1 Peer Review and Its Discontents

What Is Peer Review and Is It Any Good?

'Peer review' is the system by which manuscripts and other scholarly objects are vetted for validity, appraised for originality, and selected for publication as articles in academic journals, as academic books ('monographs'), and in different forms.[1] Since an editor of an academic title cannot be expected to be an expert in every single area covered by a publication and since it appears undesirable to have a single person controlling the publication's flow of scientific and humanistic knowledge, there is a need for input from more people. Manuscripts submitted for consideration are shown to external expert advisers ('peers') who deliver verdicts on the novelty of the work, criticisms or praise of the piece, and a judgement of whether or not to proceed to publication. A network of experts with appropriate degrees of knowledge and experience within a field are coordinated to yield a set of checks and balances for the scientific and broader research landscapes. Editors are then bound, with some caveats and to some extent, to respect these external judgements in their own decisions, regardless of how harsh the mythical 'reviewer 2' may be (Bakanic, McPhail, and Simon 1989; Bornmann and Daniel 2008; Fogg and Fiske 1993; Lock 1986; Petty, Fleming, and Fabrigar 1999; Sternberg et al. 1997; Zuckerman and Merton 1971).

The premise behind peer review may appear sound, even incontrovertible. Who could object to the best in the world appraising one another, nobly ensuring the integrity of the world's official research record? Yet, considering the system for even a few moments leads to several questions. What is a peer and who decides? What does it mean when a peer approves somebody else's work? How many peers are required before a manuscript can be properly vetted? What happens if peers disagree with one another? Does (or should) peer review operate in the same fashion in disciplines as distinct as neuroscience and sculpture? Particle physics and social

[1] As we go on to note, this reductive definition of peer review has come under increasing pressure in recent years as organisations have questioned the reliability of using such judgements for the *selection* of material.

geography? Math and literary criticism? When academics rely on publications for their job appointments and promotions, how does peer review interact with other power structures in universities? Do reviewers act with honour and integrity in their judgements within this system?

Consider, as just one example, the question of anonymity in the peer-review process. Review is meant to assess the work itself, not the authors. If the identity of the authors is available to reviewers, though, then might not they give an easy ride to people they know or allow personal disputes to affect their judgement negatively? It is also possible that radically incorrect work might be erroneously perceived as truthful when it comes from an established figure within a discipline or that bold new and correct work might be incorrectly rejected because it comes from new or unusual quarters (Campanario 2009).

Yet simply removing the names of authors is not itself necessarily a solution. When one is dealing with small pools of experts, this can provide a false sense of security (Fisher, Friedman, and Strauss 1994; Godlee, Gale, and Martyn 1998; Sandström 2009; Wang and Sandström 2015). If a reviewer knows that the work was part of a publicised funded project, for instance, it could be possible to guess with some accuracy the authors' identities. In niche sub-fields, researchers usually all know one another and the areas in which their colleagues are working (Mom, Sandström, and Besselaar 2018; Sandström and Hällsten 2008).

On the other side of this process, what about the identity of the reviewer? Should the authors (or even the readership of the final piece) be told who has reviewed the manuscript (Pullum 1984; van Rooyen, Delamothe, and Evans 2010)? There are arguments for both positive and negative answers to this question. When people are anonymous, they may be more able to speak without constraint. A junior postdoctoral researcher may be *capable* of reviewing the work of a senior professor but might not be able to criticise extensively the work for fear of career reprisals were their identity to be revealed (this also raises the question, though, of what we mean by 'peer'). Yet we also know that the cover of anonymity can be abused. Anonymous reviewers, it is assumed, may be more aggressive in their approach and can even write incredibly hurtful *ad hominem* attacks on papers (Silbiger and Stubler 2019).

