Axiomathes (2022) 32 (Suppl 3):S1205–S1219
https://doi.org/10.1007/s10516-022-09647-x

ORIGINAL PAPER

Measuring the Agreement of Mathematical Peer Reviewers

Benedikt Löwe[1,2,3]

Received: 1 March 2022 / Accepted: 15 October 2022 / Published online: 21 December 2022

Abstract
We investigate the possibility of arguing for or against the philosophical position that mathematics is an *epistemic exception* on the basis of agreement data from the mathematical peer review process and argue that Cohen's κ, the standard agreement measure used for inter-rater agreement, is unable to detect epistemic exceptionality from peer review data.

1 Introduction

1.1 The Epistemic Exceptionality of Mathematics

There is no dispute about the fact that mathematicians disagree frequently about many mathematical issues, e.g., whether a mathematical result is interesting, which of two proofs of the same theorem is more natural than the other, or what the most salient features of a proof are.

It is a common belief, however, that on matters of mathematical correctness, mathematicians will be able to reach an agreement. This does not entail that mathematicians cannot disagree about the correctness of a proof; instead, the common

✉ Benedikt Löwe
loewe@math.uni-hamburg.de

[1] Institute for Logic, Language and Computation, Universiteit van Amsterdam, Postbus 94242, 1090 GE Amsterdam, The Netherlands

[2] Fachbereich Mathematik, Universität Hamburg, Bundesstraße 55, 20146 Hamburg, Germany

[3] Churchill College, Lucy Cavendish College, and Department of Pure Mathematics and Mathematical Statistics, University of Cambridge, Storey's Way, Cambridge CB3 0DS, England, UK

belief is that with a sufficient amount of effort and discussion of details, they will eventually reach agreement. This view is implicit in Leibniz's famous *calculemus* quote:[1]

> Quando orientur controversiae, non magis disputatione opus erit inter duos philosophos, quam inter duos computistas. Sufficiet enim calamos in manus sumere sedereque ad abacos, et sibi mutuo (accito si placet amico) dicere: calculemus.

Beyond mathematics, for science in general, the ability to resolve disputes and reach a consensus is seen as a fundamentally desirable hallmark of the scientific method. In popular discussions about science, an inability of experts to agree on whether a statement is correct or not is seen as major failure. This general sentiment underlies debates such as the discussion of the Sokal hoax[2] or the public discourse on scientific consensus, e.g., debates about vaccinations or climate change.[3]

We shall refer to the ability to reach consensus in principle as *intersubjective stability*; whether science is intersubjectively stable is very closely connected to the stronger claim of the objectivity of science (Reiss and Sprenger 2020). Following Hanson (1958), Kuhn (1962, 1977), and Feyerabend (1975), the received view in the philosophy of science community embraces the principle that science is *theory-laden* and is generally sceptical of strong claims of scientific objectivity. As a consequence, while intersubjective stability of judgments is in practice a very important indicator of scientific quality, the overall expectations of degrees of agreement are modest.

This situation changes drastically if we move from the empirical sciences to mathematics: many mathematicians and philosophers of mathematics alike believe in a very high degree of intersubjective stability in judgments about mathematical correctness.[4] The standard argument for this belief, sometimes called *derivationism*, assumes that correct mathematical proofs are warranted by formal derivations corresponding to them;[5] the latter are entirely surveyable objects whose correctness can be established with certainty. Therefore, a disagreement between

[1] The quote is usually cited to highlight Leibniz's conviction that argumentation can be transformed into some form of computation; here, we emphasise instead the underlying presumption that the *computistas* have no fundamental need for a *disputatio* in order to resolve controversies.

[2] The physicist Alan Sokal submitted a paper, deliberately constructed to be flawed, to the postmodern cultural studies journal *Social Text* in order to performatively criticise their standards of checking scientific correctness (Sokal 1996; Sokal and Bricmont 1997).

[3] The amount of public discussion about the certainty of science has been vast in the last few years, mostly due to the public debates about pandemic policies and vaccinations; we give an opinion piece of Ewan Birney (2021), the Deputy Director General of the European Molecular Biology Laboratory (EMBL) as a token example.

[4] This standard view is disputed by some authors who provide empirical evidence to discuss the level of agreement of mathematicians about correctness claims and their ability and willingness to reach consensus; cf. Inglis and Mejía-Ramos (2009), Inglis et al. (2013), Weber and Mejía-Ramos (2015), Weber and Czocher (2019).

[5] Azzouni (2004) considers informal mathematical proofs as indicators for the existence of formal derivations.

two mathematicians about the correctness of a proof can be resolved *in principle* by *calamos in manus sumere sedereque ad abacos*.[6] Derivationism relies on the deductive nature of mathematics; this has been used to claim that mathematics is an *epistemic exception* among the sciences: its deductive nature gives mathematicians a categorically different epistemic access to judgments about the correctness of mathematical argument.[7] Hence, the claimed epistemic exceptionality of mathematics has been used to explain the reported intersubjective stability as well as the perceived lack of lasting disagreements and scientific revolutions in mathematics.[8]

1.2 Peer Review

The perceived differences between the sciences in general and mathematics in particular also apply to the discussion about the purpose, nature, and measures of success of peer review.

