# Jury Theorems for Peer Review*

Marcus Arvan[†] Liam Kofi Bright[‡] Remco Heesen[§]

August 21, 2020

**Abstract**

Peer review is often taken to be the main form of quality control on academic research. Usually journals carry this out. However, parts of math and physics appear to have a parallel, crowd-sourced model of peer review, where papers are posted on the arXiv to be publicly discussed. In this paper we argue that crowd-sourced peer review is likely to do better than journal-solicited peer review at sorting papers by quality. Our argument rests on two key claims. First, crowd-sourced peer review will lead on average to more reviewers per paper than journal-solicited peer review. Second, due to the wisdom of the crowds, more reviewers will tend to make better judgments than fewer. We make the second claim precise by looking at the Condorcet Jury Theorem as well as two related jury theorems developed specifically to apply to peer review.

*All authors contributed equally. We thank Justin Bruner, Allard Tamminga, Boudewijn de Bruin, Kevin Zollman, Michael Morreau, and an audience at the University of Groningen for valuable comments and discussion. RH's research was supported by the Netherlands Organisation for Scientific Research (NWO) under grant 016.Veni.195.141.

[†]University of Tampa. Email: `marvan@ut.edu`.

[‡]London School of Economics and Political Science. Email: `liamkbright@gmail.com`.

[§]University of Western Australia and University of Groningen. Email: `remco.heesen@uwa.edu.au`.

# 1 Introduction

Peer review is supposed to secure an epistemic benefit. By ensuring that only work that has been validated by multiple experts is allowed into the academic literature, peer review is commonly thought to function as a quality control that prevents us from wasting time on poor work. Rather than have to wade through every half-baked flight of fancy, a discerning researcher may simply peruse peer-reviewed journals and read only that which passes peer review. However, in this essay we argue that an open, 'crowd-sourced' approach to peer review is more likely to reliably identify high-quality work compared to traditional, journal-solicited peer review.

The widely practices journal-solicited approach to peer-review filters the quality of academic work through a small number of experts—typically, a few editors and one to three outside referees. The normative assumption that appears to underlie this practice is the belief that a small number of experts (reading anonymized submissions) are *the best mechanism* for distinguishing between high-quality and low-quality work. In this system, quality assessment occurs in two stages: first, in pre-publication peer review, which sorts papers into journals; and second, by a journal's readership post-publication. We take it that the first stage is intended to provide a proxy for the second, i.e., the more long-term assessment by the field. In this paper, we argue that a 'crowd-sourced' approach to peer review that bypasses the first stage—immediately opening up papers for evaluations by large numbers of readers—is likely to *more reliably* evaluate paper quality than the traditional model of peer review. In brief, we use the Condorcet jury theorem (Condorcet 1785) and some closely related mathematical results to argue that a large number of evaluators is more likely to produce an accurate quality assessment of a paper than a smaller number of evaluators.

While we offer up some specifics of an open, crowd-sourced peer-review system, this paper is not intended to provide a full outline of such a system. In §3–6 we offer enough of the details to make our comparative case. In §7, while considering some objections, we gesture to some further features one may wish to incorporate into such a system. For instance, one possibility we note is that in addition to crowd-sourcing from the academic community at large, it might be desirable to have a core of expert reviewers whose assessments are recorded separately. So, while we do not rest our case on such specifics of a crowd-sourced model, we are often (as in this case) supportive of particular proposals for how such a system might work. In our view, an experimental attitude to crowd-sourcing peer review will be a much better way to work out the details than any argument we could provide here.

We close this introduction by relating our argument to the previous literature. There is renewed interest in the epistemic benefits secured by large numbers of diverse agents (List and Goodin 2001, Hartmann and Sprenger 2012), including in the social epistemology of science (Heesen et al. 2019, O'Connor and Bruner 2019, Singer 2019). Our intent is to bring this literature to bear on a concrete problem in the social epistemology of science, namely peer review. Further, given the replication crisis, there has recently been interest in systematic failures of peer review (Romero 2016, Heesen 2018). Our paper offers a thoroughgoing solution to these problems. Like those who argue we should eliminate peer review in the context of project funding (Avin 2019), we think we should abandon the idea that a small number of experts can reliably predict which ideas will be worth reading. We are not the first to suggest opening up peer review (see, e.g., Gibson 2007, Nosek and Bar-Anan 2012, Heesen and Bright forthcoming). However, we offer a novel argument for its epistemic benefits. If our argument is sound, *nothing* should get a potentially deceptive stamp of authority through journal-solicited peer review. Instead, to gauge the quality of academic work, we should rely on the long-run and aggregated views of many diverse researchers.

## 2 Assumptions of Peer Review

Our purpose is to compare the present system of journal-solicited pre-publication peer review against a crowd-sourced model. However, we expressly do not engage in an all-things-considered comparison. Rather, we focus on one goal of peer review (central to its defenders): namely the selection of high-quality papers. Thus we set aside other goals, such as improving the quality of papers.

In this section we argue that *if* the present system of peer review really helps us pick out high-quality papers (however imperfectly), then research quality and the peer reviewers who assess it must satisfy certain assumptions. Our argument's structure is loosely analogous to a transcendental deduction: we argue that without satisfying these assumptions, the idea that any form of peer review could successfully select for quality makes no sense. The two assumptions are *competency* and *intersubjectivity*.

Let us begin with the competency assumption. We assume that researchers are at least better than chance at picking out papers of high quality or ranking papers according quality. The significance of this assumption is just that quality is the sort of thing which a peer reviewer is capable of discerning. If this were false, then the current system could do no better than a system of random publication. So quality is the sort of thing which researchers can and do discern, and respond appropriately to given their reviewing task.

The second assumption is that there exists broad (if rough) intersubjective agreement about what constitutes quality. The idea is that for any given academic paper, there is a unique notion of quality: one way of being the best version of that paper which readers and reviewers should track. Note that we are not saying that there is only one type of quality for all papers. Rather, we are assuming that once you fix facts about a paper's topic and the type of impact it is intended to have, *then* there is a uniquely best way of fulfilling the paper's purpose.

Even when relativized in this way, a unidimensional notion of quality invites skepticism. Kuhn (1977, chapter 13) emphasized that there are many respects in which a scientific theory (and by extension a scientific paper) might be judged good or bad, and different reviewers will weigh these differently (see also Okasha 2011, Heesen 2019, for a more formal approach to this issue).

