Article

S|S|S

Social Studies of Science
2024, Vol. 54(2) 184–209

DOI: 10.1177/03063127231200274
journals.sagepub.com/home/sss
Sage

Checking correctness in mathematical peer review

Christian Greiffenhagen[1,2]

Abstract

Mathematics is often treated as different from other disciplines, since arguments in the field rely on deductive proof rather than empirical evidence as in the natural sciences. A mathematical paper can therefore, at least in principle, be replicated simply by reading it. While this distinction is sometimes taken as the basis to claim that the results in mathematics are therefore certain, mathematicians themselves know that the published literature contains many mistakes. Reading a proof is not easy, and checking whether an argument constitutes a proof is surprisingly difficult. This article uses peer review of submissions to mathematics journals as a site where referees are explicitly concerned with checking whether a paper is correct and therefore could be published. Drawing on 95 qualitative interviews with mathematics journal editors, as well as a collection of more than 100 referee reports and other correspondence from peer review processes, this article establishes that while mathematicians acknowledge that peer review does not guarantee correctness, they still value it. For mathematicians, peer review 'adds a bit of certainty', especially in contrast to papers only submitted to preprint servers such as arXiv. Furthermore, during peer review there can be disagreements not just regarding the importance of a result, but also whether a particular argument constitutes a proof or not (in particular, whether there are substantial gaps in the proof). Finally, the mathematical community is seen as important when it comes to accepting arguments as proofs and assigning certainty to results. Publishing an argument in a peer-reviewed journal is often only the first step in having a result accepted. Results get accepted if they stand the test of time and are used by other mathematicians.

Keywords

certainty, error, mathematics, proofs, peer review, replication, scientific community

[1]The Hong Kong Polytechnic University, Hong Kong, China
[2]Department of Economics and Social Sciences, Télécom ParisTech, France

Correspondence to:
Christian Greiffenhagen, Department of Applied Social Sciences, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong SAR, China.
Email: christian.greiffenhagen@polyu.edu.hk

Mathematics is often seen as different from other academic disciplines: While the natural sciences rely on empirical evidence to back up their claims, mathematics depends on arguments in the form of written proofs (Barany & MacKenzie, 2014, p. 121; MacKenzie, 2001, p. 2; Singh, 2011, p. xii). A mathematical paper is therefore 'self-contained', and it is possible, at least theoretically, to replicate the results of a paper simply by reading a paper rather than having to go the laboratory and reproduce the experiments (see Peterson & Panofsky, 2021). This has sometimes promoted an idealized image of mathematics as the paradigm of (absolute) certainty (Russell, 1956, p. 4; Shapin, 1994, p. 322; Wiener, 1915, p. 568). It is widely acknowledged, however, that the mathematical literature is far from certain; rather, it is full of mistakes that range from simple typos to problems that invalidate the proof. In a frequently quoted remark, Davis (1972, p. 262), describes the situation thus: 'A past editor of the *Mathematical Reviews* once told me—somewhat in jest—that 50% of all mathematics papers printed are flawed.' While this statistic may be an exaggeration, early in the twentieth century, Lecat (1935) identified 500 published errors made by 330 mathematicians. There are also many examples of famous mistakes, such as Alfred Kempe's 1879 supposed proof of the four-colour conjecture (see MacKenzie, 1999, p. 18), or Hans Rademacher's argument apparently disproving the Riemann hypothesis (Davis & Hersh, 1981, p. 56; Krantz, 2011, p. ix). Of course, many mistakes may never be discovered, instead staying hidden (Zahler, 1976, p. 98).

Thus, while pure mathematics does rely on proof rather than empirical evidence, this distinction does not guarantee that what gets published in mathematical books and journals is always correct. This assertion would only be true if reading—and checking—proofs were easy. But mathematical proofs cannot be simply 'sight-read' (Davis & Hersh, 1981, p. 281); understanding them requires a lot of work (Livingston, 1986, 1999), and finding mistakes can be even more challenging. On this point, Lakatos (1976) criticizes Quine (1951, p. 87) who claims that 'the mathematician hits upon his proof by unregimented insight and good fortune, but afterwards other mathematicians can check his proof'. In contrast, Lakatos (1976, p. 4) argues that checking whether a written argument actually *is* a proof, or if it possibly contains a mistake, may be as difficult and challenging as finding the argument in the first place: 'But often the checking of an ordinary … proof is a very delicate enterprise, and to hit on a "mistake" requires as much insight and luck as to hit on a proof: the discovery of "mistakes" in informal proofs may sometimes take decades—if not centuries.'

In this article I ask how correctness in mathematics is adjudicated in practice. I examine the peer review process through 95 qualitative interviews with editors of journals, as well as a collection of more than 100 referee reports and other correspondence from the peer review process. Refereeing a paper's correctness is one key consideration in deciding whether or not a paper should be published— the other major consideration being the question of how significant or important the results reported in a paper are, which is dealt with in a separate article (see Greiffenhagen, forthcoming). In mathematics, a referee who is checking a paper can be seen as doing what Peterson and Panofsky (2021) call 'diagnostic' replication, which consists of 'evaluating the truth of a claim' (p. 586).

Checking a proof as part of peer review is only one form of reading mathematics. In Hirschauer's (2010, p. 77) terminology, it is an 'appraisal-oriented' reading, which he contrasts with a 'researcher's reading' that is 'oriented towards further use'. Barany and

MacKenzie (2014, p. 121) have described the researcher's reading approach for mathematicians, highlighting that it does not consist in reading a paper from start to end, but rather in identifying key passages in the hope that they contain ideas that help to advance the researcher's own 'instrumental research goals' (see Mejia-Ramos & Weber, 2014, p. 169). Both of these readings are different from what we might call reading for educational purposes, such as while studying for an undergraduate or graduate degree, which involves not so much doing cutting-edge research, but understanding established theorems and proofs (Greiffenhagen, 2014; Livingston, 1986).

Adopting an ethnomethodological perspective (Garfinkel, 1986; Watson, 2016), I treat checking correctness within peer review as a practical accomplishment. My aim is to contribute to sociological and historical studies of mathematics that illustrate the informal, social, and practical aspects of mathematical practice. For example, Steingart (2012), through a historical study of the classification of finite simple groups, points to the importance of face-to-face interaction in the production and communication of mathematical knowledge, showing that 'much of the work in finite simple group theory circulated via personal, often informal, communication, rather than in published proofs' (p. 185). Greiffenhagen (2014) positions the presentation of mathematical knowledge in graduate lectures as crucially dependent on material resources, in particular, large writing surfaces, such as blackboards. Barany and MacKenzie (2014) investigate research seminars and highlight that most of the mathematicians in their audiences were not interested in all the details of the proof; but rather were engaged with the materials on a more abstract level—that is, 'the talk's conceptual narrative and technical and heuristic manipulations' (p. 110). In their investigation of supervision meetings, Greiffenhagen and Sharrock (2011) show that this kind of talk was used by a supervisor when suggesting strategies for coming up with novel proofs as part of a student's doctoral thesis. By studying controversies that arose after the publication of results (and therefore after the peer review described here), MacKenzie (1999, 2001) and Rosental (2003, 2008) both discuss the importance of the community in the acceptance of particular arguments. In a famous study, Lakatos (1976) notes that the definitions of concepts (such as 'polyhedron') are not fixed but flexible, and MacKenzie through a historical study of the debates concerning the use of computers as part of a proof of the four-colour conjecture demonstrates that the very notion of what counts as a mathematical proof can be open for debate, thereby revealing 'the negotiability of "proof" itself' (MacKenzie, 1999, p. 7). Finally, Rosental (2008), in a study on the reception of a theorem about fuzzy logic, shows that the question for the community was not so much whether the theorem was correct, but what the theorem actually demonstrated.

In this article I show, first, that editors regard refereeing in mathematics as more difficult—and indeed as qualitatively different—than in other disciplines. Since a paper does not report on experiments or data gathered outside the paper, but is in a certain sense self-contained, a referee can thus be said to 'replicate' the paper. For my interviewees, this process makes checking correctness more tedious and time-consuming than evaluating soundness in other disciplines. This may also account for the fact that there is often only one referee. Second, almost every interviewee agrees that the primary responsibility for correctness lies with the author. However, this does not mean that referees are not also held accountable, since they are expected to ensure that a paper is 'likely' to be correct. Third, given that refereeing in mathematics is typically single-blind, referees will take the authors' reputations into account when checking for correctness. However, referees do so in nuanced ways, as authors only have a 'track record' in certain areas and can also gain a negative reputation if they frequently publish sloppy or even faulty papers. Fourth, the actual checking of a paper is typically not done by reading the paper line-by-line. Instead, referees often know where the 'key places' are that they will have to check in detail. Furthermore, referees may zoom in on parts in the paper that look 'suspicious' to them. Fifth, while referees only rarely find straightforward mistakes, they more frequently encounter parts in the paper that they do not understand. They may find that there are certain 'gaps' in the proof that make it difficult or even impossible for them to follow the argument. Authors will then try to fill these gaps, often successfully, but occasionally this process leads to authors withdrawing the paper. Sixth, most of my interviewees accept that it is impossible to avoid mistakes, but they believe that the community will eventually discover any mistakes in important papers. Finally, despite all its acknowledged problems, my interviewees consider peer review important for mathematics and much prefer it to publishing papers on preprint servers only, since peer review, as one referee put it, 'adds a bit of certainty'.