In the wider field of science, the general view is that the main purpose of peer review is not to serve as a mechanism to check the correctness of claims in a paper: it is not assumed that peer review can guarantee correctness; on the other hand, mere correctness is not enough to make a submission publishable.[9] Mirroring the contrast concerning intersubjective stability, the discussion of the purpose of peer review in mathematics differs from that in science in general. The *London Mathematical Society* presents a very different view of peer review from that of the learnèd societies from other disciplines (cf. fn. 9):

> 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. This gives Peer Review an especially significant role in mathematics. [...] Because of the extreme density of mathematical writing, Editors will usually expect Reviewers to take around two months unless the paper is especially long, difficult or innovative. It is not unusual for assiduous Reviewers to take several times this long. [...] Because it is necessary to invest so much effort in reading a single paper, it is extremely valuable to the com-

[6] For a discussion of the qualifier "in principle" and the tension between mathematical practice and this position, cf. Müller-Hill (2011); cf. also Tanswell (2015).

[7] Cf. Prediger (2006a, 2006b), Buldt et al. (2008).

[8] For a discussion of these latter perceptions, cf. François and Van Bendegem (2010).

[9] In evidence of this, we give two quotes from a report on peer review of the House of Commons Science and Technology Committee in the United Kingdom: The International Association of Scientific, Technical and Medical Publishers states in their text: "While peer review cannot prove that a paper is 'correct' or that the data presented is not fraudulent, it is widely accepted by both authors and readers as greatly improving the quality of reported research. (House of Commons Science and Technology Committee 2011, p. 126)". Thomas Ward, Pro-Vice-Chancellor of the University of East Anglia describes: "[P]eer review [...] acts as a filter, selecting research that seems to the reviewers sound, significant, and of value. This is not the same as a guarantee of truth, nor does it imply an independent replication of the work. (House of Commons Science and Technology Committee 2011, p. 96)".

> munity that published papers have been declared correct by experts. [...] The main benefits of Peer Review are that it ensures the correctness and clarity of the content. (House of Commons Science and Technology Committee 2011, p. 100*sq*)

This highly idealised description of the *London Mathematical Society* should be contrasted with a very sceptical view among some mathematicians whether mathematical peer review actually provides the mentioned "main benefits". In an opinion piece published in the *Notices of the American Mathematical Society*, Nathanson (2008) complains:

> Many (I think most) papers in most refereed journals are not refereed. There is a presumptive referee who looks at the paper, reads the introduction and the statement of the results, glances at the proofs, and, if everything seems okay, recommends publication. Some referees check proofs line-by-line, but many do not. When I read a journal article, I often find mistakes. Whether I can fix them is irrelevant. The literature is unreliable.

The discrepancy between the idealised picture of the referee as a correctness checker and the actual practice of mathematical peer review was analysed by Geist et al. (2010). Their data suggests that the mathematical peer review process is not fundamentally different from the peer review in other sciences: a referee is being asked to answer the questions referred to as *Littlewood's precepts*: "(1) Is it new? (2) Is it correct? (3) Is it surprising? (Krantz 1997, p. 125)". Only the second precept relates to correctness checking and there is no consensus in the mathematical community what level of detail of correctness checking is done by the referee, is expected of the referee, or even should be expected of the referee. An identified fundamental error will certainly lead to rejection of the submission, but the judgment on the other two precepts ("Is it new?" and "Is it surprising?") is much less likely to fall under the scope of epistemic exceptionality of mathematics. As a consequence, the referee's verdict is an aggregation of a judgment that possibly has epistemically exceptional status and other judgments that do not.

Since a lack of intersubjective stability in the peer review process is considered to be worrying for the scientific quality of the publication decisions, a rich literature on peer review focuses on stability of reviewer judgments, the effect of bias, and reproducibility of decisions.[10] This worry is very perspicuous in the subtitle of the seminal paper by Rothwell and Martyn (2000): *Is agreement between reviewers any greater than would be expected by chance alone?*

Rothwell and Martyn performed an agreement analysis using Cohen's κ, a statistical measure of inter-rater agreement, and found that in papers of their subject (clinical neuroscience) "there was little or no agreement between the reviewers (Rothwell and Martyn 2000, p. 1964)". A similar study with a similar result was later done for journals in information science by Wood et al. (2004).

[10] Cf., e.g., Justice et al. (1998), Nylenna et al. (1994), McNutt et al. (1990), Abby et al. (1994), Cleary and Alexander (1988).

These studies were taken up by Geist et al. (2010) who did a similar analysis for peer review processes (for conferences) in the deductive sciences (mathematics and theoretical computer science), comparing the values of Cohen's κ from clinical neuroscience, information science, and the deductive sciences. They found that the agreement values in the deductive sciences are considerably higher than those in the comparison disciplines.

In the past decade, these findings have been repeatedly discussed at workshops and conferences in the wider community usually called *Philosophy of Mathematical Practice*[11] resulting in the following question:

> Can we give an argument for or against epistemic exceptionality of mathematics on the basis of empirical data of peer reviewer agreement in the form of values of Cohen's κ? (*)

In this paper, we answer (∗) negatively. We provide a toy model for the calculation of values of κ that shows that under realistic assumptions about the practice of mathematical peer review even perfect epistemic access to the correctness of proofs would not be detectable in the value of κ.

In Sect. 2, we shall give a general description of the definition and use of Cohen's κ; in Sect. 3, we define a toy model which we then use in Sect. 4 to answer (∗) negatively.