However, consider what would be the case if our assumption were false. If there is no intersubjectively agreed unidimensional notion of quality, then it is unclear what peer review is doing. Why should the fact that reviewers like a paper give me any reason to think that I will like it? The fact that we assume that peer reviewers can assess quality and make useful judgments about what is worth spending time on belies a presupposition: namely, that for any given paper the relevant experts know what it would take to be a more or less worthy version of that paper, and that we can reasonably expect some agreement on this point. We allow, and in fact it will be essential to our argument, that this agreement may be partial and accompanied by substantive and persistent disagreement on particular points.

Moreover, the assumption of an intersubjectively agreed notion of quality is deeply entangled in many academic practices. Decisions about hiring, tenure, and promotion are at least partially based on the journals researchers have published in. This practice assumes that better journals are more likely to publish better papers, which in turn assumes that something worth caring about can be learned from peer reviewers' assessment. Thus anyone whose opinion on anything has been influenced by where something was published is implicitly committed to intersubjectivity and competence. Hence, if peer review is to make sense as a system of quality control, then the following features must be in place: Paper quality is such that for any given paper there is a unique best way it could fulfill its own potential, and the research community contains people who are competent to assess this.

While our focus is exclusively on the role of peer review in quality control, we feel it is useful at this stage to briefly discuss *fairness* considerations. Proponents of journal-solicited peer review often suggest that our current practices are the fairest method available, as a particular feature of the existing peer review process—anonymization—is vital for protecting against reviewer bias. We take no stance on anonymization. Our arguments instead support the following conditional claims: if anonymization is important for fairness in peer review, then an anonymized crowd-sourced peer review process would be superior to current processes; and if anonymization is unimportant, then a non-anonymized crowd-sourced peer review process would be superior. Either way, our point stands: an open, crowd-sourced method of peer review is likely to more accurately judge paper quality than journal-solicited peer review.

# 3 Crowd-sourcing More Reviewers

One notable feature of journal-solicited peer review is that paper quality is judged by a small number of evaluators. First, papers are often read and 'desk-rejected' by a single editor. Second, when papers are sent out for review, they are typically reviewed by anywhere from one to three referees. Contrast this system to the crowd-sourced peer review system already utilized in math and physics. In these disciplines, it is standard for unpublished papers to be posted on individuals' professional websites and on central repositories, such as the arXiv. It also appears to be a disciplinary norm for members of the academic profession to read and publicly evaluate new submissions.

Evidence for this norm in math and physics is anecdotal, as there are no formal studies of how widely arXiv preprints are discussed prior to journal publication. However, the evidence is suggestive. Prominent weblogs in both fields hosted by experts (PhDs and tenured faculty) routinely discuss and evaluate new arXiv preprints in detail, including in comment sections, where additional experts often weigh in.[1] Widespread discussions of arXiv preprints, including blog posts and subsequent arXiv preprints criticizing earlier preprints, lead to broader judgments in the discipline about the quality of particular papers before publication in any journal.

It is not merely well-known figures whose preprints are discussed. For example, in 2007, a PhD researcher in physics who had left academia posted an arXiv preprint entitled "An Exceptionally Simple Theory of Everything" (Lisi 2007). Despite having no academic post, Lisi's paper was discussed at major physics blogs[2] and in further arXiv preprints (e.g., Distler and Garibaldi 2009). This process of online discussion led to a disciplinary consensus that Lisi's paper is flawed.[3] In 2012 a well-known mathematician posted drafts of four preprints on his website, totaling around 500 pages, claiming to prove the *abc* conjecture (Mochizuki 2012a,b,c,d). A flurry of discussion on blogs and in followup preprints followed, identifying and debating potential flaws with the proof.[4]

As we see in these and other cases, the evaluation of paper quality is more widely distributed in math and physics. If it is generally expected in an academic profession—as in math and physics—that unpublished preprints should be read and discussed publicly before publication, then chances are high that under crowd-sourced peer review, the average paper will be reviewed by a larger number of reviewers than in a journal-solicited system.

It is worth noting that this does not necessarily require the overall time

[1] Examples in physics include *A Quantum Diaries Survivor* (`https://www.science20.com/a_quantum_diaries_survivor`), *Of Particular Significance* (`https://profmattstrassler.com`), and *Backreaction* (`http://backreaction.blogspot.com`). Examples in math include Terence Tao's blog (`https://terrytao.wordpress.com`) and *Persiflage* (`https://www.galoisrepresentations.com`).

[2] E.g., `http://backreaction.blogspot.com/2007/11/theoretically-simple-exception-of.html`.

[3] See `https://www.scientificamerican.com/article/wipeout-theory`.

[4] See Scholze and Stix (2018) and `https://www.galoisrepresentations.com/2017/12/17/the-abc-conjecture-has-still-not-been-proved` for discussion of the mathematical content and `https://twitter.com/andrewaberdein/status/1246553878553939980` for a timeline.

spent reviewing papers to increase. Suppose that the current disciplinary norm to volunteer one's time to review papers for journals is shifted over to the new system of crowd-sourced peer review, such that each member of the academic community volunteers exactly the same amount of time and reviews the same number of papers. Under journal-solicited peer review, journals base their decision about the quality of a paper only on the reviews solicited by that journal. For those papers that have already been rejected from other journals, all previous reviews are normally ignored. In contrast, under crowd-sourced peer review all reviews are public. Thus, if the total number of reviews remains constant, the average number of actually available reviews per paper under crowd-sourced peer review will be higher than under journal-solicited peer review.

We will assume throughout the rest of this paper that moving to crowd-sourced peer review increases the average number of reviews per paper. For a reader who thinks that crowd-sourced peer review will lower the number of reviews per paper, the arguments presented below will favor journal-solicited peer review over crowd-sourced peer review. However, they could also be read as a normative argument in favor of increasing the average number of evaluators of any given paper.

# 4 The Basic Condorcet Jury Theorem

In this section and the next two, we provide formal arguments that having more reviewers is likely to lead to superior decisions. Our argument in this section is an application of the Condorcet Jury Theorem.

