Peer review and knowledge by testimony in mathematics

Christian Geist[1], Benedikt Löwe[1,2,3] and Bart Van Kerkhove[4,*]

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

[2] Department Mathematik, Universität Hamburg, Bundesstrasse 55, 20146 Hamburg, Germany

[3] Mathematisches Institut, Rheinische Friedrich-Wilhelms-Universität Bonn, Endenicher Allee 60, 53115 Bonn, Germany

[4] Centre for Logic and Philosophy of Science, Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussel, Belgium

E-mail: cgeist@gmx.net; bloewe@science.uva.nl; bart.van.kerkhove@vub.ac.be

1 Introduction

Mathematics has been called an "epistemic exception" with a type of knowledge being categorically more secure than that of other sciences (Heintz, 2000; Prediger, 2006). At the other end of the epistemological spectrum, we have the whimsical "knowledge by testimony", disputed by some (cf. § 2) to be knowledge at all.

In this paper, we shall discuss two closely related question fields spanning the gap between these two epistemological extremes:

1. If mathematical knowledge is categorically different from other types of knowledge, and if the published mathematical research papers are a part of the written codification of this knowledge, then the level of certainty of claims made in the published literature should be higher than in other scientific disciplines. Is this true? And if so, how is this higher level of certainty achieved? A piece of mathematical text becomes part of the published literature by means of going through the process of peer review. Is mathematical peer review different from peer review in other disciplines?

2. Mathematicians refer to the published literature, sometimes without checking the proofs themselves. This is a form of knowledge by testimony; so how can the epistemic exception of mathematics survive if some of the proofs rely on pointers to the literature?

*The second and third author should like to thank the *Wissenschaftliches Netzwerk* PhiMSAMP funded by the *Deutsche Forschungsgemeinschaft* (MU 1816/5-1) for travel support.

Benedikt Löwe, Thomas Müller (*eds.*). *PhiMSAMP. Philosophy of Mathematics: Sociological Aspects and Mathematical Practice*. College Publications, London, 2010. Texts in Philosophy 11; pp. 155–178.

Received by the editors: 2 October 2009; 27 December 2009.
Accepted for publication: 11 January 2010.

A simple and naïve answer to both questions would be that the deductive nature of mathematics allows referees to check correctness of the proofs of published papers with absolute certainty, and thus the written codification of mathematical knowledge is certain knowledge, relieving us of any qualms about referring to it. However this is very far from the truth; in his opinion piece published in the *Notices of the American Mathematical Society*, Nathanson (2008) paints a dark picture of the mathematical refereeing process:

> 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.

Given that mathematical correctness of a paper is so important for the decision of whether a paper should be published or not, it might come as a surprise that there have been no studies of the mathematical refereeing process. In other fields, in particular in the medical sciences, the refereeing process is heavily scrutinized; research on the effect of peer review on the quality of papers (Goodman et al., 1994; Pierie et al., 1996; Roberts et al., 1994), on indicators for good referees (Evans et al., 1993; Black et al., 1998; Callaham et al., 1998; Nylenna et al., 1994), on referee bias (Link, 1998), on instruments that help to improve the quality of the refereeing process (Das Sinha et al., 1999; Garfunkel et al., 1990; Feurer et al., 1994), and on the question of blinding author identities and referee identities in the process (Walsh et al., 2000; Justice et al., 1998; Cleary and Alexander, 1988; Katz et al., 2002; McNutt et al., 1990; Fisher et al., 1994) abound in the medical and biological literature. In these fields, we find much more explicit rules for what is expected of the refereeing process and the individual referees than in mathematics (cf. Footnote 3).

An important *caveat* is in place here: measuring the quality of the refereeing process requires definitions of quality criteria for papers, referee reports, and referees; e.g., if you want to know whether the refereeing process improves the quality of papers, you first need to give a gauge for this quality. Let us give one such example: in (Abby et al., 1994), the authors conclude that the peer review process is successful in monitoring quality as "rejected manuscripts often were not published in other indexed medical journals". Definitions like this are circular: instead of measuring properties of the rejected papers, they rather measure whether the process is homogeneous across journals, using the refereeing process of other journals as a gauge

for the quality of the refereeing process at a given journal. These methodological issues are discussed in the meta-study (Jefferson et al., 2002); we shall not discuss them in depth in the given paper, but they will form the background of our discussions with empirical data in the later sections.

In this paper, we shall give a description of the mathematical refereeing process and its role in ascertaining that only correct results are published. In § 2, we give a brief overview of the discussion of testimony in epistemology before moving on to discussing the uses of trust and reference without checking details of proof in mathematical research in § 3.

In § 4, we give a schematic description of the mathematical refereeing process based on personal experience and a number of text sources. The description in this section is not based on any empirical research, but collects anecdotal data. For the next two sections, §§ 5 and 6, we then move to empirical data: in § 5, we give the results of a questionnaire that we sent to editors of mathematical journals with questions about the refereeing process; in § 6, we compare the results of a study on referee agreement in the neurosciences and information sciences to similar results for more mathematical conferences (actually, most of our examples are from theoretical computer science). Our empirical results can only be a very first step for understanding the mathematical refereeing process. The results of § 6 seem to indicate that there is in fact a higher degree of referee agreement in mathematical fields than in others. Further approaches and future work are discussed in our concluding § 7.

2 Knowledge by testimony

In daily life as in science, we heavily depend on reports by others. Inescapable as this may be, it nevertheless opens a deep epistemological problem. In principle, holding beliefs on the mere assertion by someone else is a precarious affair; by choosing to rely on the authority of other agents, we put ourselves at their epistemic mercy. Taking into account its ubiquity, it might surprise that testimony has only recently become a topic of major philosophical scrutiny. As a consequence, debates are still wide open, e.g., in the existing literature, there is no consensus over what exactly is to count as a proper instance of testimony and what not. Further, there are a number of complicating factors hampering a good assessment, e.g., what is the nature of the relationship between speaker and hearer, possibly illuminating or obscuring the latter's judgement. For an overview of the current debate, cf. (Adler, 2008).

Even when limiting ourselves to cases "of simple informational exchange over easily known matters, where there is little or no motivation to deceive" (Adler, 2008, § 1), serious epistemological questions remain. As the essence of the matter before us is that checking propositions for oneself is impossible,

other elements of the testimonial setting have to supply us with evidence for the trustworthiness of what is asserted.

A *default rule* for dealing with testimony is to accept assertions unless one has a special reason not to:

> Otherwise put, as long as there is no available evidence *against* accepting a speaker's report, the hearer has no positive epistemic work to do in order to justifiedly accept the testimony in question. (Lackey and Sosa, 2006, p. 4)

Limiting ourselves to the simple conversational exchanges mentioned above, in 'normal conditions', mostly no such defeaters apply. On the basis of past experience, we thus accept claims of others in these cases. This results in a large degree of uniformity in (justly) relying on testimony for situations of daily life. But also in science, having left behind 'gentlemen's culture', with peer review and reproducibility demands now seemingly constituting a system of organized skepticism rather than trust, things are at most different at the surface.[1] In this paper, we are illustrating this point with a particular view towards mathematics.

There are two major philosophical positions with respect to knowledge by testimony, the *anti-reductionist* or *non-inferentialist* stance and the *reductionist* or *inferentialist* stance. The anti-reductionists treat testimony as a fundamental source of knowledge, requiring no further justification beyond its apparent success. In this, testimony would be akin to, e.g., perception or memory. Contemporary discussions trace back to Reid (1969) and Coady (1973). Reductionists on the contrary deny that testimonial knowledge can be basic in the sense just specified, as it epistemically depends upon —and thus should always be inductively derived from— other resources, most notably sense perception and memory. In other words, reductionists demand positive reasons, not just the lack of defeaters, for accepting testimonial reports: *nullius in verba*, as the motto of the Royal Society reads. The prototype of such a thinker is David Hume.

3 Mathematical research based on trust

How much of mathematical research practice is based on testimony? Traditionalists will claim that if indeed mathematical research has proceeded on the basis of testimony, this reliance was and is *in principle* removable, i.e., all mathematicians can go through any proof in question and do it for themselves.[2] We know a substantial number of mathematicians who want

[1] Cf. (Lipton, 1998, p. 1): "Science is no refuge from the ubiquity of testimony. At least most of the theories that a scientist accepts, she accepts because of what others say".