Mathematical proofs

A mathematical proof is often conceptualized as a series of steps, each following logically from the previous ones. According to this picture, a proof is like a written calculation, where it is easy to check how the flow from one line to the next. It should thus be simple to see whether there are mistakes in a proof.

The reality is rather different. Proofs in the real world—those that are published in mathematical journals such as *Annals of Mathematics* or the *Duke Mathematical Journal*—have a very different character. They are written not as long calculations, but as rhetorical arguments to be understood by other mathematicians (Thurston, 1994, p. 162). In this sense, they can be thought of as a 'message', a communication among mathematicians (De Millo et al., 1979, p. 273; Tymoczko, 1984, p. 465). Mathematicians must therefore decide how best to write proofs so that they are easily understood—and accepted—by other mathematicians. Perhaps surprisingly, this goal means that not every detail is written down, because doing so would make proofs so long that it would be difficult for them to be taken in by other mathematicians (Krantz, 2011, pp. 30–31). In other words, when writing a proof for a journal, choosing not to write down every step is not an oversight; it is a necessary and accepted part of communicating the ideas contained in the paper (Fallis, 2003, p. 55).

When writing a proof, authors may thus use expressions like 'it follows that', 'it is easy to see that', or 'it is obvious that', which rely on an existing shared understanding of what is or is not clear based on what Livingston (1986, p. 14) has termed 'the associated practices of its proving'. This is something that novices, such as undergraduate or graduate students, struggle with. It is also the basis of jokes among mathematicians, such as this one by Buzzard (2020, p. 1792): 'One is reminded of the apocryphal story of a student asking their professor whether the fact just presented to the class as "obvious" was indeed obvious, and the professor going into deep thought to emerge 20 minutes later with the reply “yes”.’ However, it is a problem not just for novices, but for professional mathematicians as well. What may be obvious for one mathematician, may not be so obvious for another, which is nicely captured in a remark by the mathematician Boas (1995, p. 125): ‘If Bohnenblust says it’s obvious, it’s obvious. If Bochner says it’s obvious, you can figure it out in half an hour. If von Neumann says it’s obvious, you can prove it in three months if you’re a genius.’

In order to understand a mathematical proof, it is not enough to read it from beginning to end, line by line (Livingston, 1986, 1999). To follow the reasoning expressed in the paper, readers will have to think and reflect, and different readers will find different parts of the paper easy or difficult. In other words, understanding a paper requires *work* (Davis & Hersh, 1981, p. 281) and can take days, weeks, or even months. For example, Singh (2011, p. 278) describes how understanding Wiles’s proof of Fermat’s last theorem took two referees a whole summer of work, including daily emails to Wiles with clarifying questions.

While it is widely accepted that in writing a proof one should omit routine steps (AMS, 1962, p. 430), deciding what counts as ‘routine’ is not always straightforward (Harrison, 2008, p. 1399). On the one hand, in order to communicate with other mathematicians, proofs typically omit many details and can therefore be said to permit certain gaps; on the other hand, it may be precisely in these gaps that a significant error may be hiding (Nathanson, 2009, p. 8). This situation has led some mathematicians (e.g. Hales, 2008; Voevodsky, 2004) to argue that ‘formal’ proofs should replace traditional proofs. Formal proofs are written to be verified by a computer, not read by other mathematicians; they are therefore ‘unreadable’ (De Millo et al., 1979, p. 275) and ‘for the most part humanly incomprehensible’ (Avigad & Harrison, 2014, p. 73). Although formal proofs have been used as part of traditional proofs, most famously in the case of the proof of the four-colour conjecture (MacKenzie, 1999), and some researchers have called for greater use of formal proofs (Buzzard, 2020), computer verification has not become standard everyday mathematical practice for every published proof (MacKenzie, 2001, p. 316). This paper is therefore concerned with traditional proofs published in mathematical journals.

Peer review in mathematics

In mathematical parlance, peer review is referred to as ‘refereeing’ and is done by a ‘referee’, while ‘reviewing’ consists of writing summaries of already published articles for journals such as *Mathematical Reviews*, which is much more important in mathematics than in other disciplines (see Barany, 2021). Like many of the natural sciences, peer review is typically single-blind, but not double-blind: Referees know the names of the authors of the paper under review, but the authors do not know the identity of their referees. In the 1970s, the American Mathematical Society (AMS) conducted an experiment to introduce double-blind refereeing, which aroused passionate objections, including arguments that this would make it impossible to referee papers. At the time, the proposal was not adopted (Pitcher, 1988, pp. 15–17). However, almost fifty years later, the AMS made another attempt and launched double-blind review in two of its journals, *Proceedings of the American Mathematical Society* and *Representation Theory*

(American Mathematical Society [AMS], 2022). Often papers in mathematics are distributed prior to peer review on the online pre-print server arXiv (Gunnarsdottir, 2005; Jackson, 2002), and many mathematicians receive a daily news alert for papers in the subject areas that they are interested in.

Peer review involves two key judgements: first, the importance or significance of the results; second, the soundness of the methods employed or, in the case of mathematics, the correctness of the proofs (Polanyi, 1951, p. 53). In the most prestigious mathematics journals, these two judgements are often split into two steps (see Greiffenhagen, forthcoming). Editors start by soliciting several 'quick opinions' about the importance of the results. Only after a positive evaluation do they ask a referee to check their correctness (Crowley et al., 2011, p. 1128).

Peer review in mathematics has rarely been studied, although mathematicians have written about peer review in autobiographies, guidebooks, and professional journals (e.g. Crowley et al., 2011; Krantz, 2005; Pollatsek, 2018). There have been a few small-scale studies, primarily from the perspective of the philosophy of mathematics and educational studies of mathematics. Geist et al. (2010) conducted a survey of 13 editors about the referee process, revealing only half of the editors expected referees to check the correctness of all the details of a proof. Furthermore, about half of the editors thought that referees did these checks well and did this better at the 'top-level' journals than at the 'mid-level' journals. Mejia-Ramos and Weber (2014) surveyed 54 mathematicians about how they refereed proofs and found a variety of refereeing behaviours and expectations with discrepancies in whether they check every detail of a proof or trust certain authors, for example. Andersen (2017, 2020), using qualitative interviews with seven mathematicians, found that referees typically do not check every detail and do take into account the reputation of the author. Referees often compare the results of the proof against what they know to see whether the results are reasonable or not and typically only check in detail results that look surprising or suspicious. Finally, Despeaux (2011) carried out an analysis of referee reports of mathematical papers submitted to the Royal Society from 1832 to 1900, finding that referees made judgements about the importance of mathematics. In making those judgements, referees sometimes took extra-mathematical aspects into account, for example, they may have decided to publish 'a piece of "bad" (or at least not good) mathematics, because of the status of the author' (p. 14).

This project is based on qualitative interviews about the peer review process conducted with editors of mathematical journals, following the methodology of Lamont's (2009) study of panellists of multidisciplinary fellowship competitions in the humanities and social sciences. I focussed on editors for several reasons. First, editors have, in contrast to referees, a broader understanding of how the peer review process is organized at a given journal, specifically regarding how manuscripts are handled. Second, editors could tell me how referees are assigned. Third, since it is ultimately editors who alone or as members of an editorial board make decisions about a manuscript, they are in the best position to tell me about the criteria for evaluation. In the interviews, I asked editors to take me through the journey of a manuscript at their journal. My questions considered four topics: the organization of peer review, the checking of correctness, the judgements involved in peer review, and general views about peer review.

Overall, I conducted 95 interviews for this project, of which 87 were with editors of journals and 8 were with people working for publishers of journals. More than half of the interviews (49) were done as physical meetings as part of several fieldwork trips to Europe, the US, and Hong Kong. The rest of the interviews (46) were carried out through video calls or over the phone. The editors with whom I talked to were working mainly in North America (39) and Europe (37), some were based in Asia (9), and two were working in Africa (1) and Australia (1).