The Condorcet Jury Theorem shows that, subject to three assumptions, the judgments of a jury—a group charged with voting on the truth of a proposition (where a majority vote wins)—have a greater probability of accuracy the greater the number of people in the jury. The first assumption is that there is a *correct answer*: the proposition the jury is judging is either true or false. The second assumption is that every member of the jury has some probability of voting for the correct truth-value of the proposition that is *probabilistically independent* of the other jury members' vote. Finally, the third assumption is that the *average*[5] probability that any individual in the jury votes correctly is greater than .5. The theorem then says that adding more voters to the pool (keeping the average probability of voting correctly constant) *makes it progressively less likely* that the majority vote for the wrong conclusion.

Here is a brief, intuitive illustration of the theorem. Suppose the average probability that a juror votes for the right answer is .51. If only 100 people serve on the jury, then the most likely result is that 51 jury members will vote for the correct answer and 49 for the wrong answer. If, however, just *one* additional jury member votes wrongly, then the result will be a tie. And if *two* additional jury members vote wrongly, then the jury will vote 51-49 for the *wrong* verdict. So it is not unlikely for this jury to go wrong (this happens with a probability of approximately .38, with an additional .08 probability of a tie). Now consider a jury of $100,000$. If the average probability of a correct vote remains .51, the most likely result is that $51,000$ jury members will vote for the right verdict and $49,000$ for the wrong verdict. Consequently, a *thousand* additional jury members would have to make a mistake to shift the jury from the single most likely outcome to the wrong result. But this occurs only with a probability of around one in ten billion.

The theorem also shows that in the limit (an infinite-sized jury), the majority will vote for the correct answer with probability one (i.e., 100 % of the time). The relevant point for our purposes, however, is the *comparative* claim: the more jury members there are, the more likely it is that a majority of them will vote for the correct answer. This is important in light of our assumption that the typical article in a crowd-sourced peer review model will be read and evaluated by more people than under journal-solicited peer

[5]This generalizes the original theorem by allowing individual probabilities to vary.

review.

To see how the Condorcet Jury Theorem plausibly supports crowd-sourced peer review, compare the assumptions of the theorem to those discussed in §2. The theorem's first assumption is that the proposition being judged is true or false. In the case of peer review this proposition would be something like 'This paper is of high quality'. This aligns closely with the second assumption we argued peer review must satisfy, i.e., that there is an *intersubjective* quality standard for a paper on a particular topic. For the moment we are assuming that peer reviewers give (only) a binary judgment of quality: thumbs up or thumbs down. One of the motivations of the models in §5–6 is to consider more informative, graded reviewer judgments.

Now consider the Condorcet theorem's second assumption: that every jury member has an *independent probability* of voting for the correct result. This assumption does not correspond directly to one of the assumptions from §2. Under crowd-sourced peer review, we can imagine reviewers' judgments becoming correlated due to reviewers being able to read other reviews, whereas under journal-solicited peer review, the active hand of an editor may likewise induce correlation of reviewers' judgments. So whether the independence assumption is satisfied may well depend on the mechanism by which the different peer review systems are implemented. We will say more about steps a crowd-sourced peer review model could take to ensure reviewer independence in §7.1 and §7.3. For now we emphasize that independence is assumed in the Condorcet Jury Theorem, and hence the real-world applicability of our argument in this section hinges on providing a mechanism to guarantee it.

Finally, consider the Condorcet theorem's third assumption: that on average voters' probability of voting for the correct answer is better than chance. This corresponds to the first assumption we argued that peer review must satisfy: that reviewers are *competent* at picking out high-quality papers.

Our claim is that, for peer review to reliably select papers for publication on the basis of quality, two of the three assumptions of the Condorcet Jury Theorem must be satisfied. Moreover, the third assumption (independence) will be satisfied by crowd-sourced peer review if the latter is carefully implemented (we defer our discussion of this to §7). But then a crowd-sourced peer review model is more reliable than a journal-solicited peer review model. For we have argued in §3 that crowd-sourced peer review will tend to base evaluations on the judgment of a larger jury than journal-solicited peer review. And by the Condorcet Jury Theorem a larger jury is more likely to arrive at an accurate evaluation.

# 5 A Jury Theorem for Reviewer Scores

In the previous section we argued that the crowd-sourced method of peer review is superior to the present system on the basis of an 'off the shelf' application of the Condorcet Jury Theorem. While we find this argument convincing, we recognize that the basic Condorcet model is highly idealized and thereby open to objections. In this section we provide a new model intended to be more tailored to the specifics of peer review, and show that an analogous theorem holds in this model. This shows that the jury theorem is robust against certain changes in its assumptions, thus strengthening our argument.

The basic Condorcet model assumes that agents make a binary judgment on a single proposition. In contrast, real peer reviewers provide more nuanced judgments. These may come in the form of numerical scores or qualitative reasons for the reviewer's verdict. This section considers a model of peer review where reviewers only provide a numerical score; we add qualitative reasons in the next section.

Whereas in the previous section the goal was to evaluate the truth value of the proposition 'This paper is of high quality', now the goal is to *rank* papers, with those ranked highest most recommended to the attention of other researchers. By the intersubjectivity assumption, for any two papers, one can accurately be said to be of higher quality than another.

Each review consists of a numerical score, which is the reviewer's estimate of the paper's quality. We write $q_i$ for the quality estimate provided by reviewer $i$. By the competence assumption, reviewers tend to give higher scores to better papers. But as in the previous section, we assume that there is some random variation in reviewer scores, reflecting individual reviewer biases and idiosyncrasies. Also as in the previous section, we assume that this variation is independent across reviewers, so reviewer scores can be modeled as independent random draws from a large pool of potential reviewers or reviewer scores (we refer again to §7.3 for more discussion of the independence assumption).

In this setup, we can represent the competence assumption by assuming that, on average, reviewers agree on the quality of a paper (that is, for each paper there is a number $q$ such that $\mathbb{E}[q_i] = q$ for all $i$). And we can represent reviewer biases and idiosyncrasies by assuming that there is some random variation around this average ($\mathrm{Var}[q_i] = \sigma^2 > 0$ for all $i$). The intersubjectivity assumption is reflected in the different average for each paper (if $q_i$ and $r_i$ are reviewer $i$'s scores for distinct papers, the former is intersubjectively better if $\mathbb{E}[q_i] > \mathbb{E}[r_i]$).