[2] As an illustrative anecdote, let us report that an American mathematical logician teaches his graduate students to read mathematical papers as follows: read the statement

to understand all proofs that form a part of their papers and who will reprove even classical statements to be completely sure of their own results based on them; but we also know that many mathematicians are not as meticulous and accept results from the published literature as black boxes in their own research. Many mathematicians tend to trust the experts and (in Auslander's words) "[t]his is the case even if we haven't read the proof, or more frequently when we don't have the background to follow the proof." (Auslander, 2008, p. 64)

The fact that written mathematical proofs are not complete formal derivations is acknowledged by many authors dealing with the epistemology of mathematics. Fallis (2003) discusses gaps in mathematical proofs; some of them are *enthymematic gaps* where the author has checked all details and omits them from the published paper for reasons of style or brevity (cf. (Heintz, 2000, p. 170), where the author compares the original set of notes written by Hirzebruch with the much terser final publication); others are what Fallis calls an *untraversed gap*:

> A mathematician has left an untraversed gap whenever he has not tried to verify directly that each proposition in the sequence of propositions that he has in mind (as being a proof) follows from previous propositions in the sequence by a basic mathematical inference. (Fallis, 2003, pp. 56–57)

Fallis notes that "there are [...] cases where it is considered acceptable for a mathematician to leave an untraversed gap" (Fallis, 2003, p. 58). In general, research mathematicians agree with Fallis's observation. Referring to Almgren's proof that establishes the regularity of minimizing rectifiable currents up to codimension two, Hales writes:

> The preprint is 1728 pages long. Each line is a chore. He spent over a decade writing it in the 1970s and early 1980s. It was not published until 2000. Yet the theorem is fundamental. [...] How am I to develop enough confidence in the proof that I am willing to cite it in my own research? Do the stellar reputations of the author and editors suffice, or should I try to understand the details of the proof? I would consider myself very fortunate if I could work through the proof in a year. (Hales, 2008, pp. 1370–1371)

Nathanson (in the cited opinion piece in the *Notices of the American Mathematical Society*) is more critical of this described practice:

> Many great and important theorems don't actually have proofs. They have sketches of proofs, outlines of arguments, hints and intuitions

of the theorem, cover its proof with a sheet of paper, and then try and prove the theorem yourself.

> that were obvious to the author (at least, at the time of writing) and that, hopefully, are understood and believed by some part of the mathematical community. But the community itself is tiny. In most fields of mathematics there are few experts. [...] In every field, there are 'bosses' who proclaim the correctness or incorrectness of a new result, and its importance or unimportance. Sometimes they disagree, like gang leaders fighting over turf. In any case, there is a web of semi-proved theorems throughout mathematics. (Nathanson, 2008)

4 The mathematical refereeing process

We hope to have convinced the reader in § 3 that there are at least some relevant instances of references to testimony in mathematical research: even if some mathematicians check meticulously whether all of the theorems that their results depend on are correct, not all do it, and so, if a research mathematician uses a theorem from the literature, the correctness of the result depends not only on the accuracy of the refereeing process of the paper he or she uses, but also on the refereeing processes of the papers used by that paper, and so on.

So, if mathematicians are relying on these iterated refereeing processes, how much security does the refereeing process in mathematics actually generate? In this section, we shall describe the mathematical refereeing process. There is hardly any (systematic) discussion about this topic[3] and therefore, our description here is largely based on the guidelines given in the excellent books by Steven Krantz (1997, 2005) on mathematical writing and publishing and the personal experience of the second author of the present paper as an author, referee, journal editor, and book editor of mathematical papers.

In mathematics, papers are mostly published in journals; conferences and their proceedings volumes play a subordinate role. In the case of journal

[3]Cf. (Auslander, 2008, p. 65): "The issue of the refereeing process —real and ideal— in mathematics is fascinating and largely unexplored. Gossip on this topic abounds but I know of no systematic study."

We should mention that there are meta-discussions about the refereeing process as part of the discussions about major changes of the mathematical publishing process: some mathematicians do not like the role of commercial publishers in the publishing process, and would like to replace the current process with a web-based alternative, leading to changes also to the refereeing process; cf. (Birman, 2000; Jackson, 2002, 2003; Borwein et al., 2008). However, these discussions rarely touch the epistemological issues relevant here, and so we shall not discuss them further.

We should also like to mention the highly interesting case study (Weintraub and Gayer, 2001) in which the authors analyse the refereeing process of a particular result from mathematical economics with access to the reports and the paper (Thompson, 1983) in which an author displays the history of the rejection of one of his papers, raising "questions about the role of the referee in the professional development of a mathematician" (Thompson, 1983, p. 661).

publishing, the journal editor typically asks a single referee to write a report on a given submission; these referees are experts and have often published on material very closely related to the material in the submission.[4] Often, the referee is personally known to the editor, allowing the editor to read between the lines of the report. It is not standard practice to give referees many instructions apart from a deadline:[5] it is understood that referees know what is expected of them. Somewhat in contrast to our empirical findings of § 5, it is accepted by authors that the refereeing process takes more than six months. Many authors consider it inappropriate to remind an editor before six months after submission have passed, and some even do not ask about the status of their papers before a year has passed.

Ideally, referee reports "should address Littlewood's three precepts: (1) Is it new? (2) Is it correct? (3) Is it surprising?" (Krantz, 1997, p. 125). The level of detail of referee reports varies a lot: many reports are very short (less than one page of text), but some can be very long, sometimes longer than the submission itself. Typically, even for longer reports, the core of the report (the statement of the recommendation and the argument for this recommendation) is rather short, and the bulk of the report consists of detailed comments to be considered for revisions. Reports recommending rejection tend to be much shorter, sometimes only a few lines.[6] In general, it seems fair to say that the default decision for mathematical journals is *reject*: in order for an editor to accept a paper for a mathematical journal, the referee has to give arguments supporting acceptance. Editors will only very rarely overrule a referee's recommendation to reject.[7]

Mathematicians disagree about the amount of detail checking that has to be done by the referees. While some (few) mathematicians think that checking the correctness of the proofs is the main task of the referee, others disagree with this and consider mathematical correctness the problem of the author rather than that of the referee.[8] Methodologically, this is an

[4] "There are several parameters to consider [to find a good referee]: (i) the referee should be an expert in the subject area, (ii) the referee should be dependable, (iii) the referee should not be prejudiced, (iv) the referee should be someone who can get the job done." (Krantz, 2005, p. 119)

[5] "[The] guidelines [...] may certainly suggest a time frame for the refereeing process. [...] It is not often that the instructions to the referee will give detailed advice on what points to address." (Krantz, 2005, p. 121)

[6] Krantz writes that "[a] typical referee's report is anywhere from one to five pages (or, in rare instances, even more)." (Krantz, 1997, p. 125).

[7] This may be a special situation in the mathematical review process. In his account of the peer review process, Gross reports that "editors generally assume that the rejection of a paper depends on a clear negative decision on the part of both referees; a split decision ordinarily favors the authors" (Gross, 1990, p. 134). However, in an endnote commenting on this statement, Gross says: "It is on this point that journals in the humanities deviate most; in the case of split decisions they are inclined to reject." (Gross, 1990, p. 215)

[8] Cf. (Auslander, 2008, p. 65): "[S]tandards of refereeing vary widely. Some papers [...]

important issue: the anonymity of the referee means that the reader of a paper does not know which type of refereeing treatment the paper has received before acceptance.

To conclude this section with illustrating statements of mathematicians, let us give two excerpts from an interview study performed by Eva Müller-Hill (2010). This study provided a qualitative extension of the quantitative work reported on in (Müller-Hill, 2009; Löwe et al., 2010) in which mathematicians were supposed to assess the knowledge of protagonists in a fictitious story about mathematical proofs. The interviews were not primarily concerned with the mathematical refereeing process; we give two excerpts (extensively rewritten and sometimes reworded in grammatical English from the original transcript of spoken English) that are relevant here and corroborate the positions of Auslander and Krantz:[9]

> Let's say a famous mathematician comes up with a paper, and I have to referee it. Then I am preoccupied with the fact that he is a very well known mathematician, and so that it probably will be ok. And then, you say "yes, this really seems plausible, but I'm not really sure if it's true" and you end up with the question "is this because I don't have enough knowledge?" And then there's time pressure and you have other things to do when they ask you to referee this 50 pages paper. Then you have a tendency of believing that it is correct, and you think "he's publishing it, not I, so it's his responsibility that it is correct".