Editors talked about generalist journals in mathematics (such as *Annals of Mathematics, Mathematische Zeitschrift*, or *Proceedings of the American Mathematical Society*) and many different specialist journals in pure mathematics (such as *Journal of Functional Analysis, Journal of Geometry, Journal of Group Theory, Journal of Lie Theory, Journal of Number Theory, Journal of Symbolic Logic*, or *Journal of Topology*). Many of my interviewees were editors of more than one journal, and different interviewees were sometimes editors of the same journal (sometimes from different time periods). To get a sense of the distribution of the journals that editors talked about, I used the ranking of the Australian Research Council (n.d.): A* (= 'one of the best in its field'), A (= 'very high quality'), B (= 'solid, though not outstanding'), and C (= 'does not meet the criteria of the higher tiers'). A* journals were mentioned 36 times, A journals 42 times, B journals 19 times, and C journals 8 times, while editors also talked about 6 journals that were not listed in the rankings. As a consequence, there was an emphasis on 'higher-ranking' (A* or A) journals rather than 'lower-ranking' (B or C) journals in my interviews.

After each interview, I asked my interviewees whether they would be willing to share any materials (emails, quick opinions, technical reports) to which they had access to in their functions as editor, reviewer, or author. Although I was not able to access the archives of any particular journal, I was able to assemble a variety of primary materials. In total, I collected 120 quick opinions, 100 technical reports, and 50 emails (e.g. asking for quick opinions or accepting/rejecting a paper).

Analysis

The special character of refereeing in mathematics

Given what I have already outlined, it is unsurprising that all of my interviewees emphasized that checking the correctness of the proofs is extremely difficult and slow. They stated that 'refereeing a paper in math is usually a lot of work; sometimes it's an incredibly large amount of work' and they saw it as sometimes 'an unbelievably tedious process'. Indeed, one interviewee claimed that refereeing a paper is 'almost as hard as writing your own paper if you do it well'.

The difficulty of checking a paper can, of course, sometimes be seen as a result of bad writing. However, my interviewees emphasized that checking a mathematical paper was inherently difficult. For example, one referee, after spending a long time refereeing a paper, wrote in the report:

> Over several months, I was reading through 53 pages of condensed (but surprisingly well-written text) and can say with confidence that … [t]his result is important and deep. The proofs

> are correct. … The proof, and even the formulation of the result, is difficult. Even the formulation of the principal result, when written in detail, takes two pages—please see below. But this reflects the real difficulty of the matter.

Here, the difficulty of reading through the paper is not seen as something that could be solved through better writing. Instead, it is seen as in the nature of the proof and the result.

One of the consequences of the amount of time that refereeing takes has been a division of labour at 'elite' journals, where often senior mathematicians are asked to write quick opinions judging the significance of a paper and younger mathematicians are asked to do the technical refereeing of checking the correctness of the results. One of my interviewees said: 'Typically you ask a big shot to comment on the importance of the paper and you ask a lesser known but very competent mathematician to check all the details.' The presumption behind this distinction is that more senior scholars may not be willing to take the time to read a paper carefully:

> You use young people as referees because they're still energetic and willing to actually read the paper that's written. Older people will have more of a tendency, if they look at a lemma and they can see how to prove it, they won't read the proof in the paper; they'll just move on. I don't know, because they're tired or whatever. So younger people, in my experience, they tend to write much more thorough referee reports.

Indeed, one senior (and very famous) interviewee openly admitted: 'I don't referee for anybody, on the grounds that I'm too busy, but I'll give a quick opinion.'

When asked how long they would spend on refereeing a paper, my interviewees replied that on average it would take them roughly eight to ten hours, or one or two working days. However, many also recounted instances where they had spent much longer, stating for example, 'I've spent 50 hours on a paper, that's not unusual', or 'I remember a period of maybe six weeks in which I did nothing else but referee a paper.'

Refereeing in mathematics thus seems to take much longer than in other disciplines, where reported times range between two and six hours (e.g. Armstrong, 1997, p. 65; Weller, 2001, p. 157). This can also be seen in the time that typically elapses between submission to a preprint server such as arXiv and the journal publication, which is very different between disciplines. As Larivière et al. (2014, p. 1163) have shown, this delay is shortest in physics (less than half a year) but longest in mathematics (more than a year).

Indeed, some of my interviewees argued that other disciplines do not have the equivalent of *technical* refereeing in mathematics:

> The trouble is, as you're certainly aware, refereeing in mathematics is quite a bit different than refereeing in the other disciplines. The other disciplines don't really have the analogue of a *technical* referee, you don't go into someone's lab and verify all of their techniques. So the technical refereeing can be a *huge* job.

The interviewee here makes an explicit contrast between peer review in mathematics and peer review in other disciplines. In the experimental sciences, a paper can be seen as a report of data gathered through experiments outside the paper. If a referee really wanted to check all the details, they would have to visit the original laboratory and perhaps redo the experiments. However, the situation is remarkably different in mathematics. Papers in mathematics are therefore often treated as ‘self-standing’ (Macbeth, 2021, p. 13), since, as Barany and MacKenzie (2014, p. 121) put it, they ‘are seen in principle to contain all the apparatus required for their verification’. Therefore, it should—at least in theory—be possible to follow the argument from start to finish. One my interviewees noted:

> In theory, a mathematics paper should include everything that’s needed to understand the conclusion start to finish, whereas in say the lab sciences, a paper can tell you what problem they were trying to solve, what experimental set up they had, what the data looked like, how they eliminated certain kinds of error and then, to some extent, you just have to trust them that their apparatus actually works the way they said it did.

Of course, a mathematical paper is not really self-contained, since it relies on results and techniques from other papers. As one of my interviewees acknowledged, ‘The paper is not always completely self-contained. Sooner or later the author will write “it is well known”, or they will refer to another paper.’ Put differently, a mathematical paper cannot be read by anyone; it is only accessible to those with the necessary expertise and training. It is therefore very important for editors to find the right referee for a paper under review. Sometimes there are only a few people in the world who have the necessary expertise to understand the details in a paper. One of my interviewees claimed that in some cases ‘maybe five people in the world can really read through all the details of the paper’.

In the experimental sciences, referees have to take a lot on trust, something that Shapin (1984) has referred to as ‘virtual witnessing’. In mathematics, papers are written to allow what we might call ‘direct witnessing’. The reader of a mathematical paper, given sufficient time, has all the resources necessary to verify the proofs and theorems in the paper themselves (see the next sections on how, in practice, referees in mathematics also trust, for example, the expertise of authors). In that sense, a referee replicates or reproduces the paper (see Hagstrom, 1965, p. 74). As my interviewees argued, ‘there is a sense in which if a referee is checking the mathematical details in earnest, that’s like repeating the experiments that are done in experimental sciences’, or ‘I think publishing in mathematics is perhaps not the same as it is in some other fields. In a way, the reading of the paper is a form of reproducing the results.’ That said, a reader is replicating the reasoning of the *discovered* proof, not the work of *discovering* the proof in the first place.

These features of mathematics may perhaps explain why, in most cases, editors only ask for *one* report (Crowley et al., 2011, p. 1128; Krantz, 2005, p. 14). As explained by one of my interviewees: ‘Mathematics is only one report because it takes too much effort to read a paper and write a report. It’s a waste of time to duplicate these efforts.’ Indeed, some editors explicitly questioned whether it is reasonable to ask for more than one report: ‘So there are some journals where the editors require two reports. I don’t like to do that because refereeing a paper is a burdensome task and there are a lot of my colleagues around the world who get more than their fair share of papers to referee.’ Only in the cases of especially noteworthy results or at ‘top’ journals do editors ask for two or more reports. However, if a paper relies on methods from different sub-fields, editors may ask for referees from each sub-field.

Checking correctness, in contrast to judging significance, is seen as having an objective character:

> You can be far more definite about whether something is right or not. There's still scope for judgement as to whether something is interesting or important or significant, rather than boring or whatever, but it still feels much more resolute than social sciences.

Since technical refereeing is considered objective, it may also be acceptable to have technical referees who are close to the author, or possibly even suggested by the author, which would not be acceptable for quick opinions or in many other disciplines:

> It may also happen, if the quick opinion is really positive, then I can go back to the author and say, 'I've had a strongly positive quick opinion, now I just need to make sure that the article is correct. Do *you* know someone who has read your article very carefully and can promise that it's correct?' … Let's say, if your colleague down the hall is asked to comment on how important your work is, that's not so good, but whether your work is correct is an objective or very close to objective judgment.

This perception also explains why in mathematics contact between referees and authors may be acceptable during the technical referee process.

Responsibility for correctness

All of my interviewees agreed that the primary responsibility for correctness falls on the author: 'In the end it's the author's responsibility to make it correct. If there's a mistake, it's not the referee's fault.' This understanding is also confirmed in official guidelines of mathematical societies (AMS, 2015; European Mathematical Society, 2012, p. 4). As one of my interviewees expressed: 'It cannot be anybody else because the author is the one who by several orders of magnitude knows better the substance of the papers than anybody else.'