Given differing quality estimates from reviewers that are each taken to be competent, it seems reasonable for a journal editor or arXiv reader to take the average of these estimates to be her best estimate of the relative quality of a paper. Averaging in this way has been defended in the literature on combining forecasts (Clemen 1989, Armstrong 2001, especially p. 422) and peer disagreement (Elga 2007, Christensen 2007, Cohen 2013), while (weighted) linear averaging more generally has also been widely defended by formal epistemologists (Lehrer and Wagner 1981, Martini and Sprenger 2017, Pettigrew 2019). So the quantity of interest that will be used to make decisions under either journal-solicited or crowd-sourced peer review is the average of $n$ reviewer scores $q_1, q_2, \ldots, q_n$, which may be written $\frac{1}{n}\sum_{i=1}^{n} q_i$.[6]

Because individual reviewer scores are equal in expectation, so is the average reviewer score (that is, $\mathbb{E}[\frac{1}{n}\sum_{i=1}^{n} q_i] = \mathbb{E}[q_1]$). Perhaps more importantly, the random variation in the average reviewer score will decrease as the number of reviewers increases, according to the formula $\mathrm{Var}[\frac{1}{n}\sum_{i=1}^{n} q_i] = \sigma^2/n$. This means that the more reviewers there are, the smaller the probability that the average reviewer score will be much different from its expectation. Since better papers have a higher expectation, the probability that two papers are ranked incorrectly will similarly decrease.

This gives us a clear analogy to the Condorcet Jury Theorem. Previously, increasing the number of reviewers increased the probability of a correct verdict, whereas here increasing the number of reviewers increases the probability that papers are ranked correctly. To complete the analogy, note that the random variation will reduce to zero in the limit as the number of reviewers becomes infinite, meaning that papers are ranked correctly with probability one.

Once again, granted the assumption that crowd-sourced peer review will have on average more reviewers per paper than journal-solicited peer review, this yields an argument in favor of crowd-sourced peer review as more likely to yield accurate quality judgments.

[6] A potential objection here is that the average is only a meaningful quantity if reviewer scores are assumed to be measured on a cardinal scale, whereas arguably such judgments only have ordinal significance (see Tal 2020, section 3.2 for background on this classification). This need not always be a problem in practice. For example, if the underlying distribution is symmetric, the mean coincides with the median, which is ordinally meaningful, so the average will 'accidentally' track something meaningful. But to address this worry more fully we might take the median of reviewer scores as the quantity of interest instead. If we change our assumptions appropriately (individual reviewer scores share a median, with a higher median for better papers), essentially the same argument goes through, as the variance of the median of reviewer scores will also decrease with more reviewers. Moreover, if we assume reviewer scores are discrete rather than continuous, the jury theorem for the median proven by Morreau (forthcoming) applies, so our argument goes through in that setting as well. Working with the median has another advantage: it is robust against outliers. We recommend using both where feasible (see §7.1). We thank Michael Morreau for suggesting both the objection and the response.

## 6 A Jury Theorem for Reviewer Reasons

In this section we expand on the previous section's model by including reviewers' reasons for giving a particular (numerical) quality judgment. We represent these reasons by thinking of papers as having a number of features and peer reviewers as having opinions on which combinations of features make for a high-quality paper. More specifically, we assume there are $m$ features that peer reviewers evaluate for a paper on a certain topic (recall that quality standards are paper-specific). A paper is represented by its feature coordinates $x_1, x_2, \ldots, x_m$, which provide a numerical 'score' for how that paper does on each feature. We imagine that for each feature there is a 'golden mean' (possibly relative to the value of the other features) such that both more and less of that feature would make the paper worse in the eyes of the reviewer.

For example, say that for a given paper feature 1 concerns the paper's discussion of the external validity of its results, so $x_1$ indicates how the paper scores on this feature. A low value of $x_1$ might indicate that the discussion unnecessarily constrains the external validity (compared to what is justified per the scores on other features); a high value then says that the discussion generalizes the study's results too widely (i.e., in ways not sufficiently supported by the evidence). A medium value indicates a sensible discussion that avoids unsupported claims.

What might the set of features look like? Since our argument goes through regardless, we can remain agnostic between the following suggestions. First, the features might be Kuhn's criteria for theory choice: empirical adequacy, simplicity, etc. Second, the features might be some variation on those that peer reviewers are explicitly asked to score papers on by journals: novelty, methodological soundness, etc. Third, our preferred option, the features might be anything and everything peer reviewers use to evaluate papers, at as fine-grained a level as possible.

According to our intersubjectivity assumption, any given paper has an (intersubjectively agreed) quality. We conceive of both a paper's quality and any given reviewer's opinion of its quality as a function of that paper's feature coordinates $x_1, \ldots, x_m$. Quality can then be characterized as something that looks like an epistemic landscape (in the sense of Weisberg and Muldoon 2009, Alexander et al. 2015, Thoma 2015): each $m$-dimensional point $x = (x_1, \ldots, x_m)$ represents a possible paper, and the height of the landscape at that point is the quality of that paper. We define the function $f : \mathbb{R}^m \to [0, \infty)$ to describe this epistemic landscape. That is, $f(x)$ is the intersubjective quality of a paper with characteristics $x$.

In accordance with our competence assumption, peer reviewers (whether crowd-sourced or journal-solicited) can estimate the quality of a paper. However, they are not perfect. First, they may be biased in that the combinations of features they perceive to indicate high quality are different from the combinations that really constitute intersubjective quality. Second, there may be measurement error, i.e., peer reviewers may make mistakes in evaluating a paper on some or all features. We roll these two types of errors into a single bias $b_i$ for a given reviewer $i$. The bias $b_i$ is an $m$-dimensional point representing the total distortion in reviewer $i$'s estimation of quality, such that the reviewer's quality estimate for a paper with characteristics $x$ will be $f(x + b_i)$.

We denote by $\mu$ the center of mass of the epistemic landscape of intersubjective quality, and assume that this quantity exists.[7] Consequently, for any reviewer $i$, the epistemic landscape characterizing how that reviewer estimates quality also has a center of mass located at $\mu - b_i$.

As before, we assume that a journal editor or reader takes the average of

[7]More formally, we assume $\int_{\mathbb{R}^m} x_j f(x)\, dx$ is finite for each feature $j$ and define $\mu$ coordinate-wise by $\mu_j = \int_{\mathbb{R}^m} x_j f(x)\, dx / \int_{\mathbb{R}^m} f(x)\, dx$. Given our 'golden mean' approach to paper quality, this assumption is fairly innocent. It holds in particular if $f$ has a finite maximum (as it does under any reasonable formalization of the 'golden mean' approach) and the features are measured on finite scales. If the features are measured on infinite scales, whether the assumption holds depends on how quickly quality drops off away from the maximum.

these estimates to be her best estimate of the quality of a paper. So for a paper $x$ reviewed by $n$ reviewers the editor's or reader's quality estimate is $f_n(x) = \frac{1}{n}\sum_{i=1}^{n} f(x + b_i)$.