2 Cohen's κ

2.1 Cohen's κ as a Proxy for Sensitivity

Cohen's κ is a statistical measure for inter-rater agreement (Cohen 1960). It compares the degree of agreement of pairs of judgments to what would be expected if the judgments were randomly distributed. Its general formula is

$$\kappa = \frac{F - E}{1 - E}$$

where F is the observed frequency of agreement and E is the expected value of the frequency of agreement. A value of $\kappa = 1$ corresponds to full agreement ($F = 1$, i.e., every single pair of judgments is in agreement) and a value of $\kappa = 0$ corresponds to agreement according to the expected value ($F = E$).

Cohen's κ can be used as an informative proxy for the quality of a checking mechanism: suppose there is collection of samples, a property P of samples that we

[11] Recent descriptions of the work of this community can be found in Carter (2019), Hamami and Morris (2020). For a discussion of this community's scope, cf. Löwe (2018).

Table 1 Values of Cohen's κ for a test for property P (and no chance of false positives; cf. fn. 14): the rows correspond to the sensitivity of the test, the columns to the probability of a sample having property P

	50%	20%	10%	5%	2%
25%	0.14	0.21	0.23	0.24	0.25
50%	0.33	0.44	0.47	0.49	0.50
75%	0.60	0.71	0.73	0.74	0.75
80%	0.67	0.76	0.78	0.79	0.80
90%	0.82	0.88	0.89	0.90	0.90
95%	0.91	0.94	0.95	0.95	0.95
99%	0.98	0.99	0.99	0.99	0.99
100%	1.00	1.00	1.00	1.00	1.00

Table 2 Values of Cohen's κ for a test for property P with specificity $s = 0.99$: the rows correspond to the sensitivity of the test, the columns to the probability of a sample having property P

	50%	20%	10%	5%	2%
25%	0.13	0.17	0.16	0.13	0.08
50%	0.32	0.40	0.39	0.34	0.24
75%	0.58	0.66	0.64	0.58	0.44
80%	0.65	0.71	0.69	0.63	0.49
90%	0.80	0.83	0.80	0.73	0.57
95%	0.86	0.99	0.85	0.78	0.62
99%	0.96	0.94	0.90	0.82	0.66
100%	0.98	0.95	0.91	0.83	0.66

wish to check, and a checking mechanism that tests a sample whose sensitivity, i.e., the probability that a sample with property P tests positive, is not known.[12]

In a situation where the test is the only epistemic access to whether a sample has property P or not, it is difficult to determine the sensitivity of the test. One possible way to do so is to test each sample twice by independent tests and calculate the value of inter-rater agreement. Assuming independence of the tests and a random distribution of property P, the value of Cohen's κ is a remarkably good proxy for the sensitivity of the test as can be seen in Table 1.[13] If the probability of a sample having property P is 20% or less, the value of κ is very close to the sensitivity of the test.[14]

[12] In the biomedical literature, the *sensitivity* of a test is the probability that a sample with probability P produces a positive result ("no false negatives"); its *specificity* is the probability that a sample without property P produces a negative result ("no false positives"; cf. fn. 14).

[13] Table 1 shows the κ value calculated with the formulas from Sect. 3 with $p = 1$.

[14] The reason why our Table 1 does not exhibit the *first kappa paradox* of Feinstein and Cicchetti (1990) (small values of κ despite high agreement if the prevalence of property P is very small; cf. Sect. 2.2) is that our toy model suppresses the possibility of false positives; in the terminology of fn. 12, our procedure has a *specificity* of 1. If the specificity is close to, but not equal to 1, i.e., there is a small chance that something without property P tests positive producing a likely disagreement. If furthermore the prevalence of P is very small, then the overall frequency of disagreements due to false positives outweighs the effect of the true positives, thereby considerably decreasing the value of κ. If we include a parameter s for the specificity of the test in our formulas in Sect. 3, we obtain $a = ep - emp + ps - eps$, $r = 1 - a$, and

Table 3 Arbitrary descriptors for the strength of agreement by Landis and Koch (1977, p. 165).

κ	Strength of agreement
<0.00	Poor
0.00–0.20	Slight
0.21–0.40	Fair
0.41–0.60	Moderate
0.61–0.80	Substantial
0.81–1.00	Almost Perfect

2.2 Interpreting Cohen's κ

The statistical literature has a rich discussion of the strengths and weaknesses of Cohen's κ: the so called *kappa paradoxes* (Feinstein and Cicchetti 1990; Cicchetti and Feinstein 1990) describe undesirable behaviour of the function κ in practice (cf. fn. 14). Numerous variants of Cohen's κ have been proposed to deal with these unwanted features of the function, e.g., Scott's π (Scott 1955), the *G*-index by Holley and Guilford (1964), Bangdiwala's B (Bangdiwala 1985), or Gwet's AC1 (Gwet 2010). Many of the alternative measures share the basic mathematical structure of Cohen's κ and the results for our toy model in Sect. 3 would not be expected to differ substantially if we had used a different inter-rater agreement measure.