The same interviewee had to comment on the story about a fictitious world-famous expert named Jones who proved a result, submitted it, published it after a refereeing process, only to find out a few years later that the main result is wrong:

> It depends on a lot of things. Firstly, Jones is a world famous expert, so this means she's teaching at a university like Harvard. Then the paper was sent to a mathematical journal of high reputation, so,

concern famous problems, and thus have received intense scrutiny. Other papers receive more routine treatment". Auslander also reports that "referees are generally told that it is not their job to determine whether a paper is correct — this is the responsibility of the author — although the referee should be reasonably convinced. The referee is typically asked to determine whether the paper is worthwhile" (Auslander, 2008, p. 65). This is reflected in Krantz's recommendation to the referee: "While you may not have checked every detail in the paper, you should at least be confident of your opinion as to the paper's correctness and importance" (Krantz, 1997, p. 125). Cf. also our survey results in § 5.2.

[9] We should like to thank Eva Müller-Hill for the permission to include these examples. We reworded the texts from the transcripts to give the quotations the flow necessary for written texts while preserving the content and style of the original utterances. The literal quotes will be contained in Müller-Hill's dissertation (2010).

> say, *Acta Mathematica*; this tells us something about the size of the mathematical community involved. You cannot be a world famous expert on something that nobody else does. That it went to a good journal means that the journal thought of looking for good referees, so it was established more surely than that it would have been sent to the journal of a tiny mathematical society with very few members. This puts the scenario in a framework which makes it very likely that the result is correct. And of course it happens that things are not right.

5 A survey of mathematical journal editors

The description of the mathematical refereeing process given in § 4 was based on personal experience and the descriptions in (Krantz, 1997, 2005; Auslander, 2008). In the spirit of *Empirical Philosophy of Mathematics*, we aim to corroborate this personal account with empirical data.

In March 2009, we selected 27 editors of mathematical journals: 9 editors each of (what we estimated to be) top, mid or lower level journals. During the month after that, we received 13 answers (4 from top journal editors, 4 from mid level journal editors and 5 from lower level journal editors). Our modest questionnaire aimed at getting opinions of what the refereeing process is about; the questionnaire can be found in Figure 1.

5.1 Question 1. Importance of referee's tasks.

Overall, the editors agreed that Littlewood's precepts (cf. p. 161) are important. On the scale from 1 to 5, novelty gets an average score of **4.7**, correctness a score of **4.5**, and interest a score of **4.3**. The importance of whether the paper is well-written only gets an average score of **3.5**. The general tendency is that editors of higher-ranked journals give higher scores to all of the categories than editors of lower-ranked journals. The biggest difference between top-ranked journals and lower-ranked journals was in the category "is it interesting". One editor sums up our averaged findings as follows:

> [T]o be published an article must be novel (but new, enlightening proofs of older central results are publishable in exceptional cases), correct and interesting. In the process of refereeing, we try to improve the writing.

5.2 Question 2. Checking of proofs.

We received eleven answers to this question: six editors thought that the referee should check all proofs in detail; five thought that the referee should check some proofs in detail. Option (c) was not selected by anyone. As a realistic side remark, one of the editors who checked (a) wrote "but to be

1. Rate the importance of the following tasks of a referee, from 1 (not at all) to 5 (most certainly):
   (a) checking the correctness of results
   (b) estimating the novelty of results
   (c) judging whether the paper is interesting
   (d) judging whether the paper is well-written
2. Pick one answer out of the following:
   (a) I think the referee should check all proofs in detail.
   (b) I think the referee should check some proofs in detail.
   (c) I think the referee should check none of the proofs in detail.
3. How many weeks approximately do you grant a referee to write a report for an average 20-page research paper?
4. How many hours approximately do you expect a referee to spend on checking the correctness of a paper's claims?
5. What percentage of referees approximately do a good job checking the correctness of a paper's claims?

FIGURE 1. The questionnaire sent to 27 editors of mathematical journals in March 2009.

reasonable, I am happy when I find a referee doing (b)." One of the editors who did not provide an answer described a slightly non-standard procedure:

> We actually work with several referees for a given article; first an overview referee checking novelty, interest, [and] correct references. [Then, after that] if we [...] feel the paper is interesting and new, a[nother] referee [who] checks [the proofs] for correctness.

5.3 Question 3. Overall time for refereeing.

The average amount of time that our editors give their referees is about 14 weeks. The period of 3 months seems to be a standard expected length of the refereeing process (six out of thirteen responses), but two, four, and six months also occurred as answers.

5.4 Question 4. Time spent on checking correctness.

The answers to this question varied widely (from 5 to 80 hours), and a large number of the respondents refused to answer it on grounds that it depends too much on the individual paper.

5.5 Question 5. Quality of referees.

The overall average to the question how many referees do a good job checking the correctness is **52.3%**. There is a marked difference along the lines of the ranking of the journals: among the editors of top-level journals, the estimate was 61.3%, among the editors of mid-level journals, the estimate was 42.5%, and the answers in the lower-ranked journals differed too much to give a meaningful average.[10]

5.6 Some additional results.

In addition to the answers to the five questions, we received some comments that confirm parts of the description from § 4. One of the editors, when asked about the quality of the job of the referees, answered

> After [many years], I know many colleagues convenient for refereeing papers. I am sending, if possible, papers to be refereed only to those who are responsible and I can believe they do a good job. Exceptionally, I must send a paper to a [different] person.

Also the following quotation from an editor reinforces our statements from § 3 that mathematics is largely built on trust:

> There are situations where almost nothing needs be checked (e.g., the results come from a seminar where the results were checked, or I see the paper is not too good and then it is useless to check details, or the author is well-known and it is his concern to submit a correct paper). There are situations when I insist to check all the procedures (e.g., when it concerns good results from a less known author).

It is interesting to see the criteria according to which this editor decides that "almost nothing needs to be checked": it is the reputation of the author that drives decisions about how much detail has to be checked.

6 Quantitative data for conference refereeing in theoretical computer science

In § 1, we mentioned that there is a large body of research on the refereeing process in the natural and medical sciences, for instance on the question of reliability. One indicator for reliability or objectivity of the process is whether referees agree in their judgments of refereed papers. This was investigated, e.g., by Rothwell and Martyn (2000) who considered the question whether the agreement of referees is greater than it would be by mere chance on the basis of data from journals in the neurosciences, and a similar

[10] It is interesting to compare this to the quote from Müller-Hill's interview study (already mentioned on p. 163): "That it went to a good journal means that the journal thought of looking for good referees, so it was established more surely than that it would have been sent to the journal of a tiny mathematical society with very few members."

study by Wood, Roberts and Howell (2004) in the information sciences. In this section, we aim at providing the same analysis for the mathematical refereeing process.

In our description of the mathematical refereeing process in § 4 (cf. also § 5.6), we stressed that the mathematical peer review is largely a communication between an editor and one referee based on trust due to a close personal relationship. The question of referee agreement does not make sense in this situation. Therefore, we had to leave the immediate area of journal refereeing in mathematics and move to the cognate area of conference refereeing in theoretical computer science.

Theoretical computer science, as mathematics, is largely based on the deductive method, and its main results are mathematical theorems. Therefore, the epistemic character of results in theoretical computer science is comparable to that of results in mathematics. On the other hand, computer science has a rather different publication culture from mathematics. While journal publications in theoretical computer science follow the mathematical refereeing process described in § 4, computer science developed a distinctive culture of refereed conference publications. The highest-ranking conferences clearly outrank some of the good journals of the field in terms of reputation. Due to the time pressure of the production schedule, the refereeing process here is markedly different from the mathematical refereeing process.

The following description is not meant to be an empirically verified description of the refereeing process in theoretical computer science, but a generalized personal account of the second author, based on his experience in programme committees for conferences in this field. Some of the claims (e.g., the ones on the tendencies about the length and depth of the referee reports) certainly will require an empirical and methodologically clean analysis in the future.