Also as before, we assume that crowd-sourced peer review will lead (on average) to more reviewers per paper than journal-solicited peer review. The question then is whether this improves the quality estimate. For a given paper $x$, this translates in the model to the question whether $f_n(x)$ gets closer to $f(x)$ as $n$ increases. Depending on the shape of the landscape and reviewers' biases, this may be true for some values of $x$ and false for others. What we would like to know, then, is whether for an *arbitrary* paper the quality estimate gets closer to the intersubjective quality with more reviewers, i.e., whether the function $f_n$ as a whole becomes more similar to the function $f$ as $n$ increases.

How do we characterize the similarity of two functions? Here we take the following approach: compare the centers of mass. The center of mass measures the central tendency of a function, giving some indication of where in the landscape the highest peaks of quality occur. This is a fairly crude measure of similarity: two functions may have the same center of mass but be dissimilar in other respects. However, it has the advantage of giving us a single number (or $m$-dimensional point, rather) for each function. This measure works well when the landscapes are single-peaked and mostly smooth, as two such landscapes with similar centers of mass will usually agree on relative judgments (i.e., which of two papers is better). The center of mass of $f_n$ is $\mu - \frac{1}{n}\sum_{i=1}^{n} b_i$, which we compare to $\mu$, the center of mass of $f$.

Now we just need to worry about how the reviewer biases are distributed. We assume that we can treat these as random variables. This need only be true in a subjective sense: the bias of a given reviewer is random insofar as you do not know in advance which reviewer and hence which bias will be selected. We assume that expected bias is zero (loosely speaking, this says that bias is equally likely to be in any direction) and expected variation in bias is finite.[8]

It follows that in expectation the center of mass of estimated quality is equal to the center of mass of intersubjective quality ($\mathbb{E}[\mu - \frac{1}{n}\sum_{i=1}^{n} b_i] = \mu$), that is, on average there will be no bias at all. But assuming the biases of different reviewers are independent (see §7.3 for discussion), we also get that the probabilistic variation in the center of mass of estimated quality decreases with the number of reviewers ($\mathrm{Cov}[\mu - \frac{1}{n}\sum_{i=1}^{n} b_i] = \Sigma/n$). This means that the center of mass of estimated quality is more likely to be far away from the center of mass of intersubjective quality if there are fewer reviewers, and more likely to be close if there are more. Moreover, since the variation reduces to zero in the limit, $\mu - \frac{1}{n}\sum_{i=1}^{n} b_i$ probabilistically converges to $\mu$.

These results provide another close parallel to the Condorcet Jury Theorem. Estimated quality is likely to be closer to intersubjective quality as the number of reviewers increases, and they coincide (by our fairly crude measure) in the infinite limit. We conclude that crowd-sourced peer review, insofar as it tends to involve a greater number of reviewers, outperforms journal-solicited peer review even when we grant the basic assumptions (competency and intersubjectivity) that are required for journal-solicited peer review to make sense.

# 7 Replies to Potential Objections

We anticipate several objections, each focusing on a different background assumption.

[8] More formally, we assume (for all $i$) that $\mathbb{E}[b_i] = 0$ and $\mathrm{Cov}[b_i] = \Sigma$. Here, 0 is the $m$-dimensional origin, and $\Sigma$ gives the covariances between a particular reviewer's bias in each of the $m$ features (*not* the covariances between different reviewers' biases). In virtue of being the covariance matrix, $\Sigma$ is a symmetric and positive semi-definite $m \times m$ matrix. We make no assumptions on $\Sigma$ except that it is not the zero matrix.

## 7.1 Manipulation of Reviewer Scores?

Our assumptions of reviewer competence and independence of reviewer judgments would not be plausible for crowd-sourced peer review if it is overwhelmed by internet trolls with a political agenda or other forms of organized manipulation. If people base their judgment of a paper on reasons orthogonal to its quality, then our reviewer competence assumption will not be satisfied. If groups of people are mobilized to leave reviews of a paper without much thought, then our independence assumption fails. Note that the latter will be a problem for our independence assumption regardless of whether such a 'mass reviewing campaign' is ultimately motivated by scientific (e.g., a large research program ganging up on a smaller one), political, or other reasons. This is an important worry for crowd-sourced peer review given the extent to which various social media have recently been overwhelmed by such phenomena.

It is tempting to address this issue by putting tight restrictions on who is allowed to review. There are a number of ways of doing this. We might use formal requirements such as possession of a doctorate or academic employment, or social requirements such as endorsement from existing reviewers or reviews rated sufficiently helpful by other reviewers. We might apply such requirements at a system-wide level (i.e., to decide whether a given person is allowed to review anything at all) or at a subfield-specific level, e.g., requiring a doctorate in a specific subfield or endorsement from subfield specialists to be allowed to review papers in that subfield.

As explained below, there may be ways of testing whether such restrictions are necessary, and if so, what types of restrictions would function best. However, our current opinion is that such restrictions go against the spirit of the proposal of crowd-sourced peer review and will limit its advantages. Restricting who is allowed to review will lower the average number of reviews per paper, thus reducing the benefits from large numbers we have discussed. Moreover, such restrictions may reinforce existing disciplinary boundaries and subfield-level groupthink (where it exists), whereas one of the key envisioned strengths of crowd-sourced peer review is that it will be easier for disparate fields to cross-pollinate, benefit from each other's insights, and correct each other's biases.

For these reasons we favor a system in which anyone is allowed to review anything, regardless of whether they are a recognized expert or even an academic. But this does not mean giving free rein to trolls, mobs, and other manipulation. We think there are various possible measures to guard against these.

Here is a relatively simple one. For each subfield, curate a set of expert reviewers along the lines suggested above, e.g., have reviewers endorse each other's expertise in the given subfield. Here we imagine subfields to be relatively small, say, more than twenty but less than a hundred endorsed reviewers per subfield. Then, for each paper, report both the overall average reviewer score and the average *expert* reviewer score when taking into account only reviewers endorsed for that particular subfield.