In spite of the criticism and alternative proposals, Cohen's κ has developed into a standard measure for inter-rater agreement. However, interpreting the meaning of values of Cohen's κ is difficult and highly context-sensitive. It is curious that many textbooks and papers refer to the scale provided by Landis and Koch (1977) given in Table 3 even though Landis and Koch (1977, p. 164*sq*) themselves state very clearly that these descriptors are merely a convention for their paper:

> In order to maintain consistent nomenclature when describing the relative strength of agreement associated with kappa statistics, the following labels will be assigned to the corresponding ranges of kappa. Although these divisions are clearly arbitrary, they do provide useful "benchmarks" for the discussion of the specific example.

In addition to the arbitrary choice of the numerical divisions, the descriptors used (e.g., "substantial agreement" for κ between 0.61 and 0.80) should not be expected

Footnote 14 (continued)

$$F = (1-e)\left[s^2\left(1-2p+2p^2\right)+(1-s)^2+\left(2s-2s^2\right)(1-p)\right] + e\left[m^2+2m(1-m)(1-p)+(1-m)^2\left(1-2p+2p^2\right)\right]$$

while the rest of the formulas remains the same. Table 2 gives the same values as Table 1 with the modified formulas and $s = 0.99$: it is apparent that the value of Cohen's κ ceases to be a good proxy for the sensitivity as soon as the prevalence of P becomes small compared to the specificity. For more on the relationship between Cohen's κ, sensitivity, and specificity, cf. Feuerman and Miller (2005, 2008).

to work across application contexts: what counts as "substantial agreement" for a flaw check in a paperclip factory may not meet the standards of a test for infection with a highly contagious deadly disease.

3 A Toy Model

In this section, we describe a toy model for mathematical peer review in which we identify the philosophical claim that mathematics is an epistemic exception as a particular special case. Detecting epistemic exceptionality with the values of Cohen's κ therefore becomes the question whether knowledge of κ allows us to conclude that we are in that particular special case.

We emphasise that our toy model is by no means intended to be a realistic mathematical model of peer review: we make a number of simplifying assumptions that are not true of mathematical peer review in practice. However, we claim that our simplifications will only make it easier to detect epistemic exceptionality, so that if even under these simplifying assumptions, the value of Cohen's κ cannot detect the epistemic exceptionality, then this will remain true under more realistic assumptions.

We assume that we have a collection of papers submitted for review (*submissions*) and a population of peer reviewers (*referees*). The recommendation of a referee reviewing a submission will depend on whether the referee detects a fundamental flaw and, if not, whether the referee considers the paper suitable for publication in the journal to which it was submitted. We formalise this in three parameters e, m, and p:

Correctness. Submissions can either be *correct* or *incorrect*. Here, *incorrect* means that the paper is fundamentally flawed and must not be published. We ignore all of the subtle situations between these two extremes: that errors in the submission might be fixed as part of the peer review process; that papers could be partially correct and publishable after removing the errors; etc. We write e (for "error probability") for the probability that a submission is incorrect.

Mastery. This parameter is the ability of the referee to detect flaws of the type discussed in the preceding paragraph. We write m (for "mastery") for the probability that a referee is able to detect that a paper is fundamentally flawed; in that case, we shall say that the referee is *masterful*. In biomedical terminology, the mastery is the sensitivity of the flaw testing provided by the referees. As with correctness, we ignore many of the subtleties of the refereeing process: that a referee may spot a flaw in parts of the paper, but the rest of the paper is still correct and valuable; that a referee thinks they found a flaw that later turns out not to be one, etc.[15]

In general, it is reasonable to assume in all disciplines that expert referees will have a good chance of spotting fundamental flaws; the claim that mathematics is epistemically exceptional amounts to the assumption that mathematical referees are

[15] This is the justification to ignore false positives in our toy model (cf. fn. 14): one of our simplifying assumptions is that the false belief of a referee that a paper was fundamentally flawed will be rectified later in the process.

in principle always able to detect them, i.e., all referees are masterful or, in terms of the toy model, $m = 1$.

Positivity. Assuming that a paper is correct or that a referee has not identified its fundamental flaw, the referee's decision to recommend acceptance or rejection is based on many non-mathematical factors, e.g., whether the referee likes the research area, whether the referee thinks that this theorem is good enough for this journal, etc. We simplify all of these factors into a single probability that we call the *positivity* of the referee. The positivity p is the probability that a referee will recommend acceptance, provided she or he has not detected that it is *incorrect*.

Based on our parameters e, m, and p, we can now analyse what the probability of the recommendations "accept" and "reject" are:

Case 1. The paper is incorrect and the referee is masterful. This means that the referee detects the incorrectness and recommends rejection. This case happens with probability $e \cdot m$.

Case 2. The paper is incorrect and the referee is not masterful. This means that the referee does not detect the incorrectness and the recommendation depends on the positivity of the referee. Case 2 results in "accept" with probability $e \cdot (1 - m) \cdot p$ and in "reject" with probability $e \cdot (1 - m) \cdot (1 - p)$.

Case 3. The paper is correct. In this case, the mastery of the referee is irrelevant, so the referee will recommend acceptance with probability $(1 - e) \cdot p$ and rejection with probability $(1 - e) \cdot (1 - p)$.