For Burnham (1992, p. 55), the greatest change in nineteenth-century scientific publishing was not the introduction of peer review, but the fact that journals no longer published many anonymous contributions. Instead, journals started 'establishing responsibility of individual authors for their statements' (see also Biagioli, 2003, p. 256). Mathematics is a very clear example of this shift.

Interestingly, a few of my interviewees mentioned that younger mathematicians may not be completely aware of this, or may not want to take on this responsibility as an author. That is to say, they may want to use referees as 'an error-catching service' (Halmos, 1985, p. 119) in cases where they are not completely convinced of their own results. This behaviour is something that my interviewees found unacceptable:

> I've actually heard some junior people say: 'I have this great result and I can use it for this other thing, but I'm a little worried about doing that until the journal accepts it so we know it's okay'. My view is: 'It's *your* result, if you're not willing to vouch for it, who else can?'

It was more challenging for my interviewees to express what they expected from referees with respect to checking correctness. Editors frequently replied in terms of what they hoped ‘ideally’, and what they accepted ‘in practice’ (see also Krantz, 2003, p. 679):

> *Ideally* checking every line, but *in practice* I think there’s going to be a lot of, ‘okay I understand what technique they’re applying to this step of the problem, and I know from experience that this is a standard thing that’s going to work.’

Official guidelines state that ‘the referee must be reasonably convinced of the accuracy of the paper’ (AMS, 2015, p. 1), which was reflected in how editors talked about their expectations from referees. Editors wanted them ‘to convince themselves it’s most likely to be true’, so that they ‘are pretty confident that is really right’. In that sense, referees do a ‘plausibility check’, ‘with the hope to catch mistakes’.

This mentality is reflected in what referees tend to write in their technical reports. They sometimes express complete confidence in the results, making statements such as: ‘I have read the paper and I can certify that, indeed, everything is correct.’ Alternatively, they may say that they did not note any errors: ‘I could follow the argument in detail and did not find any mistakes.’ Sometimes they use weaker formulations such as ‘The proof is correct I think’, or ‘I checked the proof and the results seem correct.’ On occasion, they may even admit that they did not check every step: ‘I did not check everything in detail, but I am quite comfortable with the overall structure of the arguments.’

The editors acknowledged that there was huge variety in the amount of work that referees put into checking correctness (see Auslander, 2008, p. 65). While some reports may give the editor a lot of confidence that the referee put serious effort into checking the result, other reports were ‘too sketchy’ or only contained remarks ‘for the first half of the paper’. This discrepancy puts editors in a difficult position. On occasion, they ask for a second report. However, since they have already been holding the paper for a long time, they may accept it on the basis of what they deem an unsatisfactory report: ‘What do I do? Most of the time, I grudgingly publish the paper as is.’

I also asked my interviewees whether they thought there were differences in the care taken by referees between different levels of journals (see Geist et al., 2010, p. 165). A few interviewees did not think so: ‘I feel like the level doesn’t matter very much. I think it has much to do with *who* does the refereeing than the level of journals.’ However, quite a few expressed the view that referees at ‘lower-level’ journals do less to check correctness: ‘As you go *down* in journals, I don’t think they referee (laughs)’, or ‘I think that many of the small journals, lower-level journals, I think the refereeing is almost non-existent.’ This fits with Nathanson’s (2008, p. 773) view that ‘Many (I think most) papers in most refereed journals are not refereed.’

For higher-ranking journals, editors thought that referees would be more careful when checking the results: ‘The higher the standing of the journal, the more the editors feel that the work of the referee has to be technically correct’. Moreover, ‘when an elite journal asks me to referee I try to do an elite job of refereeing’. When referees are requested by a top journal, they know that they will themselves be judged by the leading experts in their field who, as members of the editorial board, will read their referee report. On the other hand, top journals attract top papers, which will report more interesting results. If the result is interesting and introduces novel techniques, referees will *want* to do the checking, because they want to understand the paper for their own research. That is to say, 'if this paper is of *really* great importance, then people want to check this'. In that sense, 'when you have a really good paper, it's not so hard to get a referee'.

The more important a paper, the more scrutiny it will receive during peer review, something already observed by Zuckerman (1977, p. 95) who sees this tendency as 'an efficient pattern for deploying scarce resources'. However, one of my interviewees pointed to an interesting paradox:

> Very often the lower-level journals ask lower level people who don't get asked as much. Then they've more time. It's a paradox. … So I think it often happens that in the lower level journals, papers get refereed better because of that.

While ground-breaking results will certainly be checked by several referees in great detail (Krantz, 2003, p. 679) most papers, even at top journals, do not receive much scrutiny. In those cases, the standards for refereeing clearly vary considerably.

In sum, editors saw the primary responsibility for correctness lying with the author, while still placing some accountability on referees. As one of my interviewees stated: 'I want to give numbers that add up to more than 100%. It is 100% the responsibility of the author but it is also 85% the responsibility of the referee.'

Taking account of the author

Refereeing in mathematics is typically single-blind (Pitcher, 1988), although this practice may be changing. Many of my interviewees argued that double-blind refereeing would be impractical in mathematics, partly because the field is very specialized and referees can often guess who the author might be, and partly because most papers are already available on preprint servers and referees can therefore check who the authors are.

The main reason for adopting double-blind peer review has been to battle different forms of bias (Pontille & Torny, 2015, p. 64). Such bias may lead to referees viewing papers by certain authors, such as those from famous institutions as more important or significant than those of unknown authors. Alternatively, there may be forms of gendered or racial discrimination. Indeed, this reason was given in the announcement of the introduction of double-blind refereeing by the AMS in 2022: 'Double anonymity lowers the likelihood of implicit bias and therefore supports inclusivity and diversity across mathematics' (AMS, 2022).

In single-blind refereeing, referees may not check the proofs of known authors as carefully as those of unknown ones (Andersen, 2017, p. 187; Geist et al., 2010, p. 162; Hagstrom, 1965, p. 24). My interviewees confirmed this: 'I'm sure that, depending on who the author is, people read bits more or less carefully.' However, this aspect could also be seen as an attempt to make efficient use of a scarce resource (referees' time). Since the correctness of the paper is ultimately seen as the responsibility of the author, the referee does not have to spend days or even weeks going through the details of a proof written by a mathematician who has published many papers that have withstood the test of time. As one of my editors asked: 'The risk of having a flaw in the paper, in the end … rests with the author and not with the journal, then why go through too much effort?' Another interviewee remarked:

> If we want to have the process work, then in some cases we have to say that 'A paper by X, *maybe* there's a mistake in it, but the person has a track record of publishing interesting mathematics. Even if there is a technical mistake, there still would be some interest in the paper.' So we don't have to have it refereed to the same standard of stringency as in the case of a new PhD. In mathematics, if you had to follow the *same* level of stringency with *every* paper, the process would break down. It's impossible.

Unknown mathematicians, and in particular recent PhD students, are often seen as 'problems', since they do not yet have track records. 'For an unknown or younger person, there could be a large question about: Are these claims actually true? Is what's done, does it look like it can actually prove statements like this? And that's another kind of refereeing.' In other words, unknown authors are not yet trusted: 'New authors are young, you don't trust them yet.'

Even for known authors, referees do not trust them completely. Referees emphasized that authors only have a reputation in certain areas. If they suddenly publish work in a different area, a referee may check things more carefully: 'You *know* that certain people are extremely strong experts in a certain part of the subject, however, much less in something else. Then you pay more attention on the things where they are less expert, sometimes. You want to see what they do there.'

Editors were also well aware that trusting authorial reputation has pitfalls and can backfire:

> The more difficult situation is the situation of papers by very good mathematicians, on some very difficult and technical stuff, that contains statements that everybody thinks at the end to be true, but for which the proof is not completely correct in a really deep and subtle way.

In other words, authors may make mistakes that are very difficult to spot—by the authors themselves and the referee. This issue will be discussed in the section on referees encountering problems below.

Finally, authors can develop a negative reputation instead of a positive one. That is, editors and referees may hold the view that 'certain authors make mistakes all the time' and are known for writing 'sloppy papers'. This reputation can be remarked upon in the quick opinions, as one of my interviewees noted: 'Maybe the initial opinion giver goes on to say: "The author of the paper has had some correctness issues in the past." So you have to be really careful.' Indeed, I found one quick opinion that suggested someone as a possible technical referee, but warned the editor that this person did not have a reputation for being very careful as an author:

> Regarding your suggestion as a referee, I suggest that you find out if <Name> is more careful as a referee than as an author as his papers have many places that lack sufficient detail.