This system, familiar from the film review website Rotten Tomatoes—which reports qualified reviewer scores and general audience scores—has a number of advantages. Researchers who prefer something close to journal-solicited peer review can use the expert reviewer average, whereas those favoring the wisdom of the crowds can focus on the overall average. More importantly, one can look at both. When they are similar, either there were little or no non-expert reviewers, or the non-expert reviewers tended to agree with the expert reviewers. It gets interesting when there is significant divergence between the expert reviewer average and the overall average. This could be evidence of a mob coming in to manipulate the score. However, it could also be evidence of groupthink within the subfield, exposed by the independent insights of outsiders. In any case, the divergent scores will be a signal that something is up (Nosek and Bar-Anan 2012, p. 238). Individual readers will be alerted that at least one of the scores is misleading and that blind reliance on averages is not advisable for this particular paper. Such readers would be encouraged to read and judge the paper for themselves, potentially leaving new, genuine reviews clarifying the epistemic contribution of the paper.

More generally, while we argue that using the overall average reviewer score from crowd-sourced peer review will give better quality judgments than journal-solicited peer review, it is emphatically not part of our proposal that overall average scores should be the only thing available. A lot of additional information should be made available to potential readers, so they can freely choose which metrics they think are more informative. This includes the median and mode of reviewer scores, the content and score of each individual review, the total number of reviews and the number of reviews by endorsed reviewers, the ranking of the paper relative to other papers in its subfield, and ratings of the helpfulness of individual reviewers.

Combining metrics will provide additional insight relevant to the problem of manipulation. For example, the median reviewer score is robust to manipulation as long as less than half the reviewer scores are manipulated. Thus, papers with big gaps between the average score and the median score should be treated with care. For another example, one would typically expect better papers to receive more reviews, as readers are attracted by the high score. Thus, papers with an unusually high number of reviews but a low average score should raise suspicions. The same thing goes for a paper where most of the reviews come from reviewers who have never reviewed anything else. Using this information, we think academics will be able to make use of crowd-sourced peer review to identify and read high-quality papers with minimal interference from manipulation.

There may also be technological or procedural ways to monitor and prevent internet mobbing and the like. First, statistical software might be used to detect highly-correlated votes (as might be the case when a particular article initially receives a good proportion of positive reviews only to be followed by a quick succession of overwhelmingly negative reviews). Second, reviewers might be afforded the ability to 'flag' particular reviews as suspicious.

Third, the proprietors or 'section editors' of the arXiv-like site we propose could be alerted by their site's software to a large number of reviews for a given paper being posted by visitors from particular outside websites (such as a Reddit or Twitter thread advocating for 'review bombing' the paper). Although we would not necessarily advocate deleting suspicious reviews (which Rotten Tomatoes did in early 2019)—as suspicious reviews could well be genuine—proprietors could flag individual papers as potential victims of illicit reviewer behavior, alerting other readers and reviewers to the possibility that the paper's scores may be corrupted.

Fourth, moderation to remove abusive (for instance sexist or otherwise bigoted) reviews is within the spirit of our proposal. Such reviews would not add to the evaluation of the paper. However, in addition to their intrinsic cultural or moral harm, they could harm the epistemic performance of science by contributing to some groups being less able to have their claims fairly assessed by the scientific community. Whatever potential for abuse of power exists in this content moderation system is surely no worse than what present editors and reviewers have.

To be clear, although we think these interventions might be effective, we do not commit to any particular scheme for addressing illicit reviewing practices. These matters are probably best addressed through practical experimentation of the sort that review aggregators like Rotten Tomatoes have and continue to do.

On that note, we are optimistic that our model's viability and any implementation issues could be examined empirically. One possibility would be for a central repository to roll out a 'beta' version of the system, implementing the model we outline above in a specific subfield, recruiting a batch of expert reviewers in that subfield, randomly selecting new unpublished manuscripts to receive reviews and ratings, and permitting those manuscripts to be reviewed and rated by expert reviewers and others.

If this beta is rolled out prominently, it might encourage participation from significant numbers of people in the profession. Such a trial could run for six months or longer. In addition to gaining feedback from the academic community on what works well and what does not, the implementers could collect data for statistical analysis over several years (say, 1–3 years). They might use citation counts and other measures of engagement to look at whether papers rated well by expert reviewers and general readership tend to have a greater impact on the field than low-rated papers. Or they might use textual analysis to see whether follow-up work by other authors is largely positive (constructive) or negative (critical), how this correlates with reviewer ratings, and how this compares to papers published using journal-solicited peer review. Although scientific quality is difficult to gauge, this could provide systematic statistical evidence (albeit defeasible) of whether the trial system tends to select higher-quality work than traditional peer review. Further betas (implementing tweaks in response to feedback) could be carried out and examined if the data appears promising.

## 7.2 Greater Average Competence in Journal-Solicited Peer Review?

Our arguments assumed that *reviewer competence* is randomly distributed throughout the population of possible reviewers. More specifically, in §4 we assumed that the average probability that a reviewer in journal-solicited peer review will arrive at an accurate judgment of a paper's merit is the same as the average probability of an accurate judgment from a crowd-sourced reviewer. Similarly, in §5–6 we assumed that the judgment of a randomly selected peer reviewer follows the same probability distribution whether the reviewer is journal-solicited or crowd-sourced.

However, some may doubt this. First, some might suggest that journal editors are likely to select substantially more competent reviewers than the average reviewer in the population, e.g., by commissioning reviews from the most accomplished figures in the field. These reviewers, due to their exceptional achievements, may perhaps be expected to judge a given paper's merits more accurately or with less bias.[9] Second, some might argue that insofar as journal-solicited peer review commissions reviews by specialists in the paper's field, those specialists are likely to have higher accuracy or be less biased than a pool of reviewers that includes non-specialists. If as a result of either mechanism journal-solicited peer review reliably selects reviewers who are more competent than crowd-sourced peer review, then our arguments do not go through. In order to successfully defend a Condorcet-style argument in this case, we would need to show that the *accuracy increase* generated by increasing the size of the jury pool (through crowd-sourced peer review) is *greater* than the accuracy increase generated by how journal editors select reviewers.

Our reply to this concern is two-fold. First, the balance of present evidence suggests the empirical claim that journals select better reviewers is not true. Second, there are a number of *prima facie* reasons to believe that journal-solicited reviewers are likely to be *more* biased than the population from which crowd-sourced peer review might draw.