Based on our three cases, we calculate the probabilities a and r of an acceptance or rejection recommendation and the expected value E of agreement, respectively:

$$\begin{aligned} a &= e \cdot (1 - m) \cdot p + (1 - e) \cdot p = p - emp, \\ r &= 1 + emp - p, \text{ and} \\ E &= a^2 + r^2 = 1 + 2m^2p^2e^2 - 4mp^2e + 2mpe + 2p^2 - 2p. \end{aligned}$$

Assuming we assign two referees to a given submission, we can either have two masterful referees (with probability m^2), one masterful referee (with probability $2m(1 - m) = 2m - 2m^2$), or no masterful referees (with probability $(1 - m)^2 = 1 - 2m + m^2$). If a submission is correct, the two referees will agree with probability $p^2 + (1 - p)^2 = 1 - 2p + 2p^2$, independent of the mastery of the referees. If a submission is incorrect, masterful referees will always agree, non-masterful referees will agree with probability $p^2 + (1 - p)^2 = 1 - 2p + 2p^2$, and a masterful referee will agree with a non-masterful referee with probability $1 - p$. Combining these numbers, we obtain the following formula for the frequency of agreement:

$$\begin{aligned} F :&= (1 - e)\left[1 - 2p + 2p^2\right] + \\ &\quad e\left[m^2 + 2m(1 - m)(1 - p) + (1 - m)^2(1 - 2p + 2p^2)\right] \qquad (\dagger) \\ &= 1 + 2m^2p^2e - 4mp^2e + 2mpe + 2p^2 - 2p. \end{aligned}$$

From (†), we can now calculate the value of Cohen's κ as

Table 4 Three tables of value of Cohen's κ for $e = 0.1$, $e = 0.2$, and $e = 0.3$ in the toy model: the rows correspond to the value of m and the columns to the value of p

$e = 0.1$	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
0.1	0.000	0.000	0.000	0.001	0.001	0.001	0.002	0.003	0.008	0.091
0.2	0.000	0.001	0.002	0.002	0.004	0.005	0.008	0.014	0.028	0.184
0.3	0.001	0.002	0.004	0.005	0.008	0.012	0.018	0.030	0.059	0.278
0.4	0.002	0.004	0.006	0.010	0.014	0.021	0.032	0.052	0.099	0.375
0.5	0.003	0.006	0.010	0.015	0.023	0.033	0.049	0.079	0.147	0.474
0.6	0.004	0.008	0.014	0.022	0.033	0.047	0.071	0.111	0.201	0.574
0.7	0.005	0.012	0.020	0.030	0.044	0.064	0.095	0.148	0.262	0.677
0.8	0.007	0.015	0.026	0.040	0.058	0.084	0.123	0.190	0.328	0.783
0.9	0.009	0.020	0.033	0.050	0.073	0.106	0.154	0.236	0.398	0.890
1.0	0.011	0.024	0.041	0.063	0.091	0.130	0.189	0.286	0.474	1.000

$e = 0.2$	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
0.1	0.000	0.000	0.001	0.001	0.002	0.002	0.004	0.006	0.012	0.082
0.2	0.001	0.002	0.003	0.004	0.006	0.009	0.014	0.023	0.044	0.167
0.3	0.002	0.004	0.006	0.010	0.014	0.021	0.031	0.049	0.090	0.255
0.4	0.003	0.007	0.012	0.018	0.026	0.037	0.055	0.084	0.146	0.348
0.5	0.005	0.011	0.018	0.028	0.040	0.058	0.084	0.127	0.211	0.444
0.6	0.007	0.016	0.027	0.040	0.058	0.083	0.119	0.177	0.283	0.545
0.7	0.010	0.022	0.037	0.056	0.080	0.113	0.160	0.234	0.363	0.651
0.8	0.013	0.029	0.049	0.073	0.105	0.147	0.207	0.297	0.450	0.762
0.9	0.017	0.038	0.063	0.094	0.134	0.187	0.260	0.368	0.543	0.878
1.0	0.022	0.048	0.079	0.118	0.167	0.231	0.318	0.444	0.643	1.000

$e = 0.3$	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
0.1	0.000	0.001	0.001	0.001	0.002	0.003	0.005	0.008	0.015	0.072
0.2	0.001	0.002	0.004	0.006	0.008	0.012	0.018	0.029	0.052	0.149
0.3	0.002	0.005	0.009	0.013	0.019	0.027	0.040	0.061	0.103	0.231
0.4	0.004	0.009	0.016	0.024	0.034	0.049	0.070	0.103	0.165	0.318
0.5	0.007	0.015	0.025	0.037	0.054	0.076	0.107	0.154	0.237	0.412
0.6	0.010	0.022	0.037	0.055	0.078	0.109	0.151	0.214	0.317	0.512
0.7	0.014	0.031	0.051	0.076	0.108	0.149	0.204	0.283	0.406	0.620
0.8	0.019	0.042	0.069	0.102	0.143	0.195	0.265	0.361	0.504	0.737
0.9	0.025	0.055	0.090	0.132	0.183	0.249	0.334	0.448	0.611	0.863
1.0	0.032	0.070	0.114	0.167	0.231	0.310	0.412	0.545	0.730	1.000

Values greater than 0.8 are marked in dark grey; values between 0.6 and 0.8 are marked in mid grey; the column marked in light grey is discussed in Sect. 4.4

$$\kappa = \frac{F - E}{1 - E} = \frac{m^2pe - m^2pe^2}{1 + 2mpe - m^2pe^2 - me - p}.$$

As mentioned, Table 1 shows some values of this function for the special case $p = 1$; in this case, if $e < 0.5$, the value of Cohen's κ is a remarkably good proxy for the value of m. Therefore, in this special case, we could check whether all referees are masterful by checking the value of Cohen's κ.