In sum, an author's reputation is certainly taken into account when checking correctness—up to a point. Referees take a differentiated view of reputation and only trust authors in certain areas. Furthermore, trusting authors can be seen as the other side of assigning the responsibility for correctness to authors. Referees may give some authors the benefit of the doubt, but will also hold those authors responsible for any errors that are eventually discovered. If mistakes are noted later, they could be a serious blow to a mathematician's reputation (Hagstrom, 1965, p. 27).

Checking correctness in practice

What do referees actually do to check the correctness of the results? Many of my interviewees started by saying what they would *not* do, namely read the paper line by line:

> If you're *really* asking yourself, 'Is this paper fundamentally correct?', the thing to do is *not* check it line by line, because you won't see the woods for the trees. … Suppose if a great result came out, and I just want to read the paper and find out what's new, I wouldn't read it line by line. I would somehow look for the *skeleton* of the paper, identify the steps that I thought, 'Yes, that seems plausible to me, you could do it by the techniques', and the steps where I say to myself, 'I have no idea how I would do that.' And then I would sort of work *down* trying to fill in the argument. And some parts of it I would never check, because they would seem so plausible that I could do them if I wanted. Other parts I'd spend a long time really trying to locate what it was that was new.

Referees explained that they would often start by trying to get an overview of the paper, to understand the rough structure or outline of the proof. They would then find what they consider the 'turning points' or 'key arguments' in the proof, which they would read in much greater detail (see Barany & MacKenzie, 2014, p. 122).

Part of the expertise of referees consists in being so familiar with the subject matter that they can quickly identify where such key places are, since they have worked on the topic already or are aware of what strategies have already been tried to solve the problem:

> In fields that people are familiar with, [referees] *know* what can go wrong. … So it's not a question of formally checking line by line so much as understanding where the problems are likely to appear.

One interviewee explained:

> When a question is really well known, the experts understand the bottleneck. Very often it is a question that they really thought about. Even if the paper is 70 pages, I open to page 37, because I know that *this* is where the action has to be because I thought about it before.

There are thus parts of the proof that are important for the general argument and referees will try to understand exactly how these work. In contrast, many other parts of the proof receive less scrutiny. Referees will skip over the elements they deem plausible or reasonable (Andersen, 2020, p. 237). One interviewee told of a recent example:

> The last paper I reviewed, … I read through very carefully in sections. The paper had, I think, 34 pages. That's already very long for a mathematics paper. And I have to admit, after the first part was really excellent, I simply believed the second part because it was very plausible.

Here the referee admits that different parts of the paper received different kind of scrutiny. While the first part was read ‘carefully’, the second part was not, because the referee deemed it ‘plausible’. Of course, it may be that a mistake is lurking in precisely those parts that are deemed plausible (as will be discussed in the next section).

In contrast, referees would often read parts of the paper carefully that looked ‘suspicious’ or ‘unlikely’ to them (see also Andersen, 2020, p. 239):

> The way you frequently find mistakes is you see some little place in some little lemma somewhere in the paper, it doesn’t quite fit with what you thought about how the world worked in this area, based on these other papers. Even if you can’t immediately point to a mistake, you can usually start looking at it, and eventually find, okay, there is some problem here.

It may be that a result ‘seems a bit too strong’ or that the methods that are used seem too weak: ‘You *know* that these techniques won’t be strong enough to prove the theorem that is claimed.’ Alternatively, there may be something in the writing that raises suspicions:

> As one reads the paper, very often, if you read a paper you can see: Oh, all of a sudden there’s something a little lackadaisical in the write-up and the level of details has changed and maybe that’s indicative of something. Of course at that point I would want every word to be checked.

In sum, when checking a paper, not every part gets equal treatment. Instead, referees will try to understand the broad structure of the argument, before zeroing in on what they consider to be the key places in it. Indeed, very often referees already have an understanding on where these key places will be, since they have worked on the problem themselves or are aware of what kind of methods have been used previously to try to solve it. In contrast, they often skip over details of the paper that seem plausible or reasonable (which may thus be where mistakes remain undiscovered). Finally, if something in the paper looks suspicious, for example a result that is ‘too good to be true’ or a technique that should not be strong enough to prove a particular result, referees will read it in great detail.

Referees encountering problems: Mistakes and gaps

When going through a paper, referees may encounter different kinds of problems in the written argument. To start with, every paper has typos or what might be called trivial mistakes. Consequently, there is almost always a long list of these at the end of the report (sometimes several pages long):

> I have found several typos and mathematical mistakes of the same level of triviality; I am attaching the list (everything in it is so minor that there is definitely no need for me to see the paper again).

On the other end of the spectrum, there can be serious mistakes—problems that might mean that the proof no longer works. However, it is incredibly rare that peer review reveals such serious mistakes: ‘There are mistakes that kill the result and those are rare actually. At least I haven't encountered too many.’

When it comes to referees encountering problems, the more common situation is not finding a mistake but thinking that the proof does not provide enough details. This is typically expressed as a ‘gap’ in the proof, which is a complicated issue. As described above, mathematicians agree that to write proofs that are clear—easily understood—it is not necessary for authors to write down every detail possible. Indeed, if a proof is too detailed, it may make understanding *more* difficult for the reader, as remarked upon in one report:

> Some parts of the paper are rather lengthy, which makes the paper not easy to read. In my opinion the paper could have been shortened without loss of rigor, given that many of the methods have been explained in detail at other places (by one of the authors).

Yet, in practice, referees frequently complain that the authors did not provide enough detail. This does not necessarily mean that referees believe that there is a serious problem. However, they do think that the current version of the proof is incomplete in some way. The reports contain many examples:

> In spots the authors are not careful and leave too much to the reader (making it difficult to verify).

> The lack of detail makes it extremely difficult to read when the proofs are more complicated and the reader has to figure out all the details on her or his own.

In these reports, the referees want the author to provide more details in the written argument to render the proof easier to follow. The referees here do not necessarily think that there is a problem with the argument. They complain that the current version makes it difficult for them to verify the proof—that too much work is left to the reader (or referee).

In contrast, referees may indicate in their report that they do not understand how a particular step works. I was able to find many examples of this in the written reports:

> There are some steps in these proofs which I cannot follow. I give details below.

> I could not understand the proof of Case 2 of Theorem 7.1. This is the most complicated proof of the paper.

> I still cannot follow the argument on pp.7-8. [in the second round of review]

Here referees indicate that they have identified a place in the argument that they deem incomplete, but in a more serious manner. Despite trying to work out the argument, they were not able to do so (again see Andersen, 2020, p. 239). From the perspective of the referee, it may be unclear how serious this issue is. Is this something that can be fixed through more detail or has the referee identified an actual error or oversight? Put differently, is this perhaps more of a problem of the *referee*, in that the referee is not able to understand something that the authors—and perhaps other referees—think is obvious? Indeed, a referee—in a second round—may suddenly understand a step that they previously thought was problematic. One referee wrote in the second round of revision: 'I'm not sure why I objected to this proof before! The result in Cases 1-4 does indeed seem straightforward to verify.' Some referees acknowledged this uncertainty:

> The proof has a gap (see below); I don't know how serious it is.
>
> Theorem 2.11; perhaps the theorem itself is correct, but I do not understand one place in the proof. In the middle of the proof, the authors write …. Thus, please supply the authors' argument here. If it fails, then one of the main results of Theorem 2.24 collapses.

None of my interviewees suggested that authors left these gaps deliberately, trying to get away with parts that they as authors were unsure about. However, it may be precisely in the places that authors treat as obvious and unproblematic that mistakes are hidden. As one of my interviewees put it: 'So in fact the mistakes have to be hidden somewhere, because at least the author cannot see them.'

A good referee then can, on occasion, identify a place in the proof that has been overlooked by the author as requiring detailed justification. It is not that the referee is able to necessarily identify anything like an error as authors are typically very careful; it is something more like an 'oversight'. However, a mistake may be lurking in such 'gaps':

> So in fact, the natural type of mistake is a gap. It's a spot where a miracle happens, and from the point of the referee, everything was clear before, and it's clearer after, but there is a kind of spot where by a miracle you can move from line three to line four, but it's a miracle in the eyes of the referee. And the referee tries to generally understand like any reader would, how do you move from there to there, but what's written doesn't suffice.

When referees identify a part where they get stuck, very often the author can deal with it through more careful writing. However, on occasion, the referee has stumbled on an issue that the authors had not previously considered.

A good example is the communication between Andrew Wiles' who proved Fermat's last theorem, and one of the referees of that proof, Nick Katz. They would frequently email questions that Wiles could easily answer (Singh, 2011, p. 278), which thus turned out not to be gaps. However, at one point they found an issue that Wiles was not able to answer and which turned out to be not 'just a minor difficulty but a fundamental flaw' (p. 279). This problem was something that had not only been missed by Wiles, but indeed by Katz himself when he initially listened to Wiles giving a lecture course on the proof (Singh, 2011, p. 280).