Arguments that journal-solicited reviewers are likely to be more competent to evaluate papers than readers at large tend to come from the armchair. However, two sources of empirical evidence collectively cast doubt on this intuition. First, empirical studies on the quality of journal-solicited reviews suggest very low interrater reliability (Lee et al. 2013, pp. 5–6; Bornmann 2011, p. 207). Interrater reliability measures the level of agreement between different reviewers judging the same paper, which is relevant here because disagreement imposes an upper bound on reviewer accuracy. In one study, interrater reliability barely exceeded chance (Kravitz et al. 2010). In terms

[9]See `https://philosopherscocoon.typepad.com/blog/2018/12/incentivizing-better-reviewer-behavior.html`, where David Bourget notes that journal editors may aim to select more accomplished, senior scholars as reviewers for this reason.

of the basic Condorcet model, this corresponds to probabilities of voting correctly barely exceeding .5.

Second, anecdotal reports suggest that reviewers at highly-selective journals *routinely misjudge* papers that larger audiences have judged more accurately. Gans and Shepherd (1994) report how a variety of classic (including Nobel Prize-winning) economics articles were systemically rejected by top-ranked journals in the field. Anecdotally, this also happens in academic philosophy—for instance, Jason Stanley reported that four of his articles rejected from multiple highly-ranked journals are now among the twenty most-cited articles in those very journals since 2000.[10] This phenomenon is further illustrated by a study in which twelve articles were submitted to the same highly-ranked psychology journals that already published them (Peters and Ceci 1982). Of the nine that made it past desk-rejection, eight of the papers were *rejected* without reviewers or editors realizing that the journal had already published them—in many cases on the basis of "serious methodological flaws".

How can we square the intuitive thought that prestigious journals will select the most competent reviewers with the empirical research and anecdotes indicating that journal-solicited reviewers are highly unreliable? Although we can only speculate, there are reasons to think that journal-solicited reviewers are likely to be more biased than average. First, insofar as journals tend to commission comparatively accomplished reviewers, these reviewers plausibly have particular biases. These may include a bias for their own views and work (as they have a vested interest in their views remaining influential) and biases for particular arguments (e.g., by authors they admire or are personally acquainted with).

Second, the journal-solicited peer review system arguably introduces its own biases. First, when reviewers know that the journal they are reviewing

[10] `https://philosopherscocoon.typepad.com/blog/2015/09/stanley-on-peer-review.html`.

for has a high rejection rate, their presumption may be that they should recommend rejection unless they are convinced the paper is excellent. This potentially lowers the risk of accepting bad papers but increases the risk of rejecting good papers. Anecdotally, this is borne out in journal-solicited peer review—as illustrated again by significant numbers of seminal economics papers being rejected by journals. Second, journal editors plausibly have grounds to err on the side of rejection as well, as publishing a bad paper may harm the journal's reputation. Finally, by explicitly selecting 'specialists' (people who have already published in the area) to review papers, journal-solicited peer review runs a serious risk of *groupthink*. Indeed, consider two causal antecedents of groupthink: *group cohesiveness* and *insularity* (Janis 1972). Both are arguably embedded in journal-solicited peer review insofar as editors seek out specialist reviewers—individuals who have chosen to work in a similar area, attend conferences together, and share unpublished work among each other. Conversely, journal-solicited peer review does not appear to regularly involve a practice empirical evidence suggests serves to *prevent* groupthink: the stimulation of *intellectual conflict* (Turner and Pratkanis 1994) through the inclusion of outside perspectives (Janis 1972, pp. 209–215). Because crowd-sourced peer review would not only invite more people to review papers, but also give reviewers the opportunity to *contest* each other's reviews (see §7.3), it would be more likely to generate the kind of intellectual conflict necessary for combating groupthink. Similar arguments have been made by philosophers under the label of *epistemic diversity*: researchers actively pursuing opposing theories or methodologies is often fruitful (Feyerabend 1975, Lakatos 1978, Longino 1990, Kitcher 1993, Zollman 2010).

Thus, there are reasons to believe that, even if journal editors recruit 'the best reviewers' (which we have no clear empirical evidence for), journal-solicited peer review introduces biases that are likely to be less pronounced or more evenly distributed in a general population of readers in the discipline.

The 'beta' experiment suggested at the end of §7.1 could help support (or undermine) these claims.

To be clear, readers in the general academic population will tend to have biases of their own. Like journal-solicited reviewers, they plausibly have vested interests in advancing their own views, idiosyncrasies, and so on. The point of our argument is not primarily that crowd-sourced reviewers are less biased than journal-solicited reviewers. Our argument is (1) there is no clear evidence that journal-solicited reviewers are more accurate or less biased than the reviewer pool at large; (2) there is *some* anecdotal evidence they may in fact be less accurate or more biased; but more importantly (3) whatever biases there are in an academic population, our three jury theorems suggest that in a larger jury, these biases tend to *cancel each other out* more. With a small jury (e.g., two reviewers), distorting biases are much more likely to produce the wrong verdict.

Consequently, we submit that there is no clear support for the proposition that journal-solicited reviewers are more accurate in their judgments of paper quality than academics at large. Given that the evidence is unclear, we believe it is more appropriate to assume that accuracy and bias *are* randomly distributed in an academic population—unless and until clear evidence is provided to the contrary.

## 7.3 Failures of Independence in Crowd-Sourced Peer Review?

Another worry is that the jury theorems hold only when votes are probabilistically independent. This assumption seems plausible for journal-solicited peer review: each reviewer judges a given paper without knowledge of what other reviewers think. Conversely, with crowd-sourced peer review, one (influential) reviewer's evaluation of a paper may affect others—potentially generating a snowball effect. When votes in a jury are correlated, the collective competence of a jury may be *lower* than the competence of individual jurors

(Kaniovski and Zaigraev 2011).

Our reply begins by noting some technical points. First, correlated votes only undermine the Condorcet Jury Theorem when reviewer competence is low and correlation between their opinions high (Kaniovski and Zaigraev 2011). Second, for the results in §5–6 to be undermined, an even stronger condition needs to hold: the correlation needs to *increase* systematically with the number of reviewers. If reviewer opinions are correlated but the correlation coefficient is constant, the first and more important part of our theorems—that the probability of a judgment close to the correct one increases with the number of reviewers—still holds, even if the second part—that this probability goes to one in the limit—fails. Third, as Estlund (1994) points out, the mere fact that early reviewers might influence the opinion of later reviewers is not necessarily inconsistent with *probabilistic* independence of reviewer judgments. Estlund thus shows that the 'correlation' (in the intuitive sense) induced by early influential reviewers need not entail the specific type of correlation that would undermine our argument.