In general, we give the numerical values of κ in Table 4 for the values $e = 0.1$ (first block; i.e., 10% of all submissions are fundamentally flawed), $e = 0.2$ (second block; i.e., 20% of all submissions are fundamentally flawed), and $e = 0.3$ (third block; i.e., 30% of all submissions are fundamentally flawed). An inspection of the tables yields that κ ceases to be a good proxy for m if $p \neq 1$: the values of κ are below 0.2 unless the positivity is 0.8 or higher or the mastery is 0.7 or higher. Values in the Landis–Koch category of "substantial agreement" (0.61 or higher) require either $p = 1$ or $p \geq 0.9$ and $m \geq 0.9$; only values in the column of $p = 1$ reach the Landis–Koch category of "almost perfect agreement".

We should like to emphasise that our toy model is not specifically about mathematical peer review or even about peer review processes. It is a general model for any decision making process that aggregates a correctness check (with no false positives) and a subjective judgment into a single binary answer: submissions identified as flawed are to be rejected and all others will receive a verdict on the basis of the subjective judgment. The subjective judgment is represented by our parameter p: if $p = 1$, we revert to the case of a mechanical check of a property P from Sect. 2. Table 4 shows that if a non-trivial amount of subjectivity is injected into the aggregation, the value of Cohen's κ drops substantially.[16]

4 Interpretation and Discussion

4.1 Masterful, but Resource-Bounded Referees

Before inspecting the numerical details of the values of Cohen's κ, we should like to discuss an interpretation issue with the attempt to argue for epistemic exceptionality from quantitative data of κ:

Even in a world of epistemically exceptional referees, they remain human beings and are resource-bounded. As mentioned in the report of the *London Mathematical Society* (Sect. 1.2), checking the details of a proof requires a great deal of time and energy and not every referee can or is willing to invest that time and energy. It is therefore conceivable that we live in a world of epistemic exceptionality where nevertheless not every referee uses their special epistemic powers. In fact, we know that there are referees who do not check the proofs; in the survey of editors of mathematical journals by Geist et al. (2010), only half of the editors reported that they "think the referee should check all of the proofs in detail" and one of them was realistic enough to comment "to be reasonable, I am happy when I find a referee [who checks some proofs in detail] (Geist et al. 2010, p. 163*sq*)".[17]

The methodological issue is that the values of Cohen's κ are unable to distinguish between a situation where all referees are masterful and can *in principle* detect errors with absolute certainty, but only 70% of them invest the time and energy to do so, and a situation where mathematics is not epistemically exceptional and referees just have an average mastery level of 0.7. This presents a structural problem with any argument for or against epistemic exceptionality on the basis of values of Cohen's κ.

[16] In this context, the remark by Feinstein and Cicchetti (1990, p. 548) concerning their first kappa paradox is illustrative: "The presumptive reason for the chance correction in κ is that the measuring instruments are human observers, rather than inanimate technologic (sic!) procedures, and that the subjective responses of the observes may sometimes agree by chance."

[17] Cf. also Andersen (2017).

4.2 Realistic Ranges of the Parameters for Correctness and Positivity

In our argument concerning question (∗) (cf. Sect. 4.3), it would be methodologically satisfying to use good estimates for the actual values of the parameters e and p in the practice of mathematical peer review based on solid quantitative data. However, such data does not exist, so we need to give realistic ranges for p and e on the basis of personal experience.

Correctness. Based on many years of experience as an editor and managing editor of mathematical journals, the present author expects that only a small number of submissions is fundamentally flawed in a way that makes the publication of any version of the submission impossible on the grounds of correctness alone. The chosen values $e = 0.1$, $e = 0.2$, and $e = 0.3$ in Table 4 reflect that.

Positivity. If we fix a particular mathematical research journal and have access to collection of editorial decisions, we could estimate the value of p for that journal by counting the number of rejections based on criteria that do not involve correctness, e.g., the scope of the journal, the level of difficulty, novelty, or relevance of the result. We expect that this empirically determined value of p depends significantly on the choice of the journal. While this data could be obtained in principle, as far as we know, no robust quantitative study has been done to date. Personal experience with mathematical research journals suggests that a substantial number of submissions is rejected and, given that fundamental flaws are rare (see above), a value of $p = 0.5$ or below is a realistic assumption.

4.3 Negative Answer to Question (∗)

With the realistic ranges for e and p, we can now make question (∗) precise in terms of the toy model: if e and p are in the realistic range, is it possible to determine from the values of Cohen's κ whether the toy model is epistemically exceptional (i.e., whether $m = 1$)?

The answer is "no": inspecting the values in Table 4 in the columns for $p = 0.5$ or below, we see that the values are all below 0.2. There is no categorical difference between the values at $m = 1$ from the values at $m < 1$. E.g., in an epistemically exceptional world ($m = 1$) with $p = 0.45$ and $e = 0.2$, the value of Cohen's κ is 0.141; this is the same as the value in a world with $m = 0.8$, $p = 0.5$, and $e = 0.295$. Similarly, in an epistemically exceptional world ($m = 1$) with $p = 0.3$ and $e = 0.1$, the value of Cohen's κ is 0.041; this is the same as the value in a world with $m = 0.8$, $p = 0.35$, and $e = 0.13$ or $m = 0.6$, $p = 0.34$, and $e = 0.28$.