A computer science conference has a programme committee that is responsible for the selection of papers. Papers are submitted to a conference about half a year before the conference; after the submission deadline for the conference, the chair of the programme committee assigns papers to members of the programme committee, often after a phase of bidding during which the members can announce their preferences for being assigned certain papers. Typically, each paper is assigned to three or four members of the programme committee who then take a role similar to that of an editor in the mathematical peer review process. An assigned member of the programme committee can either decide to write a referee report herself or find a so-called *subreferee* who will write a referee report. Typically, the referee has three to four weeks to complete the report. Reports come with a numerical score on a scale fixed in advance by the chairs of the programme

committee and tend to be shorter than reports in the mathematical journal refereeing process described in § 4: some can be as detailed as in the journal peer review, but others can just consist of a couple of lines. While it is preferred that the referees check the mathematical details, it is acceptable to submit a referee report stating "I did not check the details", "I did not have the time to check the details" or even "I am not an expert in the area and didn't follow the proofs".

After the referee reports are in, the chairs initiate the so-called *PC Session*, a time period of about a week during which the members discuss the papers and referee reports electronically. During this period the chairs moderate the discussion of the members, propose to accept certain papers and reject others, and in some cases announce votes on particular decisions. During the session, it is not uncommon to ask subreferees for more clarification and further comments on their reports.

In past years, the refereeing process for computer science conferences has been uniformized strongly by the use of conference submission software. An important system used is the EasyChair system of Andrei Voronkov which was used for 1313 conferences in 2008 and for 2186 conferences in 2009.

6.1 Available data

From the papers (Rothwell and Martyn, 2000; Wood et al., 2004) serving as our comparative data set from non-mathematical areas of science, we obtained the data on two journals in the neurosciences (neuro1 and neuro2) and two conferences in the information sciences (infor1 and infor2). In these four cases, the majority of refereed papers had two referees, and the studies (Rothwell and Martyn, 2000; Wood et al., 2004) reduced their data set to the subset of those submissions. The journals neuro1 and neuro2 used the categories "accept", "accept after revision", "reject"; the conferences infor1 and infor2 used "accept", "accept with minor revisions", "accept with major revisions", and "reject".[11]

We obtained anonymized data for eight conferences for which the refereeing was organized via the EasyChair system. One of these conferences was purely mathematical (math1), the others were in theoretical computer science (comp1 to comp7). In the following, we shall call the original data from (Rothwell and Martyn, 2000; Wood et al., 2004) *non-mathematical conferences* and our new data *mathematical conferences*.

Each of the conferences used a slighly different grading scales, and we decided to use a three-category scale 1 ("accept"), 0 ("accept with revisions" or "borderline"), −1 ("reject") as the most natural common grading scale.

[11] Note that while the data of (Wood et al., 2004) had four categories, the analysis was done in terms of two categories.

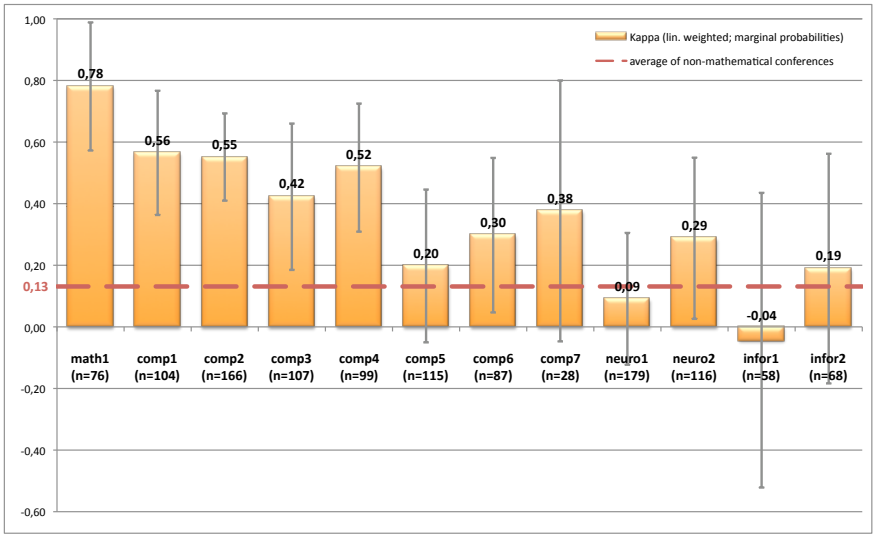

FIGURE 2. Comparison of Kappa values (partly averaged; linearly weighted; using marginal probabilities) to the average value of the non-mathematical conferences. The error bars show the induced 95% confidence intervals.

We needed to translate the various scales used into this three-category set-up.

The major difference between the data used by (Rothwell and Martyn, 2000; Wood et al., 2004) and our new data was that the mathematical conferences used variable numbers of referees (the minimum number was two and the maximum number was seven). We shall discuss how we dealt with this in § 6.2.

The data sets used for this calculation (i.e., the data translated to our three-category scale as described) can be found in Appendix B..

6.2 Method

Following (Rothwell and Martyn, 2000; Wood et al., 2004), we did our analysis by means of Kappa statistics. Cohen's Kappa is a standard measure for the analysis of inter-rater reliability (two raters) proposed in (Cohen, 1960). The Kappa value is scaled such that 0 represents the level of agreement that would have been expected by chance, and 1 represents perfect agreement between the raters. Various suggestions for improvement of the method of Kappa statistics have been made (Brennan and Prediger, 1981; Cohen, 1960; Sim and Wright, 2005): e.g., it is strongly recommended by Sim and Wright (2005) to use a weighted Kappa in the case of ordinal categories.

We decided to use a linearly weighted Kappa (making disagreement by one category count as half an agreement).[12] We did not use the results of the analysis from (Rothwell and Martyn, 2000; Wood et al., 2004) directly, but instead used their data and recalculated the Kappa values to make sure that we used exactly the same method for all conferences investigated.

As mentioned, the new conferences had a variable number of referees (between two and seven). Such a situation cannot be handled by Kappa analysis[13] and would make the comparison between conferences rather difficult. In order to carry out this type of analysis to compare it to the result obtained from the data in (Rothwell and Martyn, 2000; Wood et al., 2004), we had to choose two referees per article. We let a computer pick two different referees per paper uniformly at random and iterated this procedure 100 times; every time computing Kappa value as well as its 95% confidence interval as induced by the confidence intervals of the observed level of agreement, which were computed using the standard formula for a proportion on a 95% confidence level and applying the usual continuity correction. The values we report here are the arithmetic means of these 100 computations.[14]

6.3 Results

We have visualized the results of our analysis in Figure 2, which exhibits the Kappa values obtained (with 95% confidence intervals) for each conference (as bars) and compares them to the average Kappa of the non-mathematical conferences (dashed line at 0.13).

The values for Kappa range from -0.04 (infor1) up to 0.78 (math1) and all mathematical conferences have Kappa values that are strictly higher than the threshold of 0.13 (average Kappa of the non-mathematical conferences). Furthermore, with one exception, all mathematical conferences have strictly higher Kappa values than all four non-mathematical conferences.

7 Conclusion and future work

We started this paper with the question: If mathematics is an epistemic exception, shouldn't the mathematical literature be more reliable than that of other fields; and if so, how does the refereeing process contribute to this?

The answer given in this paper is somewhat ambiguous: we argued that a lot of mathematical research uses black boxes from the literature without checking the proofs, and claimed that there are serious issues with the reliability of the literature. Looking into the mathematical refereeing process, we saw that it is not universally expected that referees check the correctness

[12] The definition of the Kappa we used is given in Appendix A..

[13] Fleiss (1971) introduced a new version of Kappa in that is able to treat fixed larger numbers of raters (Fleiss' Kappa), but not variable numbers of raters.

[14] We attempted to avoid the randomization by choosing the extreme two raters in any given set, but this did not generate realistic results.

of all claims in the papers, and this was corroborated by our survey study in § 5. But the survey study also showed that editors have a lot of trust in their referees. In our empirical study in § 6, we saw a different facet of this: compared to other fields where referees do not come to the same conclusion much more often than they would by pure chance, the agreement between referees in fields based on the deductive method is higher, indicating (but not proving) that the degree of objectivity is higher.