Further, how can we tell the standards demanded of a publisher without knowledge of the individuals used to assess the manuscripts? As Kathleen Fitzpatrick notes, conditions of anonymity limit our ability to investigate the review process thoroughly. For 'in using a human filtering system', she writes, 'the most important thing to have information about is less the data that is being filtered, than the human filter itself: who is making the decisions, and why' (Fitzpatrick 2011, 38). It is also clear that errors are only likely to be caught if the selection of peers is up to scratch (although as we note later, there are even some problems with this assumption). Group social dynamics may also affect decision-making in this area (Olbrecht and Bornmann 2010; van Arensbergen, van der Weijden, and van den Besselaar 2014). Gender biases also play a role (Besselaar et al. 2018; Biernat, Tocci, and Williams 2012; Helmer et al. 2017; Kaatz, Gutierrez, and Carnes 2014). Anonymity in the review process, just the first of many concerns, is far more complicated than it might at first appear (for more on this debate, see Brown 2003; DeCoursey 2006; Eve 2013; Godlee 2002; Ross-Hellauer 2017; Seeber and Bacchelli 2017; R. Smith 1999; Tattersall 2015; van Rooyen et al. 1999).

It also seems abundantly clear that the peer-review process is far from infallible. Every year, thousands of articles are retracted (withdrawn) for containing inaccuracies (Brainard and You 2018), for conducting unethical research practices, and for many other reasons (for more on this, see the 'Retraction Watch' site; see Brembs, Button, and Munafò 2013 for a study that found that impact factor correlates with retraction rate). On occasion, this has had devastating consequences in spaces such as public health. Andrew Wakefield's notorious retracted paper claiming a link between the mumps, measles, and rubella (MMR) vaccine and the development of autism in children was published in perhaps the most prestigious medical journal in the world, *The Lancet* (Wakefield et al. 1998). The work was undoubtedly subject to stringent pre-publication review and was cleared for publication. Yet the article was later retracted and branded fraudulent, having caused immense and ongoing damage to public health (Godlee, Smith, and Marcovitch 2011). It is, alas, always easier to make an initial statement than subsequently to retract or to correct it. As a result, a worldwide anti-vaccination movement has seized upon this circumstance as evidence of a conspiracy. The logic uses the supposed initial validation of

peer review and the prestige of *The Lancet* as evidence that Wakefield was correct and that he is the victim of a conspiratorial plot to suppress his findings. Hence, when peer review goes wrong, the general *belief* in its efficacy, coupled with the prestige of journals founded on the supposed expertise of peer review, has damaging real-world effects.

In other cases, peer review is problematic for the delays it introduces. Consider research around urgent health emergencies, such as the Zika virus or newly proposed treatments for the 2019 novel coronavirus. Is it appropriate and ethical to wait several days, weeks, or even months for expert opinion on whether this information should be published when people are dying during the lag period? The answer to this depends on the specific circumstances and the outcome, which can only be known after publication. On the one hand, if the information is published, without peer review, and it turns out to be correct and solid without revision, then the checks and balances of peer review would have cost lives. On the other hand, if the information published is wrong or even actively harmful, and there is even a chance that peer review could have caught this, one might feel differently. These are but a few of the problems, dilemmas, and ethical conundrums that circulate around that apparently 'simple' concept of peer review.

In this opening chapter, we describe the broad background histories of peer review and its study. This framing chapter is designed to give the reader the necessary surrounding context to understand the historical evolution and development of peer review. It also introduces much of the secondary criticism of peer review that has emerged in recent years, questioning the usual assumption that the objectivity (or intersubjectivity) of review is universally accepted as the best course of action to ensure standards. Finally, we address the merits of innovative new peer-review practices and disciplinary differences in their take-up (or otherwise). While there are certainly cross-disciplinary implications for our work, it has been in the natural sciences that the benefits and rewards of these new approaches have been most heavily sold.

The Study of Peer Review

Despite the aforementioned challenges, the role of peer review in improving the quality of academic publications and in predicting the impact of manuscripts through criteria of 'excellence' is widely seen as essential to the

