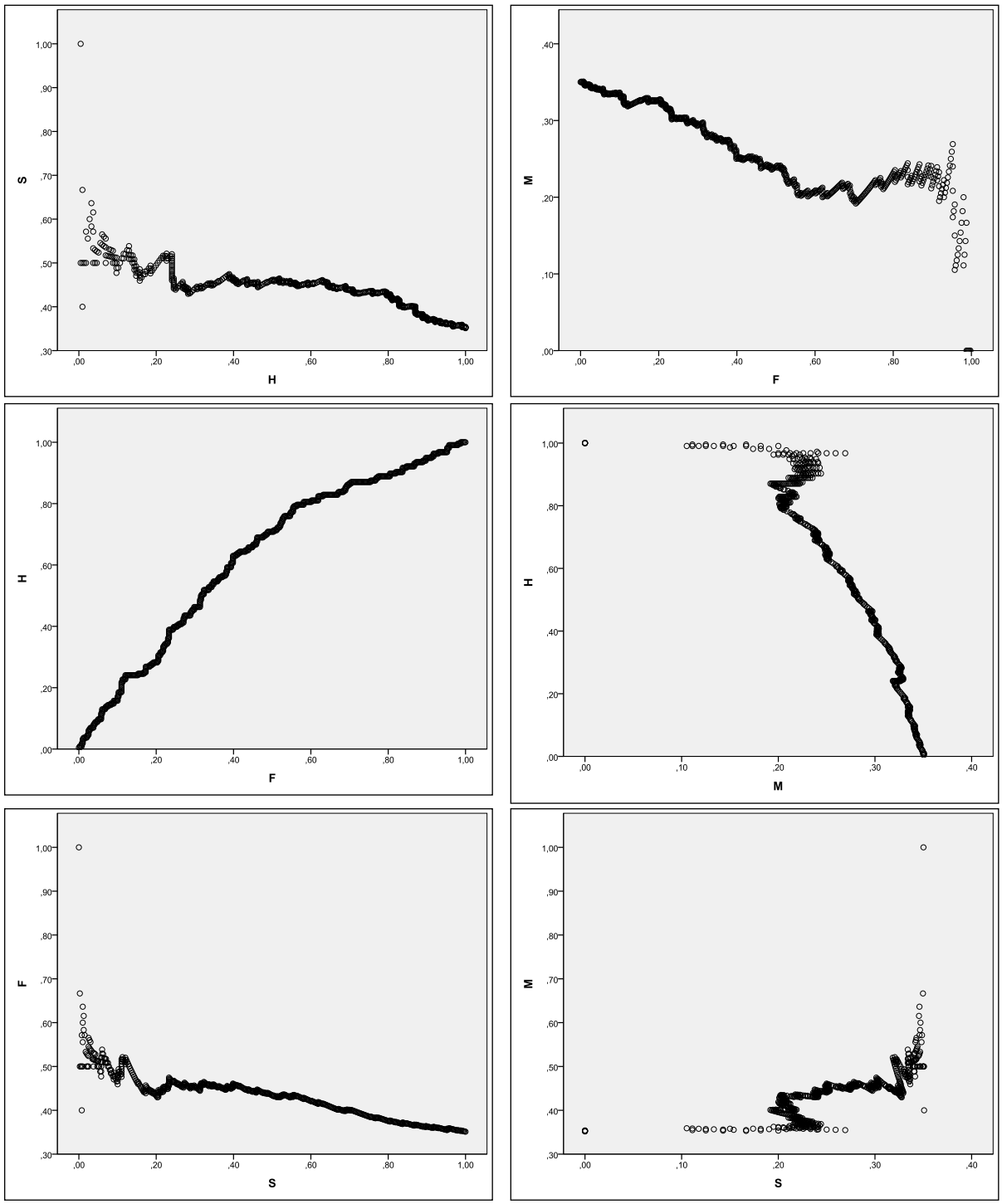

Fig. 1: Scatterplots of interrelations between S, H, M, and F (*n*=615 accepted or rejected, but published elsewhere manuscripts; sorted randomly – by manuscript identifier – in case of equal reviewers' ratings)

The H-F curve is close to a concavely increasing curve as predicted in Egghe (2008). The (almost) concavely decreasing H-M curve was predicted in Egghe (2008) as well while it was mentioned there that the M-F curve, F-S curve and M-S curve are irregular, which is also the case here.