We also saw that the empirical methods employed in § 6 do not really fit the core of the mathematical refereeing business: a more qualitative study is necessary, analogous to the field work of Greiffenhagen (2008) on the process of graduate student supervision in mathematics. A possible source of data could be the new and unusual journal *Rejecta Mathematica*, a journal that

> publishes only papers that have been rejected from peer-reviewed journals in the mathematical sciences [...] [together with] an open letter from its authors discussing the paper's original review process, disclosing any known flaws in the paper, and stating the case for the paper's value to the community. (Wakin et al., 2009, p. 1)

It remains to be seen how useful *Rejecta Mathematica* will be as a source for studying the mathematical refereeing process (the published letters from the authors in its inaugural issue do not give much insight in the original refereeing process). We should also like to mention the interesting paper (Weintraub and Gayer, 2001) (cf. Footnote 3): written by economists, it reports on how a mathematical result (the Arrow-Debreu theorem) became accepted by the whole community within a very short time, even though the refereeing process was anything but unproblematic. One of the reviewers, the mathematician Cecil G. Phipps, originally selected by the associate editor to "thoroughly check the mathematics of the argument",[15] strongly objected to the publication of the paper. But the editors decided to accept the paper on the basis of one very short referee report of just a few lines and four typographical corrections (Weintraub and Gayer, 2001, p. 430) and a detailed comment of the associate editor who, however, had decided not to do a "thorough checking of the mathematics".[16] Weintraub and Gayer raise interesting questions like "Did the persuasion occur before Arrow and Debreu submitted the article for publication at *Econometrica*?" (Weintraub and Gayer, 2001, p. 440). Case studies like this, either dealing with historical cases like the Arrow-Debreu theorem or with a refereeing process accompanied as it is going on, would be the natural next step.[17]

[15] From a letter by associate editor Nicholas Georgescu-Roegen to the managing editor Robert Strotz, dated 8 October 1953, quoted after (Weintraub and Gayer, 2001, p. 431).

[16] From the same letter; quoted after (Weintraub and Gayer, 2001, p. 434).

[17] Note that also Gross's study of the peer review process in his (Gross, 1990, § 9) is based on textual data from actual referee/editor and author/editor exchanges.

Bibliography

Abby, M., Massey, M. D., Galandiuk, S., and Polk Jr., H. C. (1994). Peer review is an effective screening process to evaluate medical manuscripts. *Journal of the American Medical Association*, 272:105–107.

Adler, J. (2008). Epistemological problems of testimony. In Zalta, E. N., editor, *Stanford Encyclopedia of Philosophy*. Fall 2008 Edition.

Auslander, J. (2008). On the roles of proof in mathematics. In Gold, B. and Simons, R. A., editors, *Proof & Other Dilemmas: Mathematics and Philosophy*, pages 61–77. Mathematical Association of America, Washington DC.

Birman, J. S. (2000). Scientific publishing: A mathematician's viewpoint. *Notices of the American Mathematical Society*, 47(7):770–774.

Black, N., van Rooyen, S., Goglee, F., Smith, R., and Evans, S. (1998). What makes a good reviewer and a good review for a general medical journal? *Journal of the American Medical Association*, 280:231–233.

Borwein, J. M., Rocha, E. M., and Rodrigues, J. F., editors (2008). *Communicating Mathematics in the Digital Era*. AK Peters, Natick MA.

Brennan, R. L. and Prediger, D. J. (1981). Coefficient κ: Some uses, misuses, and alternatives. *Educational and Psychological Measurement*, 41:687–699.

Callaham, M. L., Wears, R. L., and Waeckerle, J. F. (1998). Effect of attendance at a training session on peer reviewer quality and performance. *Annals of Emergency Medicine*, 32:318–322.

Cleary, J. D. and Alexander, B. (1988). Blind versus nonblind review: Survey of selected medical journals. *Drug Intelligence and Clinical Pharmacy*, 22:601–602.

Coady, C. A. J. (1973). Testimony and observation. *American Philosophical Quarterly*, 10:149–155.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37–46.

Das Sinha, S., Sahni, P., and Nundy, S. (1999). Does exchanging comments of Indian and non-Indian reviewers improve the quality of manuscript reviews? *National Medical Journal of India*, 12:210–213.

Evans, A. T., McNutt, R. A., Fletcher, S. W., and Fletcher, R. H. (1993). The characteristics of peer reviewers who produce good-quality reviews. *Journal of General Internal Medicine*, 8:422–428.

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

Feurer, I. D., J., B. G., Picus, D., Ramirez, E., Darcy, M. D., and Hicks, M. E. H. (1994). Evaluating peer reviews pilot testing of a grading instrument. *Journal of the American Medical Association*, 272:98–100.

Fisher, M., Friedman, S. B., and Strauss, B. (1994). The effects of blinding on acceptance of research papers by peer review. *Journal of the American Medical Association*, 272:143–146.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Garfunkel, J. M., Ulshen, M. H., Hamrick, H. J., and Lawson, E. E. (1990). Problems identified by secondary review of accepted manuscripts. *Journal of the American Medical Association*, 263:1369–1371.

Goodman, S. N., Berlin, J., Fletcher, S. W., and Fletcher, R. H. (1994). Manuscript quality before and after peer review and editing at annals of internal medicine. *Annals of Internal Medicine*, 121:11–21.

Greiffenhagen, C. (2008). Video analysis of mathematical practice? Different attempts to 'open up' mathematics for sociological investigation. *Forum: Qualitative Social Research*, 9(3):art. 32.

Gross, A. G. (1990). *The Rhetoric of Science*. Harvard University Press, Cambridge MA.

Hales, T. C. (2008). Formal proof. *Notices of the American Mathematical Society*, 55(11):1370–1380.

Heintz, B. (2000). *Die Innenwelt der Mathematik. Zur Kultur und Praxis einer beweisenden Disziplin*. Springer, Vienna.

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

Jackson, A. (2003). The digital mathematics library. *Notices of the American Mathematical Society*, 50(8):918–923.

Jefferson, T., Wager, E., and Davidoff, F. (2002). Measuring the quality of peer review. *Journal of the American Medical Association*, 287:2786–2790.

Justice, A. C., Cho, M. K., Winker, M. A., Berlin, J. A., and Rennie, D. (1998). Does masking author identity improve peer review quality? A randomized controlled trial. *Journal of the American Medical Association*, 280:240–242.

Katz, D. S., Proto, A. V., and Olmsted, W. W. (2002). Incidence and nature of unblinding by authors: Our experience at two radiology journals with double-blinded peer review policies. *American Journal of Roentgenology*, 179:1415–1417.

Krantz, S. G. (1997). *A Primer of Mathematical Writing. Being a Disquisition on having your ideas recorded, typeset, published, read, and appreciated*. American Mathematical Society, Providence RI.

Krantz, S. G. (2005). *Mathematical Publishing. A Guidebook*. American Mathematical Society, Providence RI.

Lackey, J. and Sosa, E., editors (2006). *The Epistemology of Testimony*. Oxford University Press, Oxford.

Link, A. M. (1998). US and non-US submissions: An analysis of reviewer bias. *Journal of the American Medical Association*, 280:246–247.

Lipton, P. (1998). The epistemology of testimony. *Studies in History and Philosophy of Science A*, 29(1):1–31.

Löwe, B., Müller, T., and Müller-Hill, E. (2010). Mathematical knowledge: A case study in empirical philosophy of mathematics. In Van Kerkhove, B., De Vuyst, J., and Van Bendegem, J. P., editors, *Philosophical Perspectives on Mathematical Practice*, volume 12 of *Texts in Philosophy*, pages 185–203. College Publications, London.

McNutt, R. A., Evans, A. T., Fletcher, R. H., and Fletcher, S. W. (1990). The effects of blinding on the quality of peer review. A randomized trial. *Journal of the American Medical Association*, 263:1371–1376.

Müller-Hill, E. (2009). Formalizability and knowledge ascriptions in mathematical practice. *Philosophia Scientiae*, 13(2):21–43.

Müller-Hill, E. (2010). *Die epistemische Rolle formalisierbarer mathematischer Beweise. Formalisierbarkeitsbasierte Konzeptionen mathematischen Wissens und mathematischer Rechtfertigung innerhalb einer sozioempirisch informierten Erkenntnistheorie der Mathematik*. PhD thesis, Rheinische Friedrich-Wilhelms-Universität Bonn.

Nathanson, M. B. (2008). Desperately seeking mathematical truth. *Notices of the American Mathematical Society*, 55(7):773.

Nylenna, M., Riis, P., and Karlsson, Y. (1994). Multiple blinded reviews of the same two manuscripts. Effects of referee characteristics and publication language. *Journal of the American Medical Association*, 272:149–151.