research endeavour. As a term that first entered critical and popular discourse around 1960 but also as a practice that only became commonplace far later than most suspect, 'peer review' is sometimes described as the 'gold standard' of quality control, and the majority of researchers consider it crucial to contemporary science (Alberts, Hanson, and Kelner 2008; Baldwin 2017, 2018; Enslin and Hedge 2018; Fyfe et al. 2019; Hames 2007, 2; Moore et al. 2017; Mulligan, Hall, and Raphael 2013, 132; Shatz 2004, 1). Indeed, peer review is much younger than many suspect. In 1936, for instance, Albert Einstein was outraged to learn that his unpublished submission to *Physical Review* had been sent out for review (Baldwin 2018, 542). Yet, despite its relative youth, peer review has nonetheless become a fixture of academic publication. This raises the question, though, of *why* this might be the case. For surprisingly little evidence exists to support the claim that peer review is the *best* way to pre-audit work, leading Michelle Lamont and others to note the importance of ensuring that 'peer review processes . . . [are] themselves subject to further evaluation' (Lamont 2009, 247; see also LaFollette 1992). Indeed, there are long-standing criticisms of the validity of peer review, exemplified in Franz J. Ingelfinger's notorious statement that the process is 'only moderately better than chance' (Ingelfinger 1974, 686; see also Daniel 1993, 4; Rothwell and Martyn 2000) and Drummond Rennie's (then deputy editor of the *Journal of the American Medical Association*) 'if peer review was a drug it would never be allowed onto the market' (cited in Richard Smith 2006a, 2010). However, the status function declaration, as John Searle (2010) puts it, of peer review is to institute a set of institutional practices that allow for the selection of quality by a group of empowered, qualified experts (see also Rachar 2016). Peer review as conducted within universities resonates, in many senses, as a type of 'total institution' as defined by Christie Davies and Erving Goffman: a 'distinctive set of organizations that are both part of and separate from modern societies' (Davies 1989, 77) and a 'social hybrid' that is 'part residential community, part formal organization' (Goffman 1968, 22).

Yet research into peer review processes can be difficult to conduct. At least one of the challenges with such studies is that there is always the risk of seeking explanations for the accepted constructions and logics of peer review, rather than recognising the contingency of their emergence. This

has not, however, prevented a burgeoning field from emerging around the topic (Batagelj, Ferligoj, and Squazzoni 2017; Tennant and Ross-Hellauer 2019). Certainly, following the influential work of John Swales (1990), an ever-increasing number of studies have examined the language and mood of published academic articles, grant proposals, and editorials (Aktas and Cortes 2008; Connor and Mauranen 1999; Giannoni 2008; Harwood 2005a, 2005b; Shehzad 2015; these examples are drawn from Lillis and Curry 2015, 128). This is not surprising; after all, as van den Besselaar, Sandström, and Schiffbaenker note, '[l]anguage embodies normative views about who/where we communicate about, and stereotypes about others are embedded and reproduced in language' (2018, 314; see also Beukeboom and Burgers 2017; Burgers and Beukeboom 2016). Indeed, a number of existing studies have examined the linguistic properties of peer review reports written by the authors themselves (Coniam 2012; Woods 2006, 140–6; for more, see Paltridge 2017, 49–50).

Yet, as examples of some of the difficulties that hinder the study of these documents, consider that peer-review reports are often owned and guarded by organisations that wish to protect not only the anonymity of reviewers but also the systems of review that bring them an operational advantage – anonymous review comes with several benefits for organisations that are in commercial competition with one another. In particular, and as just one instance, since reviewers often work for multiple publication outlets (that is, the same reviewers can review for more than one journal), the claim of one publisher to have a more rigorous standard of review or better quality of reviewer than other outlets could be damaged were review processes open, non-anonymous, and subject to transparent verification. (It is also clear that top presses can publish bad and incorrect work and that excellent work can appear in less prestigious venues (Shatz 2004, 130).) Further, these organisations often do not have conditions in place that will allow research to be conducted upon peer-review reports. The earliest studies of peer review, therefore, generally used survey methodologies rather than directly interrogating the results of the process (Chase 1970; Lindsey and Lindsey 1978; Lodahl and Gordon 1972). As an occluded genre of writing that nonetheless underpins scientific publication, relatively little is known about the ways in which academics write and behave, at scale, in their reviewing practices.