4.4 Cohesion

One of the implicit assumptions in our toy model that could be criticised is the fact that referee judgments on correct papers are made randomly according to the positivity p. In the calculation of the value of F in (†), this is reflected in the two factors $p^2 + (1 - p)^2 = 1 - p^2 + 2p^2$ corresponding to the probability of agreement of

Table 5 Values of Cohen's κ for $e = 0.2$ and $p = 0.5$ with additional cohesion parameter c, using the formula (‡): the rows correspond to the value of m and the columns to the value of c

$e = 0.2$ $p = 0.5$	0.5	0.6	0.7	0.8	0.9	0.95	1
0.1	0.002	0.194	0.387	0.579	0.772	0.868	0.964
0.2	0.006	0.192	0.378	0.564	0.750	0.843	0.936
0.3	0.014	0.195	0.375	0.555	0.735	0.826	0.916
0.4	0.026	0.201	0.377	0.552	0.728	0.816	0.903
0.5	0.040	0.212	0.384	0.556	0.727	0.813	0.899
0.6	0.058	0.227	0.394	0.565	0.734	0.818	0.903
0.7	0.080	0.247	0.414	0.581	0.747	0.831	0.914
0.8	0.105	0.271	0.437	0.603	0.768	0.851	0.934
0.9	0.134	0.300	0.465	0.631	0.797	0.880	0.963
1.0	0.167	0.333	0.500	0.667	0.833	0.917	1.000

The column corresponding to $c = 0.5$ is marked in light grey as it corresponds to the same column in Table 4

two referees on a paper with no detected flaw. One could argue that as part of a social community of mathematicians sharing similar values, agreement about judgments of novelty or relevance is more likely than expected if it were a pure chance decision. We can model this in our toy model by replacing the two occurrences of $p^2 + (1-p)^2 = 1 - p^2 + 2p^2$ with a new parameter $c \geq p^2 + (1-p)^2$ for *cohesion*, i.e.,

$$F := (1-e)c + e\big[m^2 + 2m(1-m)(1-p) + (1-m)^2 c\big]. \qquad (‡)$$

The results, for $e = 0.2$ and $p = 0.5$, are given in Table 5: the light grey column is $c = p^2 + (1-p)^2 = 0.5$ and matches the column $p = 0.5$ in the second block of Table 4 (also marked in light grey).

Not surprisingly, the values of Cohen's κ are increasing quickly as c increases, but we also observe that a high degree of cohesiveness minimises the effect that the parameter m has on the value of κ; as a consequence, it becomes even harder to detect epistemic exceptionality.

Cohesion effects like this are expected in general in most academic disciplines; unpublished work of Greiffenhagen (2021) suggests that these effects are stronger in mathematics than in other disciplines.

The numerical values that were obtained by Geist et al. (2010) for conferences in the deductive sciences were considerably higher than the values in clinical neuroscience and information science and also considerably higher than would be predicted by our toy model. One possible explanation for these findings could be that we observe the strong(er) cohesion effect in mathematics at work. Making this precise and giving an argument that this is the case would require a detailed and comparative empirical study of the mathematical peer review process.

Funding Open Access funding enabled and organized by Projekt DEAL.

References

Abby M, Massey MD, Galandiuk S, Polk HC Jr (1994) Peer review is an effective screening process to evaluate medical manuscripts. J Am Med Assoc 272:105–107

Andersen LE (2017) On the nature and role of peer review in mathematics. Account Res 24:177–192

Azzouni J (2004) The derivation-indicator view of mathematical practice. Philos Math (III) 12:81–105

Bangdiwala SI (1985) A graphical test for observer agreement. Bull Int Stat Inst 51(1):307–308. (Proceedings of the 45th session of the International Statistical Institute, Vol. 1. Held in Amsterdam, August 12-22, 1985)

Birney E (2021) Opinion: scientific disagreements aren't new to covid. European Molecular Biology Laboratory. (EMBL Communications, News, 5 January 2021)

Buldt B, Löwe B, Müller T (2008) Towards a new epistemology of mathematics. Erkenntnis 68:309–329

Carter J (2019) Philosophy of mathematical practice-motivations, themes and prospects. Philos Math (III) 27:1–32

Cicchetti DV, Feinstein AR (1990) High agreement but low kappa: II. Resolving the paradoxes. J Clin Epidemiol 43:551–558

Cleary JD, Alexander B (1988) Blind versus nonblind review: survey of selected medical journals. Drug Intell Clin Pharm 22:601–602

Cohen J (1960) A coefficient of agreement for nominal scales. Educ Psychol Meas 20:37–46

Feinstein AR, Cicchetti DV (1990) High agreement but low kappa: I. The problems of two paradoxes. J Clin Epidemiol 43(6):543–549

Feuerman M, Miller AR (2005) The kappa statistic as a function of sensitivity and specificity. Int J Math Educ Sci Technol 36(5):517–527

Feuerman M, Miller AR (2008) Relationships between statistical measures of agreement: sensitivity, specificity and kappa. J Eval Clin Pract 14:930–933

Feyerabend PK (1975) Against method. Verso, London

François K, Van Bendegem JP (2010) Revolutions in mathematics. More than thirty years after Crowe's "Ten Laws''. A new interpretation. In: Löwe B, Müller T (eds) PhiMSAMP. Philosophy of mathematics: sociological aspects and mathematical practice, Texts in Philosophy, vol 11. College Publications, London, pp 107–120