Pierie, J., Walvoort, H., and Overbeke, A. (1996). Readers' evaluation of effect of peer review and editing on quality of articles in the Nederlands Tijdschrift voor Geneeskunde. *Lancet*, 348:1480–1483.

Prediger, S. (2006). Mathematics: Cultural product or epistemic exception? In Löwe, B., Peckhaus, V., and Räsch, T., editors, *Foundations of the Formal Sciences IV. The History of the Concept of the Formal Sciences*, volume 3 of *Studies in Logic*, pages 271–232. College Publications, London.

Reid, T. (1969). *Essays on the Intellectual Powers of Man.* MIT Press, Cambridge MA.

Roberts, J. C., Fletcher, R. H., and Fletcher, S. W. (1994). Effects of peer review and editing on the readability of articles published in Annals of Internal Medicine. *Journal of the American Medical Association*, 272:119–121.

Rothwell, P. M. and Martyn, C. N. (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.

Sim, J. and Wright, C. C. (2005). The kappa statistic in reliability studies: Use, interpretation, and sample size requirements. *Physical Therapy*, 85(3):257–268.

Thompson, R. C. (1983). Author vs. referee: A case history for middle level mathematicians. *American Mathematical Monthly*, 90(10):661–668.

Wakin, M., Rozell, C., Davenport, M., and Laska, J. (2009). Letter from the editors. *Rejecta Mathematica*, 1(1):1–3.

Walsh, E., Rooney, M., Appleby, L., and Wilkinson, G. (2000). Open peer review: A randomized controlled trial. *British Journal of Psychiatry*, 176:47–51.

Weintraub, E. R. and Gayer, T. (2001). Equilibrium proofmaking. *Journal of the History of Economic Thought*, 23(4):421–442.

Wood, M., Roberts, M., and Howell, B. (2004). The reliability of peer reviews of papers on information systems. *Journal of Information Science*, 30:2–11.

Appendix A. Calculation of the Kappa value

The data sets we use are assigning two reviewer recommendations for each of a set of N papers; derived from this, we consider functions $\mathbf{d} : N \to \{-1, 0, 1\}$ picking out the first and the second reviewers recommendation from the data set. To simplify notation, we introduce the following functions on $\{-1, 0, 1\} \times \{-1, 0, 1\}$, representing “equality” and “distance 1”:

$$\delta_{x,y} := \begin{cases} 1 & x = y \\ 0 & \text{otherwise} \end{cases} \quad \text{and} \quad \xi_{x,y} := \begin{cases} 1 & |x - y| = 1 \\ 0 & \text{otherwise.} \end{cases}$$

We fix $\mathbf{a}, \mathbf{b} : N \to \{-1, 0, 1\}$ as given from our data set and write $\mathbf{c}_i$ for the constant function with value $i \in \{-1, 0, 1\}$. We then define

$$\Delta_{\mathbf{a},\mathbf{b}} := \sum_{i=0}^{N-1} \delta_{\mathbf{a}_i,\mathbf{b}_i},\ \Xi_{\mathbf{a},\mathbf{b}} := \sum_{i=0}^{N-1} \xi_{\mathbf{a}_i,\mathbf{b}_i},\ \text{and}$$

$$\eta_i := (\Delta_{\mathbf{a},\mathbf{c}_i} + \Delta_{\mathbf{b},\mathbf{c}_i})/2N.$$

Then the *observed agreement* is defined as

$$O_{\mathbf{a},\mathbf{b}} := \frac{1}{N}\Delta_{\mathbf{a},\mathbf{b}} + \frac{1}{2N}\Xi_{\mathbf{a},\mathbf{b}},$$

and the *expected agreement* as

$$E_{\mathbf{a},\mathbf{b}} := (\eta_{-1})^2 + (\eta_0)^2 + (\eta_1)^2 + \eta_1\eta_0 + \eta_0\eta_{-1}.$$

Finally, the Kappa value is given by

$$\kappa_{\mathbf{a},\mathbf{b}} := \frac{O_{\mathbf{a},\mathbf{b}} - E_{\mathbf{a},\mathbf{b}}}{1 - E_{\mathbf{a},\mathbf{b}}}.$$

Appendix B. The data used in § 6

A semicolon separates the different papers, whereas a comma separates the recommendations for a paper by different reviewers.

math1 ($n = 76$)

1, 1; 1, 1; 1, 1; 1, 1; 1, 1; 1, 1; 1, 1; 1, 1; 1, 1; 1, 0; 1, 0; 1, 0; 1, 0; 0, 0; 1, −1; 1, −1; 1, −1; −1, −1; −1, −1; −1, −1; −1, −1; −1, −1; −1, −1; −1, −1; −1, −1; −1, −1; −1, −1; −1, −1; −1, −1; −1, −1; −1, −1

comp1 ($n = 104$)

1, 1, 1; 1, 1; 1, 1; 1, 1; 1, 1; 1, 1, 1; 1, 1; 1, 1, 1; 1, 1; 1, 1, 1; 1, 1; 1, 1, 1; 1, 1, 1; 1, 1; 1, 1, 1; 1, 1, 1; 1, 1, 1; 1, 1; 1, 1; 1, 1, 1; 1, 1, 1; 1, 1; 1, 1, 1; 1, 1; 1, 1; 1, 1; 1, 1, 1; 1, 1, 1; 1, 1, 1; 1, 1, 1, 1; 1, 1, 1; 1, 1, 1; 1, 1, 1; 1, 1, 1; 1, 1, 0; 1, 1, 1; 1, 1, 1; 1, 1, 1; 1, 1, 0; 1, 1, 1; 1, 1, 1; 1, 1, 1; 0, 1, 1; 1, 1, 1; 1, 1, 0; 1, 0, 1; 1, 1, 0; 1, 1, 0; 0, 1, 1; 0, 1, 1; 0, 1, 1; 0, 1, 1; 0, 0, 1, 1; 1, 0, 1, 1; 0, 1, 0, 0; 1, 1, 0; 1, 0, 1; 1, 1, −1, 1; 0, 1, 0, 0; 0, 0, 1; 0, 0, 1, 0; 0, 1, 0; 0, 1, 0; 1, 1, 0, 0; 1, 1, 0, 0; 0, 0, 0; −1, 1, 1; 0, 1, 0, 0; 0, 0, 1, 1; −1, 1, 1; 0, 1, 0; 0, 0, 0; 0, 0, 1; 0, 0, 1; 0, 0, 1; 0, 0, 0; 1, 0, 0; 0, 0; 0, 0, 0, 0; 0, 0, 0; 0, 1, 0; 1, 0, 0; −1, 0, 1, 0; −1, 1, 0; 0, 0, 0; 0, 0, −1; 0, 0, −1; −1, 0, −1, 1; −1, 0, −1, −1, 0; −1, 0, −1, 0; 0, −1, 0; −1, 0, 0; −1, 0, −1; −1, −1; −1, 0, −1; −1, −1, 0; 0, −1, −1; −1, −1, −1; −1, −1, −1, −1; −1, −1, −1; −1, −1, −1; −1, −1, −1; −1, −1, −1; −1, −1, −1; −1, −1, −1