Sometimes it is easy for authors to fill a gap, but sometimes it takes a long time. And sometimes, in trying to fill a gap in one part of the paper, a new problem appears in another: 'The author will fill this gap, but by filling the gaps they have changes everywhere. Many times it happens like this.' After two or three revisions referees may start to be unwilling to engage with the paper. One report said, 'I should say that my patience with this paper is running low. If the next revision has similar problems and presents the same careless write-up, this will be my last round with this paper.' And the editor may agree with this, as one of my interviewee recounted an exchange with an editor who said to him:

'You know what? Three strikes and the author's out, right. If you can't tell whether it's true or not after *putting* in as much time as you've put in, then we just move it.'

The issue of gaps is therefore not always easily resolved. Occasionally, the authors can remain convinced that that they have a proof, while the referee remains unconvinced. Perhaps the most surprising issue is that in such a situation neither side may have sufficient evidence to convince the other. The argument provided by the author is in some way deemed insufficient by the referee, yet they have not been able to find a clear mistake that shows that there really is a problem with the proof or theorem. Instead, the referee can only say that they remain unconvinced. One of my interviewees worked with a colleague on a paper for several weeks; they did not think the paper was correct, but neither were they able to find a mistake. At the time of the interview, the process was still going. Similarly, another editor told me of a situation where they tried to get quick opinions for a paper but received no replies. Eventually, they spoke to someone at a party and found out the reason: 'The community is dubious about the result, but we didn't find a mistake yet.'

In sum, referees do not check every step of the proof, but they are good at identifying potential problems and gaps (Thurston, 1994, p. 169). Sometimes, authors can respond with enough detail to convince the referee and 'fill' the gap. Other times, authors may not be able to do so and may withdraw the paper. Very occasionally, the issue remains unresolved: the author remains convinced that they have an argument that constitutes a proof, but the referee remains unconvinced.

Discovering mistakes after peer review

As indicated in the introduction, the fact that the published literature contains mistakes is widely acknowledged among mathematicians. My interviewees also confirmed this: 'No one pretends that papers certified as correct are correct' and 'being published in a journal is not a certificate for correctness'. Yet senior mathematicians still say that they might have to remind graduate students of this fact. There are stories of doctoral students relying on the (published) result of someone else and only discovering at a late stage that there are mistakes in that paper:

> One of my students, for example, was finishing his PhD, and he was at the blackboard explaining his result, but I found something is wrong. This cannot be true. He told me, 'I checked all my computations, it was correct'. But he was relying on the paper of somebody else, which was published, and this paper was containing a mistake. So I think a good philosophy is to check all papers.

Most of my interviewees argued that serious mistakes are quite rare: 'Large mistakes don't seem to happen too often'. A recent study found that, among all disciplines, mathematicians express the greatest certainty about the resilience of their knowledge base (Ambrasat & Heger, 2020, p. 15). Indeed, one piece of evidence that serious mistakes are relatively rare lies in the fact that if they do occur, they become a 'tell-able story', especially if it happens at a top journal. When my interviewees did remember an example, they would often mention a specific journal like the *Annals* (*of Mathematics*) or

Inventiones (*Mathematicae*). For example: ‘Certainly in the *Annals* there are instances where years later somebody says “hey this is false” (chuckles).’

Mistakes are typically discovered when someone is trying to use the result in their own work, or prove the same result using different techniques. In other words, they are discovered when mathematicians do a ‘researcher’s reading’ rather than an ‘appraisal-oriented reading’ (Hirschauer, 2010) as they do during peer review. In the terminology of Peterson and Panofsky (2021), they are discovered when mathematicians are doing ‘integrative’ rather ‘diagnostic’ replication, which is ‘concerned with incorporating findings from a study for one’s own purposes’ (p. 586). Interestingly, it is therefore often the authors themselves who discover mistakes in their own papers: ‘I would say the people who typically find the error, I would say are usually the authors themselves, more than anybody else.’ I heard stories where an author discovered a mistake after a paper was accepted, but before or while it was being published—and had to withdraw the paper:

> Then, two days ago, I get this email from the author, whom I know. He said: ‘Dear, I'm sorry to say that such and such lemma turns out to have a serious error in it and I don’t know if the lemma is true anymore. … I hope it’s not too late to withdraw the paper.’

Here an author withdraws a paper that had already been accepted. One interviewee had done this himself as an author:

> In my case I submitted a paper to *Inventiones* a long time ago and I found a mistake when looking at the galley proofs. … It was a serious mistake, and the referee did not find it. Unfortunately, I had to write to *Inventiones* that I don’t want to publish this paper of mine. … The mistake killed the paper, but the result was correct, it was proved by someone two years later.

The more a paper is used by other people (or the authors themselves), the more scrutiny it will receive, and the more likely it is that mistakes will be discovered: ‘The important stuff, other people read it, and if there’s something wrong, it’s found out quickly.’ In contrast, those papers that are published but never read or used may contain undiscovered mistakes: ‘If the paper is not studied, the mistake may remain there.’ This has led some mathematicians to question whether peer review is really necessary:

> Some of my colleagues claim: … If something is important and interesting, then if it is just put on arXiv people will read it and will try and understand this and to apply this, and then if there is something wrong there they will soon find it and it will be known. But if it is not interesting then nobody will read it and it doesn’t matter whether it is correct or not.

However, most of my interviewees thought that peer review remained important, for both judging the importance of papers and checking their correctness.

The continued importance of peer review for checking correctness

When asked whether they thought that peer review in mathematics is still working, many of my interviewees answered that they thought it performs ‘reasonably well’. However, they also emphasised that the system is ‘under stress’, in particular because of what they consider an explosion of papers: ‘too many papers, not enough referees’. Consequently, many editors stated that it was increasingly difficult to find referees. A few editors thought that ‘the system is bursting’, indeed that ‘the system is collapsing’. However, the more common view was that ‘it’s not a crisis yet’ and that ‘it works *still*, so to speak’.

Most editors argued that peer review adds a lot of value. The presentation of papers is often substantially improved during peer review. However, one mentioned value of peer review was that it ensures that a paper is read carefully by at least one person (see Hirschauer, 2010, p. 96). As one my interviewees put it: ‘I think peer reviewing contributes significantly to the fact that papers are read, I'm exaggerating now, at least once, by one person.’ Despite all its faults, my interviewees argued that peer review therefore makes a significant contribution to the certification of results:

> Peer review adds *a bit* of certainty.

> If it goes through the peer review process, that’s not a guarantee that absolutely everything is correct, but it means it’s pretty plausible.

While publication in a journal is not a guarantee of correctness, almost all my interviewees trusted published papers more than those that had only been uploaded to the preprint server arXiv. This difference becomes visible when a paper under review relies on the results in another paper that has not yet been published (or accepted) in a journal. In this situation, some journals will not accept the paper under review:

> If the paper relies on something that’s not at least accepted for publication somewhere, then we say, ‘no’.

> I personally think that it’s completely unacceptable to publish a paper that relies on things that have not been published.

In sum, despite acknowledging the problems of peer review, almost all of my interviewees were keen to keep it as ‘some sort of quality assurance’ or ‘hygienic practice’ without which ‘math will be a jungle’. They saw no better alternative to peer review (see also London Mathematical Society [LMS], 2011). In the words of one my interviewees: ‘I guess as Winston Churchill said: Democracy is the worst form of government except for everything else. So refereeing and peer review is the worst form of checking a paper other than anything else.’

Conclusion

Using interviews with editors of mathematics journals and a corpus of referee reports, I have shown first, that mathematical peer review produces not unqualified certainty, but more certainty than otherwise; second, that this certainty increases or decreases after publication, as the mathematical community uses and validates or questions the result; and, third, that disagreement between referees and authors can sometimes be resolved productively but sometimes remains intractable.

First, while mathematicians acknowledge that mistakes do slip through the peer review process and that peer review therefore does not guarantee the correctness of published papers, they nevertheless value it. In their opinion, peer review 'adds *a bit* of certainty' and having a paper peer reviewed means that 'it's reasonably likely that the paper is correct'. As a consequence, mathematicians regard published peer-reviewed papers as more certain than those uploaded to the preprint server arXiv. There are other aspects apart from peer review that can add certainty to a paper; in particular, the reputation of the author, which can be 'good' (for being 'careful' and writing 'correct' papers) or 'bad' (for writing 'sloppy' papers). Yet, in terms of increasing certainty, perhaps the most important element is time.