Geist C, Löwe B, Van Kerkhove B (2010) Peer review and knowledge by testimony in mathematics. In: Löwe B, Müller T (eds) PhiMSAMP. Philosophy of mathematics: sociological aspects and mathematical practice, Texts in Philosophy, vol 11, pp. 155–178

Greiffenhagen C (2021) The disciplinary specificity of peer review: 'quick opinions' in mathematics. (Presentation at Mathematical Cultures Practices XI, 11 July 2021)

Gwet KL (2010) Handbook of inter-rater reliability. The definitive guide to measuring the extent of agreement among raters, 2nd edn. Advanced Analytics, Oxford

Hamami Y, Morris R (2020) Philosophy of mathematical practice: a primer for mathematics educators. ZDM Math Educ 52(6):1113–1126

Hanson NR (1958) Patterns of discovery: an inquiry into the conceptual foundations of science. Cambridge University Press, Cambridge

Holley JW, Guilford JP (1964) A note on the G index of agreement. Educ Psychol Meas 72:323–327

House of Commons Science and Technology Committee (2011) Peer review in scientific publications. Eighth report of session 2010–2012. Volume II. Additional written evidence. The Stationery Office

Inglis M, Mejía-Ramos JP (2009) The effect of authority on the persuasiveness of mathematical arguments. Cogn Instr 27:25–50
Inglis M, Mejía-Ramos JP, Weber K, Alcock L (2013) On mathematicians' different standards when evaluating elementary proofs. Top Cogn Sci 5:270–282
Justice AC, Cho MK, Winker MA, Berlin JA, Rennie D (1998) Does masking author identity improve peer review quality? A randomized controlled trial. J Am Med Assoc 280:240–242
Krantz SG (1997) A primer of mathematical writing: being a disquisition on having your ideas recorded, typeset, published, read, and appreciated. American Mathematical Society, Providence
Kuhn TS (1962) The structure of scientific revolutions. University of Chicago Press, Chicago
Kuhn TS (1977) Objectivity, value judgment, and theory choice. In: Kuhn TS (ed) The essential tension: selected studies in scientific tradition and change. University of Chicago Press, Chicago, pp 320–339
Landis JR, Koch GG (1977) The measurement of observer agreement for categorical data. Biometrics 33:159–174
Löwe B (2018) Philosophy or not? The study of cultures and practices of mathematics. In: Ju S, Löwe B, Müller T, Xie Y (eds) Cultures of mathematics and logic. Selected papers from the conference in Guangzhou, China, 9–12 November 2012. Trends in the History of Science Birkhäuser, pp 23-42
McNutt RA, Evans AT, Fletcher RH, Fletcher SW (1990) The effects of blinding on the quality of peer review. A randomized trial. J Am Med Assoc 263:1371–1376
Müller-Hill E (2011) Die Rolle formalisierbarer Beweise für eine philosophische Theorie mathematischen Wissens. Formalisierbarkeitsorientierte Konzeptionen mathematischen Wissens und mathematischer Rechtfertigung innerhalb einer sozio-empirisch informierten Erkenntnistheorie der Mathematik (doctoral dissertation). Rheinische Friedrich Wilhelms-Universität Bonn
Nathanson MB (2008) Desperately seeking mathematical truth. Not Am Math Soc 55(7):773
Nylenna M, Riis P, Karlsson Y (1994) Multiple blinded reviews of the same two manuscripts: effects of referee characteristics and publication language. J Am Med Assoc 272:149–151
Prediger S (2006a) Mathematics-cultural product or epistemic exception? In: Löwe B, Peckhaus V, Räsch T (eds) Foundations of the Formal Sciences IV. The history of the concept of the formal sciences. Studies in Logic, vol 3 College Publications, London, pp 271–272
Prediger S (2006b) Wider den Mythos der epistemischen Sonderstellung der Mathematik-Platonische Ideen oder Produkte einer kulturellen Praxis? Erwägen-Wissen-Ethik 17(3):384–386
Reiss J, Sprenger J (2020) Scientific objectivity. In: Zalta E (ed) Stanford encyclopedia of philosophy. (Winter 2020 Edition)
Rothwell PM, Martyn CN (2000) Reproducibility of peer review in clinical neuroscience is agreement between reviewers any greater than would be expected by chance alone? Brain 123(9):1964–1969
Scott WA (1955) Reliability of content analysis: the case of nominal scale coding. Public Opin Q XIX:321–325
Sokal AD (1996) Transgressing the boundaries: toward a transformative hermeneutics of quantum gravity. Soc Text 46–47:217–252
Sokal AD, Bricmont J (1997) Impostures intellectuelles. Éditions Odile Jacob
Tanswell F (2015) A problem with the dependence of informal proofs on formal proofs. Philos Math (III) 23(3):295–310
Weber K, Czocher J (2019) On mathematicians' disagreements on what constitutes a proof. Res Math Educ 21(3):251–270
Weber K, Mejía-Ramos JP (2015) The contextual nature of conviction in mathematics. Learn Math 35(2):9–14
Wood M, Roberts M, Howell B (2004) The reliability of peer reviews of papers on information systems. J Inf Sci 30:2–11