comp2 $(n = 166)$

1, 1, 1, 1; 1, 1, 1; 1, 1, 1; 1, 1, 1; 0, 1, 1, 1; 1, 1, 1; 1, 1, 1; 1, 1, 1; 1, 0, 1; 1, 1, 0; 1, 1, 1; 1, 1,
0; 1, 1, 1; 1, 1, 1; 1, 1, 1; 0, 1, 1; 0, 1, 0, 1; 1, 1, 1; 1, 1, 1; 0, 0, 1, 1; 1, 1, 0; 1, 1, 0; 1, 1, 1; 1,
1, 0; 1, 1; 1, 1, 1; 1, 1, 1; 1, 1, 0; 1, 1, 1; 1, 1, 0; 1, 0, 1; 0, 0, 1, 1; 1, 1, 0, 1, 1; 1, 1, 1, 0; 1, 1,
0, 1; 1, 1, 1, 1, 0, 0; 0, 1, 1; 1, 0, 1; 0, 1, 1; 1, 0, 1; 0, 1, 1; 1, 0, 0; 1, 1, 0; 1, 1, 0; 0, 1, 1; 0, 1,
1; 1, 1, 0, 0; 1, 1, 0, 1; 1, 0; 0, 0, 1; 1, 0, 1, 0; 0, 1, 0; 1, 1, 0; 1, 0, 0, 1; 0, 1, 0; 1, 1, 1, 0, 0; 1,
0, 0; 1, 0, 0; 0, 0, 0, 1; 0, 0, 0; 0, 1, 0; 0, 1, 0, 0; 0, 0, 1; 1, 0, 0; 1, −1, 1, 0; 1, −1, 1; 0, 0, 0;
1, 0, 0; 0, 0, 1; 1, 0, 0; 0, 1, 0; 0, 0; 1, 0, 1, −1, 0; 0, 0, 0; −1, 1, 0; 0, 0, 0; 0, 0, 1; 0, 0, 0; 0,
1, 0, −1; −1, 0, 0; 0, 0; 0, 0, 1; −1, 0, 1; 0, −1, 1, 0; −1, 0, 1; 0, 0, −1, 1; 0, 0; 0, 0; 0, −1, 0;
−1, 1, −1; 0, −1, 0; 0, −1, 0; 0, 0, 0; 1, −1, −1; 1, −1, 0; 0, −1, 0, −1; −1, 0; 0, 0, 0; −1, 0,
0, −1, −1; 1, −1, −1; −1, −1, 0; −1, −1, 0; 1, −1, −1, −1; −1, 0, 0; 0, 0, −1, −1; −1, −1,
−1; −1, −1, 0, −1; −1, 0, −1; 0, −1, −1; −1, 0, −1; −1, 0, −1; −1, 0, −1; −1, −1, 1; −1, 0;
−1, 0, −1; −1, −1, −1; −1, −1, −1; 0, −1, −1; −1, −1, −1; 0, −1, −1; −1, −1, −1; −1, −1,
−1; −1, 0, −1; 0, −1, −1; −1, −1; −1, −1, −1; −1, −1, −1; −1, −1, −1; −1, −1, −1; −1,
−1, −1; −1, −1, −1; −1, −1, −1; −1, −1, −1; −1, −1, −1; −1, −1, −1; −1, −1, −1; −1, −1,
−1; −1, −1, −1; −1, −1, −1; −1, −1, −1; −1, −1, −1; −1, −1, −1; −1, −1, −1; −1, −1, −1;
−1, −1, −1; −1, −1, −1; −1, −1, −1; −1, −1, −1; −1, −1, −1; −1, −1, −1; −1, −1, −1; −1,
−1, −1; −1, −1, −1; −1, −1, −1; −1, −1, −1; −1, −1, −1; −1, −1, −1; −1, −1, −1; −1, −1,
−1; −1, −1, −1; −1, −1; −1, −1, −1; −1, −1, −1; −1, −1, −1; −1, −1, −1; −1, −1

comp3 $(n = 107)$

1, 1, 1, 1, 1; 1, 1, 1; 1, 1, 1; 1, 1, 1; 1, 1, 1; 1, 1, 1; 0, 1, 1, 1; 1, 1, 0; 0, 1, 1; 1, 1, 1; 0, 1, 1, 1;
1, 1, 1; 0, 1, 1, 1, 1; 1, 1, 1, 1; 1, 1, 1, 0, 1; 0, 1, 1, 0, 1, 1; 1, 1, 1; 1, 0, 1, 1; 1, 1, 1, 0; 0, 1, 1,
0; 1, 0, 1, 1; 1, 1, 0; 1, 0, 1; 0, 1, 1; 0, 1, 0, 1, 1, 0; 1, 0, 0; 1, 0, 1; 1, 1, 1, 0, 0; 1, 0, 0, 1; 0, 0,
1, 1, 1; 0, 0, 1, 1, 1; 1, 1, 0, 1, 0; 1, 1, 0, 0, 0; 0, 1, 1, 0, 1; 1, 0, 0; 1, 1, 0; 1, 1, 0; 1, 0, 0; 1, 0,
1; 1, 0, 1; 1, 0, 0; 1, 0, 0; 1, 0, 0, 0; 0, 0, 1; 0, 0, 1, 1; 0, 0, 1, 1, 0; 1, 1, −1, 1, 0; 0, 0, 1, 1, 1;
0, 0, 1; 0, 1, 1, 0; 0, 1, 0; 0, 1, 1; 0, 0, 0, 1, 0; 0, 0, 1; 0, 1, 1, 0; 0, 1, 0; 0, 0, 1; 1, 0, 0, 0, 1; 1,
0, 0, 0; 0, 0, 0, 0, 1; 0, 1, 1, 0, 0; 0, 1, 0; 0, 0, 0, 0, 1, 0; 0, 1, 0; 0, 0, 0, 0, 0, 1; −1, 0, 1, 0, 1,
1; 0, 0, 0; 1, 0, 0, 0; 1, 0, 0, 0, 0; 0, 0, 0; 0, 0, 0; 0, 0, −1, 1; 1, −1, 0; 1, −1, 0; 0, 0, 1, 0; 1,
−1, 0; 0, 0, 0; −1, 0, 0, 1; 0, 0, −1, 0; 0, −1, 0; 0, 0, 0, −1, 0; 0, 0, 0; 0, 0, 0, 0; −1, 1, 0, −1,
0; 0, −1, 0, 0, 0; 0, 0, −1, 0, 1, −1; −1, 0, 1; 1, 0, −1; 0, 0, 0, 0; 0, −1, 1; −1, 0, 0, −1, 0, 1;
−1, 0, 0, 0; 0, 1, −1, 0; −1, 1, 0, 0; 0, 0, 0, −1, −1, 0; 0, 0, −1; 0, −1, 0; 0, 0, 0; 0, 0, 0, −1,
−1, 0; 0, −1, 0, 0, 0, 0; 0, −1, 0; 0, 0, 0, 0, −1; −1, 1, −1, 0; 0, −1, 0, −1, 0, 0; −1, 0, 0, −1;
0, −1, −1; 0, 0, 0, −1, −1, 0; 0, −1, 0, 0, −1, −1; 0, −1, −1; −1, 0, −1; 0, −1, −1, 0; −1, 0,
−1; −1, −1, 0; −1, −1, 0, −1; −1, −1, −1

comp4 $(n = 99)$

1, 1, 1; 1, 1, 1, 1; 1, 1, 1; 1, 1, 1, 1; 1, 1, 1; 1, 1, 1; 1, 1, 1; 1, 1, 1; 1, 1, 1; 1, 0, 1; 1, 1, 1; 1, 1,
1, 1; 1, 1, 1, 1; 1, 1, 1, 1, 0; 1, 1, 0; 0, 1, 1; 1, 1, 1; 1, 1, 1, 1; 0, 1, 1, 1; 1, 1, 1; 1, 1, 1; 1, 1, 1;
1, 0, 1, 1; 1, 0, 1; 1, 1, 1; 1, 1, 1, 1, 1; 0, 1, 1; 1, 1, 1; 1, 1, 1; 0, 1, 1, 0; 1, 0, 1, 1; 0, 1, 1; 0, 1,
1; 1, 1, 0; 0, 1, 1; 1, 0, 0; 0, 1, 1; 1, 1, 0; 1, 1, 0; 1, 1, 0; 1, 1, 0; 1, 1, 0, 0; 1, 0, 0, 1; 0, 1, 1, 0;
0, 0, 1; 1, 0, 0; 1, 0, 0; 0, 1, 0; 1, 0, 0; 1, 0, 0; 0, 0, 1; 1, 0, 0; 0, 1, 0; 0, 1, 0; 1, 0, 0; 0, 0, 1; 0,
0, 1; 0, 1, 0; 0, 0, 1, 0; 0, 1, 0, 1; 0, 0, 0; 0, 1, 0; 0, 0, 1; 0, 0, 0, 1; 0, 0, 0, 1, 0; 0, 0, 1; 0, 0, 1;
1, 0, 0; 0, 0, 1; 0, 0, 0, 1; 0, 0, 0; 0, 0, 0; 0, 0, 0, 0; 0, 0; 0, 0, 0; 1, −1, 1; 0, 0, 0; 0, 0, 0; 0, 0,
0; 0, 0, 0, 0; 0, 0, 0, 0; 0, 0; 0, 0; 0, 0, 0; 0, 0, 0; 0, 0, 0; 0, 0, 0, 0; 0, 0, 0, 0, 0; 0, 0, 0; 0, 0,
−1; 0, 0, −1; 0, −1, 0; 0, 0, −1; 0, −1, 0; 0, 0, −1; 0, 0, −1; −1, 0, −1; 0, −1, −1; −1, 0, −1;
−1, 0, −1; −1, −1, 0; −1, 0, −1; −1, 0, −1; −1, 0, −1; −1, −1, −1; −1, −1, −1; −1, −1, −1