These curves show some imperfections of the journal peer review process – if bibliometric measures are used to assess the quality of manuscripts. Take e.g. the H-F curve. We see that a H of 0.20 (i.e. 20%) yields a F of about 0.10. To require a H of 0.60 lets the F increase to around 0.40. Similarly, for the H-M curve, a H of 0.40 corresponds to a M of around 0.30. Only a slight increase of this number lets a H drop to 0.20. All this is due to the fact that the reviewers' ratings are not the same as the qualification ratings (measured by citation counts), just as in IR where the relevant documents are not always ranked on top in a retrieval ranking.

Conclusions and remarks

Journal peer review can be interpreted as an IR-process where accepted papers play the role of retrieved papers and where qualified papers play the role of relevant papers. Hence the analogue measure of precision, recall, fallout and miss can be defined. They are called here Success rate (S), Hitting rate (H), Fallout rate (F) and Miss rate (M).

On a dataset of Bornmann and Daniel (2010) on 615 submitted papers to AC-IE we can determine the accepted papers as well as the rejected papers but published elsewhere. Note that it is not easy to determine the latter ones since they are only known by the editorial office of the journal. They gave the permission to study the peer review process. It follows the difficult task to determine in both sets the qualified papers (defined as the papers with a normalised citation score above a certain threshold).

The 615 papers are ranked in decreasing order of reviewers' ratings (and randomly – according to manuscript identifier – if reviewers' ratings are the same). For each $n = 1, 2, ..., 615$ we consider the first *n* papers as "accepted" so that for each *n* we can determine S, H, F and M and hence the curves relating any two of these measures. Hereby we quantify imperfections of the journal peer review process, just as in the IR-case where we

quantify the imperfection of the retrieval process. Thus we have information on how much is one measure changing if another one changes with a certain amount.

The curves relating any two of the measures S, H, F and M are not easy to obtain (as described above). In addition to this we underline that, to the best of our knowledge, these curves are produced for the first time (even when interpreted in the IR field itself). This is so because of the following developments: Egghe (2004) re-introduced the measure miss (M) (before that we had to go back more than 25 years to find a few references to M, and they were not in English). Non of these references study curves relating M to any other measure and even after the publication of Egghe (2004) we are not aware that such curves have been constructed (except for the recall-precision curves). The only article dealing with these curves is the theoretical paper of Egghe (2008) where e.g. the shapes of the S-H, H-F and H-M curves are predicted. Fig. 1 is hence the only graph that gives practical evidence for the shapes of these curves (and is interpreted in the journal peer review framework). That we could generate these curves is basically due to the paper of Bornmann and Daniel (2010) who were able to determine the qualifiedness of a set of submitted papers.

Against the backdrop of these developments we encourage researchers to construct other curves relating any two of these four measures, in IR or in journal peer review or in any other field where the IR analogy can be made.

References

L. Bornmann and H.-D. Daniel (2009). The luck of the referee draw: the effect of exchanging reviews. Learned Publishing 22(2), 117-125.

L. Bornmann and H.-D. Daniel (2010). The manuscript reviewing process - empirical research on review requests, review sequences and decision rules in peer review. Library & Information Science Research 32(1), 5-12.

L. Bornmann and L. Egghe (2012). Journal peer review as an information retrieval process. Journal of Documentation, to appear.

L. Egghe (2004). A universal method of information retrieval evaluation: the “missing” link M and the universal IR surface. Information Processing and Management 40(1), 21-30.

L. Egghe (2007). Existence theorem of the quadruple (P,R,F,M): Precision, Recall, Fallout and Miss. Information Processing and Management 43(1), 265-272.

L. Egghe (2008). The measures precision, recall, fallout and miss in function of the number of retrieved documents and their mutual interrelations. Information Processing and Management 44(2), 856-876.

L. Egghe (2010). Study of some Editor-in-Chief decision schemes. Annals of Library and Information Studies 57(3), 184-195.

G. Salton and M.J. McGill (1987). Introduction to modern Information Retrieval. McGraw-Hill, Auckland, New Zealand.

D.M. Schultz (2010). Are three heads better than two? How the number of reviewers and editor behavior affect the rejection rate. Scientometrics 84(2), 277-292.