As another example, an absence of rules and guidelines around ownership of peer-review comments certainly contributes to the challenges. The lack of financial incentives in many cases and, therefore, contracts and agreements with reviewers have meant that no clear ownership of reviews has been established (for more on the economics and financials of scholarly communications in the digital era, see Gans 2017; Houghton 2011; Kahin and Varian 2000; McGuigan and Russell 2008; Morrison 2013; Willinsky 2009). In the absence of such statements, it can be assumed that copyright remains with the author of the reports in most jurisdictions. Publishers often do not wish to exert any dominance in this area for fear of dissuading future referees, who can be hard enough to persuade at the best of times. Reviews, therefore, exist in a world of grace and favour rather than one with any clear legal framework. Publishers also benefit from opacity in this domain in other ways. For example, by keeping poor-quality reviews hidden from sight, journals are able to build their reputations on other, less direct, criteria such as citation indices and editorial board celebrity (for more on the occluded nature of peer review, see Gosden 2003). A journal's reviewer database can also provide competitive advantage and bring value to a publishing stable beyond the status of the title itself.

That said, in spite of these difficulties, a substantial number of studies have examined peer review (for just a selection, consider Bonjean and Hullum 1978; Mustaine and Tewksbury 2008; Smigel and Ross 1970; Tewksbury and Mustaine 2012), and it would be a mistake to call the field under-researched, although the methods used are diverse and disaggregated (Grimaldo, Marušić, and Squazzoni 2018; Meruane, González Vergara, and Pina-Stranger 2016, 181). Indeed, Meruane, González Vergara, and Pina-Stranger (2016, 183) provide a good history of the disciplinary specialisations of the study of peer-review processes (PRP) since the 1960s noting that while 'PRP has been a prominent object of study, empirical research on PRP has not been addressed in a comprehensive way.' The precise volume of research varies by the way that one searches, but there appears to be up to 23,000 articles on the topic between 1969 and 2013 by one count – and this does not even include the so-called grey literature of blogs (Meruane, González Vergara, and Pina-Stranger 2016, 181). It is, then, well beyond the scope of this book to provide a comprehensive meta-review of the

secondary literature on peer review. The interested reader, though, could consult one of the many other studies that have conducted such an exercise (Bornmann 2011b; Meruane, González Vergara, and Pina-Stranger 2016; Silbiger and Stubler 2019; Weller 2001). Much, although by no means all, of this research has been critical of peer-review processes, finding our faith in the practice to be misplaced (Squazzoni 2010, 19; Sugimoto and Cronin 2013, 851–2; for more positive opinions on the process, see Bornmann 2011a; Goodman 1994; Pierie, Walvoort, and Overbeke 1996). Critics of peer review usually point to its poor reliability and lack of predictive validity (Fang, Bowen, and Casadevall 2016; Fang and Casadevall 2011; Herron 2012; Kravitz et al. 2010; Mahoney 1977; Schroter et al. 2004; Richard Smith 2006b); biases and subversion within the process (Bardy 1998; Budden et al. 2008; Ceci and Peters 1982; Chawla 2019; Chubin and Hackett 1990; Cronin 2009; Dall'Aglio 2006; K. Dickersin et al. 1987; Kay Dickersin, Min, and Meinert 1992; Ernst and Kienbacher 1991; Fanelli 2010, 2011; Gillespie, Chubin, and Kurzon 1985; Ioannidis 1998; Link 1998; Lloyd 1990; Mahoney 1977; Ross et al. 2006; Shatz 2004, 35–73; Travis and Collins 1991; Tregenza 2002); the inefficiency of the system (Ross-Hellauer 2017, 4–5); and the personally damaging nature of the process (Bornmann 2011b, 204; Chubin and Hackett 1990). For instance, and as just a sample, when Rothwell and Martyn (2000, 1964) studied the reproducibility of peer-review reports, they repeated Ingelfinger's assertion that 'although recommendations made by reviewers have considerable influence on the fate of both papers submitted to journals and abstracts submitted to conferences, agreement between reviewers in clinical neuroscience was little greater than would be expected by chance alone.' As another example, more recent work by Okike et al. (2016) unveiled a strong unconscious bias among reviewers in favour of known or famous authors and institutions in the discipline of orthopaedics when using a single-blind mode of review. This casts serious doubt on claims to be able to 'put aside' one's knowledge of an author or to act objectively in the face of conflicts of interest, although in other disciplinary spaces, it has been argued that the definition of merit, as defined in a discipline, is constructed by particular figures and that this identity *should* play a part in the evaluation of their work (Fish 1988). Additionally, Murray et al. (2018, 25) explore the relationship between gender and international

diversity and equity in peer review, concluding that '[i]ncreasing gender and international representation among scientific gatekeepers may improve fairness and equity in peer review outcomes, and accelerate scientific progress.'