comp5 $(n = 115)$

1, 1, 1; 1, 1, 1, 1; 1, 1, 1, 1; 1, 1, 1, 1; 1, 1, 1, 1; 1, 1, 1, 1; 1, 1, 1, 1; 1, 1, 1, 1; 1, 1, 1, 1, 1; 1,
1, 1, 1, 0, 1; 1, 1, 1, 1; 1, 1, 1, 1; 1, 1, 1, 1; 1, 1, 0, 1; 1, 1, 0, 1; 1, 1, 1, 1; 1, 1, 1, 1; 1, 1, 0; 1,
1, 1, 1; 1, 1, 1, 1; 1, 1, 1; 1, 1, 1, 1; 1, 1, 1, 1; 1, 1, 1, 1; 1, 1, 1, 1; 1, 1, 0; 1, 1, 1, 1; 1, 1, 1, 0,
1; 1, 1, 1, 1; 1, 0, 1, 1; 1, 1, 0; 1, 0, 1; 0, 1, 1, 0, 1; 1, 0, 1; 1, 1, 1, 0; 1, 1, 0, 0; 1, 0, 0; 0, 1, 1;
1, 0, 1; 0, 1, 1, 0; 1, 0, 0, 1; 0, 1, 0, 1; 0, 1, 0; 0, 1, 1, 1; 0, 0, 1, 1; 0, 1, 0; 1, 0, 0, 0; 1, 1, 0, 0;
0, 0, 1; 0, 0, 0, 1; 0, 0, 0, 1; 1, 0, 0, 0; 1, 0, 0, 0; 0, −1, 1, 1, 1; 0, 1, 0, 0; 0, 1, 0, 0; 0, 0, 0, 1;
0, 0, 0; 1, 1, 0, 0; 0, 0, 0, 0, 0; 0, 0, 1; 0, 0, 0; 0, 0, 0, 0; 0, 0, 0; 0, 0, 0; 0, 1, 0, 0; 0, 1, −1, 1;
1, 0, 0, −1; 0, 0, 0, −1; 0, 0, −1, 0; 0, 0, −1, 0; 0, 0, 0, −1; 0, −1, 0, 0, 0; 0, −1, −1, 0; 0, 0,
0, 0; 0, −1, 0, 0; 0, −1, 0, 0; −1, 0, 0, −1, 0, 0; 0, 1, −1, 0; −1, 0, 0, 0; −1, −1, 1; −1, 0, 0;
−1, 0, 0, 0; 0, −1, 1; 0, −1, 0, −1, −1; −1, −1, 0, 0; −1, −1, 0, 0; 0, −1, −1; 0, −1, −1, −1,
−1; 0, −1, −1; −1, −1, −1; −1, −1, −1; −1, −1, −1; −1, −1, −1, −1; −1, −1, −1, −1; −1,
−1, −1, −1; −1, −1, −1, −1; −1, −1, −1, −1; −1, −1, −1, −1

comp6 ($n = 87$)

1, 1, 1; 1, 1, 1; 1, 1, 1, 1; 1, 1, 1; 1, 1, 1; 1, 1, 1; 1, 0, 1; 1, 1, 1; 0, 1, 1; 1, 1, 1; 1, 1, 1; 1, 1, 1; 0, 1, 1; 1, 1, 1, 1, 0, 1, 1; 1, 0, 1; 1, 1, 1; −1, 1; 1, 1, −1, 1; 1, 1, 1, 0; 0, 1, 1; 0, 1, 1; 1, 1, 1; 1, 1, 1; 1, 1, 1, 1; 1, 0, 0; 1, 0, 1; 0, 0, 1, 1, 1; 1, 0, 1; 1, 0, 0; 0, 1, 0, 0; 0, 1, −1, 1; 0, 1, 0; 0, 1, 1, 0; 0, 0, 1, 1; 0, 0, 1; 1, 0, 0; 1, 0, 1, 1; 0, 1, 0; −1, 0, 0, 1; −1, 1, 1; 1, 0, 0; 0, 1, 0, 1; 0, 0, 1; 0, 0, 0; 0, 1, 0; 1, 0, 0; 0, 0, 1; 1, 0, 0; 0, 0, 1; 0, 0, 0, 0; −1, −1, 1, 1; 1, −1, 1, 0; 0, 0, 0; 0, 1, 0; −1, 0, 0; 0, 1, 0; 0, 0, 0; 0, 0, 1; 0, 0, 0; 0, 1, −1; 0, 0, −1; 1, −1, 0; 0, −1, 1; −1, 0, 0; 0, 1, −1; 0, 0, 0; 1, −1, 0; 0, 0, 0; 0, 0, 0; −1, −1, 1; −1, 0, 0; −1, 0, −1; 1, −1, −1; −1, −1, 0; 0, −1, 0; −1, 0, −1; 0, 0, −1; 0, −1, −1; −1, −1, −1; −1, −1, −1; −1, −1, −1; −1, 0, −1; −1, −1, 0; 0, 0, −1; −1, 0, 0; −1, −1; 0, −1, −1

comp7 ($n = 28$)

1, 1; 1, 1; 1, 1; 1, 1; 1, 0; 1, 1; 0, 1; 0, 0; 0, 1; 0, 1; 1, 0; 1, 0; 0, 0; 1, −1; 0, 0; 0, 0; −1, 0; −1, 1; −1, 0; 0, 0; −1, 0; 0, −1; −1, −1; −1, 0; −1, −1; −1, −1; −1, −1; −1, −1

neuro1 ($n = 179$)

1, 1; 1, 1; 1, 1; 1, 0; 1, 0; 1, 0; 1, 0; 1, 0; 1, 0; 1, 0; 1, 0; 1, 0; 1, 0; 1, 0; 1, 0; 1, 0; 1, 0; 1, 0; 1, −1; 1, −1; 1, −1; 1, −1; 1, −1; 1, −1; 0, 1; 0, 1; 0, 1; 0, 0; 0, −1; −1, 1; −1, 1; −1, 1; −1, 1; −1, 1; −1, 0; −1, −1

neuro2 ($n = 116$)

1, 1; 1, 0; 1, 0; 1, 0; 1, 0; 1, 0; 1, 0; 1, 0; 1, 0; 1, −1; 1, −1; 1, −1; 1, −1; 0, 1; 0, 1; 0, 1; 0, 0; 0, −1; 0, −1; 0, −1; 0, −1; 0, −1; 0, −1; 0, −1; 0, −1; 0, −1; 0, −1; 0, −1; 0, −1; −1, 1; −1, 0; −1, 0; −1, 0; −1, 0; −1, 0; −1, 0; −1, 0; −1, 0; −1, 0; −1, 0; −1, 0; −1, 0; −1, 0; −1, 0; −1, 0; −1, 0; −1, 0; −1, −1

infor1 ($n = 58$)

1, 1; 1, 0; 1, 0; 1, 0; 1, 0; 1, 0; 1, 0; 1, 0; 1, 0; 1, 0; 1, 0; 1, 0; 1, −1; 1, −1; 1, −1; 1, −1; 0, 0; 0, −1; 0, −1; 0, −1; 0, −1; 0, −1; 0, −1; 0, −1; 0, −1; 0, −1; 0, −1; 0, −1; 0, −1; −1, −1

infor2 ($n = 68$)

1, 1; 1, 1; 1, 1; 1, 1; 1, 0; 1, 0; 1, 0; 1, 0; 1, 0; 1, 0; 1, 0; 1, 0; 1, 0; 1, 0; 1, 0; 1, 0; 1, 0; 1, 0; 1, 0; 1, 0; 1, 0; 1, 0; 0, −1; 0, −1; 0, −1; 0, −1; 0, −1; 0, −1; 0, −1; 0, −1; 0, −1; 0, −1; 0, −1; 0, −1; 0, −1; 0, −1; −1, −1; −1, −1; −1, −1; −1, −1