Second, I have highlighted the importance of the role of the *community* when it comes to establishing the certainty of mathematical results. As Krantz (2023, p. 679) remarks, the evaluation of a paper continues after its publication and can take place over a period of years. Of course, it is not the passage of time per se that lends certainty to results, but rather the scrutiny a paper receives after publication—a paper that is never read after publication does not automatically become more or less certain. While the mathematical community treats the correctness of a paper as the primary responsibility of the author and simultaneously expects that referees do their best within reason to make sure that papers do not contain mistakes, it is ultimately the community as a whole that validates the correctness of results. As my interviewees put it: 'But because of the way mathematics works, the community as a whole has also some responsibility of assessing mathematics because all mathematics is being assessed by others.' This mirrors an observation made by Steingart (2012) in the context of studying the classification of finite groups: 'it had become the responsibility of the entire community to function as a "sieve" filtering local errors from serious ones' (Steingart, 2012, p. 201).

The more people use the results and techniques in a paper, and the longer time no significant mistakes are being discovered, the more the community's confidence in the results grows. As one of my interviewees put it 'The papers where people are relying on them, I think most of the time, the social process validates them.' This assertion mirrors a remark by MacKenzie (2001, p. 318), who argued that 'the members of the relevant specialist mathematical community, in interaction with each other, come to a collective judgment as to what counts as a mathematical proof.' As indicated by this statement, the mathematical community is not one big homogenous group, but rather consists of various specialist subfields and subdisciplines. What may be known within one subfield may not be known in another. Recall the episode of the editor above who was puzzled that they were not able to find a referee for a particular paper; only when the editor spoke to someone in the subfield did they discover that the community had 'doubts' about that paper but could not find a mistake either. This 'community knowledge' is an important resource for understanding mathematics, but it is not captured in the published pages (Bourguignon, 2011, p. 14).

This aspect is reflected in the requirements for some prizes in mathematics, according to which the prize will only be awarded two years after the paper has been published in a peer-reviewed journal. For example, in 2000 the Clay Mathematics Institute announced million-dollar prizes for solving any of the seven 'Millennium problems'. The Institute required not just publication in a peer-reviewed journal but also 'at least two years must have passed since publication' and 'the proposed solution must have received general acceptance in the global mathematics community' (Clay Mathematics Institute, 2018; s ee also Nathanson, 2008, p. 773). Similarly, the prize for solving Fermat's last theorem, bequeathed by Paul Wolfskehl in 1906, contained the following requirement: 'The award of the Prize by the Society will take place not earlier than two years after the publications of the memoir to be crowned.' (cited in Singh, 2011, p. 136). These requirements make publication in a peer-reviewed journal a necessary, but not sufficient condition (Castelvecchi, 2020, p. 177). It is the passing of time and the scrutiny the paper receives *after* publication that is the central condition for being awarded the prize.

Finally, while mathematicians may not quarrel over the results of calculations (Wittgenstein, 1953, §341), they can nevertheless disagree not only about a result's importance but its correctness. Deciding whether a written argument constitutes a proof is not always easy, contrary to some idealized accounts of mathematics, such as this official statement of the London Mathematical Society:

> Mathematics is distinguished by the fact that the results are not a matter for debate: when an argument is presented, it can be studied by other experts, who will determine whether it is correct and whether it is complete. Although it may take some time for particularly long or difficult arguments, there is no room for disagreement. (LMS, 2011)

In practice, debate and disagreement *do* arise. Referees may encounter a part of the paper that they don't understand, which will often be expressed in terms of a 'gap' in the proof. Frequently, the author can 'fill' the gap, re-writing the argument so that the referee can understand it. Sometimes the author agrees with the referee that there is indeed a gap but cannot find a way to solve it. In such a situation, a paper may be withdrawn.

Occasionally, there can be a situation where authors and readers disagree over whether the argument constitutes a proof. A recent very prominent example concerns the papers by Shinichi Mochizuki claiming to prove the abc-conjecture. They had been uploaded in 2012 to the Internet and were published in 2021 in a peer-reviewed journal, *Publications of the Research Institute for Mathematical Science*, of which Mochizuki is the chief editor. The mathematical community found these proofs 'difficult to digest' (Rittberg, 2021, p. 5588) seen as being written 'in an impenetrable, idiosyncratic style, built entirely on unfamiliar mathematical concepts' (Castelvecchi, 2020, p. 177). In response, the German mathematicians Scholze and Stix have written notes arguing that the written arguments do not constitute a proof. While this does not necessarily mean that 'abc is a theorem in Japan while still an open conjecture in Germany' (Bordg, 2021, p. 50), it is certainly the case that there is currently disagreement in the mathematical community: Some mathematicians believe the published proof is correct, others think it is flawed, and many suspend their judgement (Rittberg, 2021, p. 5589).

While a mathematical paper is self-contained in the sense that it is not a report of empirical materials gathered outside the paper through laboratory experiments or qualitative interviews, for example, the expertise necessary to understand, and more importantly to evaluate, a paper is not found in the paper itself; it resides in the mathematical community. It is the community that knows where the gaps in papers are, whether they are insignificant or significant, and whether solutions for them have been found. On rare occasions, the community may for a while disagree about the correctness of certain results, but overall mathematics is characterized by a very ‘tight’ community with an extraordinary high level of agreement on such matters. In that sense, the quotation from the London Mathematical Society above—‘there is no room for disagreement’—expresses the belief of the mathematical community that disagreements *can*, and eventually *will*, be resolved.

Acknowledgements

I am extremely grateful to all my interviewees for taking the time to share their views and experiences about peer review in mathematics with me. I also would like to thank Andrew Aberdein, Götz Hoppe, Matthew Inglis, Michael Lynch, Alex Preda, Fenner Tanswell, Keith Weber, and the editors and two anonymous reviewers from this journal, for their generous feedback and help in writing and revising this article. Special thanks to Michael Baranay and Ilana Gershon whose enthusiasm and insights were extremely helpful when I was struggling with the paper.

Funding

The author disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The work described in this paper was substantially supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. CUHK14600218).

ORCID iD

Christian Greiffenhagen https://orcid.org/0000-0002-7544-9246

References

Ambrasat, J., & Heger, C. (2020). *Barometer für die Wissenschaft: Ergebnisse der Wissenschaftsbefragung 2019/20*. Deutsches Zentrum für Hochschul-und Wissenschaftsforschung (DZHW). https://www.wb.dzhw.eu/downloads/wibef_barometer2020.pdf

American Mathematical Society. (1962). Manual for authors of mathematical papers. *Bulletin of the American Mathematical Society*, *68*(5), 429–444.

American Mathematical Society. (2015). Manual for journal editors. Retrieved December 14, 2021, from http://www.ams.org/journals/journal-manual/journal-manual.pdf

American Mathematical Society. (2022). News from the AMS: AMS adopts double-anonymous peer review for journals. https://www.ams.org/news?news_id=6983

Andersen, L. E. (2017). On the nature and role of peer review in mathematics. *Accountability in Research*, *24*(3), 177–192.

Andersen, L. E. (2020). Acceptable gaps in mathematical proofs. *Synthese*, *197*(1), 233–247.

Armstrong, J. S. (1997). Peer review for journals. *Science and Engineering Ethics*, *3*(1), 63–84.

Auslander, J. (2008). On the roles of proof in mathematics. In B. Gold & R. A. Simons (Eds.), *Proof & other dilemmas* (pp. 61–77). Mathematical Association of America.

Australian Research Council. (n.d.). *Journal rankings for ARC FoR*. Retrieved November 24, 2022, from https://www.austms.org.au/Rankings/AustMS_final_ranked.html

Avigad, J., & Harrison, J. (2014). Formally verified mathematics. *Communications of the ACM*, *57*(4), 66–75.

Barany, M. J. (2021). Abstract relations: Bibliography and the infra-structures of modern mathematics. *Synthese*, *198*(26), 6277–6290.

Barany, M. J., & MacKenzie, D. (2014). Chalk: Materials and concepts in mathematics research. In C. Coopmans, J. Vertesi, M. E. Lynch, & S. Woolgar (Eds.), *Representation in scientific practice revisited* (pp. 107–129). MIT Press.

Biagioli, M. (2003). Rights or rewards: Changing frameworks of scientific authorship. In M. Biagioli & P. Galison (Eds.), *Scientific authorship: Credit and intellectual property in science* (pp. 253–280). Routledge.

Boas, R. (1995). *Lion hunting & other mathematical pursuits*. The Mathematical Association of America.

Bordg, A. (2021). A replication crisis in mathematics? *The Mathematical Intelligencer*, *43*, 48–52.

Bourguignon, P. (2011). The role of publications in mathematical research. In *Mathematics journals: What is valued and what may change* (pp. 12–15). http://www.msri.org/attachments/workshops/587/MSRIfinalreport.pdf

Burnham, J. C. (1992). How journal editors came to develop and critique peer review procedures. In H. F. Maryland & R.E. Sojka (Eds.), *Research ethics, manuscript review, and journal quality* (pp. 55–62). ACS Miscellaneous Publications.