Special mention should also be made of the PEERE Consortium, which has, in particular, achieved a great deal in opening up datasets for study (PEERE Consortium n.d.). Aiming, with substantial European research funding, to 'improve [the] efficiency, transparency and accountability of peer review through a trans-disciplinary, cross-sectorial collaboration', the consortium has been one of the most prolific centres for research into peer review in the past half decade. Publications from the group have spanned the author perspective on peer review (Drvenica et al. 2019), the reward systems of peer review (Zaharie and Seeber 2018), the links between reputation and peer review (Grimaldo, Paolucci, and Sabater-Mir 2018), the role that artificial intelligence might play in future structures of review (Mrowinski et al. 2017), the timescales involved in review (Huisman and Smits 2017; Mrowinski et al. 2016), the reasons why people cite retracted papers (Bar-Ilan and Halevi 2017), the fate of rejected manuscripts (Casnici, Grimaldo, Gilbert, Dondio et al. 2017), and the ways in which referees act in multidisciplinary contexts (Casnici, Grimaldo, Gilbert, and Squazzoni 2017).

In terms of the language used in peer review reports, work by Brian Paltridge (2015) has examined the ways in which reviewers request revisions of authors, using a mixed-methods approach. Paltridge studied review reports for the peer-reviewed journal *English for Specific Purposes* finding a mixture of implicit and explicit directions for revision used by reviewers, making for a confusing environment in which 'what might seem like a suggestion is not at all a suggestion' but 'rather, a direction' (Patridge 2015, 14), a view echoed by Gosden (2001, 16). Somewhat in contrast to this, though, Kourilova (1998) found that non-native users of English often wrote with an honest, or brutal, bluntness in their reports for a range of sociocultural reasons. While this may come with its own challenges and be painful for authors, such bluntness is far less subject to misinterpretation than hedged attempts at avoiding offence. Comments with such a negative tone can also appear in published book reviews (Salager-Meyer 2001), but

these are often more complimentary than behind-the-scenes peer-review reports (Hyland 2004, 41–62). The 'out of sight' or occluded nature of peer review (Swales and Feak 2000, 229) tends, then, to lend itself to more critical judgements.

New Modalities of Peer Review

For some commentators, peer review is the least bad option, and any intervention is only likely to result in poorer outcomes. Others, though, have proposed a range of new methods and operational modes that are designed to address the perceived shortcomings in existing review protocols. Certainly, for the most part, this has taken place within the natural sciences and, as David Shatz remarks, 'the paucity of humanities literature on peer review . . . [is] truly striking', although he is unable to explain why this should be (2004, 4). In this section of this chapter, we outline these new experiments, mostly within this disciplinary space.

Post-Publication Review

The twenty-first century is characterised by what Clay Shirky has famously called 'filter failure' (2008). In the face of an ever-increasing abundance of material – be that scholarly, information, or news – it is apparent that we face severe difficulties in knowing where to spend our scarce attention (Bhaskar 2016; Eve 2017; Nielsen 2011). Various solutions have been posed for how this might be remedied, most of which centre on what Michael Bhaskar calls a culture of 'curation', in which whether by algorithm or by human selection, the 'wheat is sorted from the chaff' (Bhaskar 2016). What it might mean to do so appropriately is, of course, a matter of some contention. Algorithms that surface only mainstream content when we are looking for outliers represent just another problematic case of filter failure, as opposed to any viable solution.

The same problems apply to the scholarly and research literature (Eve 2020). We exist in a world where more is published than it is possible for a person to read, even in almost every niche sub-discipline. Some solutions have taken the algorithmically curated route. Hence, there have been, on the reader side, several attempts to provide automatic summaries of articles, condensing these otherwise large artefacts down into bite-size, digestible