Buzzard, K. (2020). Proving theorems with computers. *Notices of the AMS*, *67*(11), 1791–1799.

Castelvecchi, D. (2020). Mathematical proof that rocked number theory will be published. *Nature*, *580*(7802), 177.

Clay Mathematics Institute. (2018). Rules for the millennium prize problems. Retrieved June 24, 2023, from https://www.claymath.org/millennium-problems/rules/

Crowley, J., Hezlet, S., Kirkby, R., & McClure, D. (2011). Mathematics journals: What is valued and what may change. *Notices of the AMS*, *58*(8), 1127–1130.

Davis, P. J. (1972). Fidelity in mathematical discourse. *American Mathematical Monthly*, *79*(3), 252–263.

Davis, P. J., & Hersh, R. (1981). *The mathematical experience*. Birkhäuser.

De Millo, R. A., Lipton, R. J., & Perlis, A. J. (1979). Social processes and proofs of theorems and programs. *Communications of the ACM*, *22*(5), 271–280.

Despeaux, S. E. (2011). Fit to print? Referee reports on mathematics for the nineteenth-century journals of the Royal Society of London. *Notes and Records of the Royal Society*, *65*(3), 233–252.

European Mathematical Society. (2012). Code of practice. Retrieved December 14, 2021, from http://www.euro-math-soc.eu/system/files/uploads/COP-approved.pdf

Fallis, D. (2003). Intentional gaps in mathematical proofs. *Synthese*, *134*(1), 45–69.

Garfinkel, H. (Ed.). (1986). *Ethnomethodological studies of work*. Routledge.

Geist, C., Löwe, B., & Van Kerkhove, B. (2010). Peer review and knowledge by testimony in mathematics. In B. Löwe & T. Müller (Eds.), *PhiMSAMP. Philosophy of mathematics: Sociological aspects and mathematical practice* (pp. 155–178). College Publications.

Greiffenhagen, C. (2014). The materiality of mathematics: Presenting mathematics at the blackboard. *The British Journal of Sociology*, *65*(3), 502–528.

Greiffenhagen, C. (forthcoming). Judging importance before checking correctness: Quick opinions in mathematical peer review.

Greiffenhagen, C., & Sharrock, W. (2011). Does mathematics look certain in the front, but fallible in the back? *Social Studies of Science*, *41*(6), 839–866.

Gunnarsdottir, K. (2005). Scientific journal publications: On the role of electronic preprint exchange in the distribution of scientific literature. *Social Studies of Science*, *35*(4), 549–579.

Hagstrom, W. O. (1965). *The scientific community*. Basic Books.

Hales, T. C. (2008). Formal proof. *Notices of the AMS*, *55*(11), 1370–1380.

Halmos, P. R. (1985). *I want to be a mathematician*. Springer.

Harrison, J. (2008). Formal proof—Theory and practice. *Notices of the AMS*, *55*(11), 1395–1406.

Hirschauer, S. (2010). Editorial judgments: A praxeology of ‘voting’ in peer review. *Social Studies of Science*, *40*(1), 71–103.

Jackson, A. (2002). From preprints to e-prints: The rise of electronic preprint servers in mathematics. *Notices of the AMS*, *49*(1), 23–31.

Krantz, S. G. (2003). Peer review. *Notices of the AMS*, *50*(6), 678–679.

Krantz, S. G. (2005). *Mathematical publishing*. American Mathematical Society.

Krantz, S. G. (2011). *The proof is in the pudding: The changing nature of mathematical proof*. Springer.

Lakatos, I. (1976). *Proofs and refutations: The logic of mathematical discovery*. Cambridge University Press.

Lamont, M. (2009). *How professors think*. Harvard University Press.

Larivière, V., Sugimoto, C. R., Macaluso, B., Milojević, S., Cronin, B., & Thelwall, M. (2014). ArXiv e-prints and the journal of record: An analysis of roles and relationships. *Journal of the Association for Information Science and Technology*, *65*(6), 1157–1169.

Lecat, M. (1935). *Erreurs de mathématiciens des origines à nos jours* [Errors of mathematicians: From the origins to the present day]. Castaigne.

Livingston, E. (1986). *The ethnomethodological foundations of mathematics*.Routledge.

Livingston, E. (1999). Cultures of proving. *Social Studies of Science*, *29*(6), 867–888.

London Mathematical Society. (2011). Written evidence. House of Commons Science and Technology Committee, Peer Review in Scientific Publications. HC 856. The Stationary Office. https://www.parliament.uk/business/committees/committees-a-z/commons-select/science-and-technology-committee/inquiries/peer-review/

Macbeth, D. (2021). Formal proofs in mathematical practice. In B. Sriraman (Ed.), *Handbook of the history and philosophy of mathematical practice*. Springer.

MacKenzie, D. A. (1999). Slaying the Kraken: The sociohistory of a mathematical proof. *Social Studies of Science*, *29*(1), 7–60.

MacKenzie, D. A. (2001). *Mechanizing proof*. MIT Press.

Mejia-Ramos, J. P., & Weber, K. (2014). Why and how mathematicians read proofs. *Educational Studies in Mathematics*, *85*(2), 161–173.

Nathanson, M. B. (2008). Desperately seeking mathematical truth. *Notices of the AMS*, *55*(7), 773.

Nathanson, M. B. (2009). Desperately seeking mathematical proof. *The Mathematical Intelligencer*, *31*(2), 8–10.

Peterson, D., & Panofsky, A. (2021). Self-correction in science: The diagnostic and integrative motives for replication. *Social Studies of Science*, *51*(4), 583–605.

Pitcher, E. (1988). *History of the second fifty years, American mathematical society, 1939–1988*. American Mathematical Society.

Polanyi, M. (1951). Self-government of science. In *The logic of liberty* (pp. 49–67). University of Chicago Press.

Pollatsek, H. (2018). How mathematics research journals select articles. *Notices of the AMS*, *65*(1), 63–64.

Pontille, D., & Torny, D. (2015). From manuscript evaluation to article valuation: The changing technologies of journal peer review. *Human Studies*, *38*(1), 57–79.

Quine, W. V. O. (1951). *Mathematical logic* (Rev. ed.). Harvard University Press.

Rittberg, C. J. (2021). Intellectual humility in mathematics. *Synthese*, *199*(3), 5571–5601.

Rosental, C. (2003). Certifying knowledge: The sociology of a logical theorem in artificial intelligence. *American Sociological Review*, *68*(4), 623–644.

Rosental, C. (2008). *Weaving self-evidence: A sociology of logic*. Princeton University Press.

Russell, B. (1956). *Portraits from memory*. Simon and Schuster.

Shapin, S. (1984). Pump and circumstance: Robert Boyle's literary technology. *Social Studies of Science*, *14*(4), 481–520.

Shapin, S. (1994). *A social history of truth*. University of Chicago Press.

Singh, S. (2011). *Fermat's last theorem*. Harper.

Steingart, A. (2012). A group theory of group theory: Collaborative mathematics and the 'uninvention' of a 1000-page proof. *Social Studies of Science*, *42*(2), 185–213.

Thurston, W. P. (1994). On proof and progress in mathematics. *Bulletin of the AMS*, *30*(2), 161–177.

Tymoczko, T. (1984). Gödel, Wittgenstein and the nature of mathematical knowledge. *Philosophy of Science Association*, *2*, 449–468.

Voevodsky, V. (2014). *The origins and motivations of the univalent foundations: A personal mission to develop computer proof verification to avoid mathematical mistakes*. *Institute for Advanced Study*. https://www.ias.edu/ideas/2014/voevodsky-origins

Watson, R. (2016). *Analysing practical and professional texts: A naturalistic approach*. Routledge.

Weller, A. C. (2001). *Editorial peer review: Its strengths and weaknesses*. Information Today.

Wiener, N. (1915). Is mathematical certainty absolute? *The Journal of Philosophy, Psychology and Scientific Methods*, *12*(21), 568–574.

Wittgenstein, L. (1953). *Philosophical Investigations*. Blackwell.

Zahler, R. (1976). Errors in mathematical proofs. *Science*, *193*(4248), 98.

Zuckerman, H. (1977). Deviant behavior and social control in science. In E. Sagarin (Ed.), *Deviance and social change* (pp. 87–138). Sage.

Author biography

Christian Greiffenhagen is Associate Professor at the Department of Applied Social Sciences, The Hong Kong Polytechnic University of Hong Kong, where he leads the Video Analysis, Science and Technology (VAST) Research Group, which develops video-based methodologies to study the impact of scientific and technological developments on people's lives. He is also a Visiting Associate Professor at the Department of Economics and Social Sciences, Télécom ParisTech.