Questions about reviewer competence and correlation between their opinions would ideally be settled empirically. As discussed earlier, the empirical finding of low interrater reliability suggests that reviewer competence may indeed be low. We are not aware of any empirical work on correlations among reviewer opinions in a crowd-sourced model of peer review (perhaps the 'beta' experiment discussed in §7.1 could provide some). However, potentially relevant evidence comes from recent studies using surveys and prediction markets to see whether academics can predict the results of replications (Dreber et al. 2015, Camerer et al. 2016, 2018, Forsell et al. 2019). There is a close analogy here, as the surveys measure individual academics' opinions without information about what others think (as in journal-solicited peer review) while the prediction markets measure opinions in the presence of information about others (like crowd-sourced peer review). The results provide reason for optimism. Prediction markets tend to do at least as well as surveys, suggesting that information about other academics' opinions does not introduce the sort of correlation that undermines the benefits of large numbers.

We add some speculative reasons to doubt that the correlation of votes under crowd-sourced peer review would be high, or that it would increase with the number of reviewers. Academic training and incentives encourage academics to *evaluate* and *counter* arguments they find unpersuasive. Consider the kinds of discussions that occur in journals after an article or book is published—e.g., John Rawls' book *A Theory of Justice*. Some commentators were highly critical of Rawls' arguments (e.g., Hare 1973a,b, Nowell-Smith 1973, Barry 1973); others more sympathetic (e.g., Mandelbaum 1973). Scholars then began debating *each other* on particular 'flaws' in Rawls' book (e.g., Bedau 1975). Eventually, a *general consensus* emerged that Rawls' work is important yet flawed in particular ways. Crowd-sourced reviewers could be expected to do something similar: present their own evaluations of a given piece of work and *contest* others. Assuming such practices are central to crowd-sourced peer review, 'votes' will tend to be poorly correlated.

Moreover, we expect expert reviewers in particular to form their opinions independently. Recall that in §7.1 we introduced expert reviewers (based on colleagues' endorsements) whose scores potential readers may want to consider separately as a way to guard against internet trolls. Now a genuine expert cannot just be somebody who happens to have more true beliefs than average about a domain: such a person would lack what has been called *contributory expertise* in the literature (Collins et al. 2016). The contributory expert must also know the methods and heuristics one ought to adopt to reliably arrive at true beliefs about the topic (Licon 2012, p. 451). We take it that knowledge of these methods is meant to make the contributory expert someone whose beliefs are relatively independent of their peers, conditional on the truth. This point is strengthened by Estlund (1994), who observes that in the presence of even fairly high degrees of deference to influential reviewers, the kind of reviewer independence needed for our theorems to apply can be maintained as long as reviewers add at least a modicum of their own (truth-tracking) insight.

If this is granted, then at least the set of expert reviewers will satisfy the independence condition, and hence all the conditions for our jury theorems to apply. Assuming that the average number of expert reviewers under crowd-sourced peer review is at least as high as the average number of reviewers under journal-solicited peer review, it follows (at minimum) that basing one's opinion on the average expert reviewer score is better than basing one's opinion on journal-solicited peer review.

We also think that, barring cases of internet mobs, correlation among non-expert reviewers will tend to be low, and so basing one's opinion on the overall average score is even better than basing one's opinion on the average expert reviewer score (as this will make for an even larger jury). Whether we are right about this will become clear over time as we experiment with crowd-sourced peer review.

# 8 Conclusion

We have argued that if the presuppositions which peer review is based upon are correct, then three jury theorems suggest that an open, crowd-sourced model of post-publication peer review would do better at directing researchers towards better work. This leaves two major questions open. First, are the presuppositions of peer review correct? The brief arguments we gave for them here should be supplemented with more sustained, and often empirical, socio-epistemic inquiry. However, it should be noted that if these presuppositions turn out to be false, journal-solicited peer review is arguably even less defensible than our argument suggests. Second, is it actually desirable to direct researchers towards the best work? This might seem good for individual researchers but that is not yet to argue for its socio-epistemic optimality (Mayo-Wilson et al. 2011). We leave both these projects for future work.

We conclude by noting that the practical difficulties with implementing our proposal are substantial but not insurmountable. Online forums for public peer review (such as the arXiv in physics) already exist, even serving as a primary point of publication in some fields. It is not beyond our capacities to add the necessary features on some expanded version of these venues. What is presently lacking is the will. We thus hope that our paper goes towards building this will, such that researchers will become able to take further and more systematic advantage of the combined wisdom of the academic community.

# References


Jason McKenzie Alexander, Johannes Himmelreich, and Christopher Thompson. Epistemic landscapes, optimal search, and the division of cognitive labor. *Philosophy of Science*, 82(3):424–453, 2015. URL `https://doi.org/10.1086/681766`.

J. Scott Armstrong. Combining forecasts. In J. Scott Armstrong, editor, *Principles of Forecasting: A Handbook for Researchers and Practitioners*, pages 417–440. Kluwer, New York, 2001.

Shahar Avin. Centralized funding and epistemic exploration. *The British Journal for the Philosophy of Science*, 70(3):629–656, 2019. URL `https://doi.org/10.1093/bjps/axx059`.

Brian M. Barry. *The Liberal Theory of Justice: A Critical Examination of the Principal Doctrines in a Theory of Justice by John Rawls*. Clarendon Press, Oxford, 1973.

Hugo Adam Bedau. Review of the liberal theory of justice: A critical examination of the principal doctrines in a theory of justice by John

Rawls by Brian Barry. *Philosophical Review*, 84(4):598–603, 1975. URL https://doi.org/10.2307/2183864.

Lutz Bornmann. Scientific peer review. *Annual Review of Information Science and Technology*, 45(1):197–245, 2011. URL https://doi.org/10.1002/aris.2011.1440450112.

Colin F. Camerer et al. Evaluating replicability of laboratory experiments in economics. *Science*, 351(6280):1433–1436, 2016. URL https://doi.org/10.1126/science.aaf0918.

Colin F. Camerer et al. Evaluating the replicability of social science experiments in *Nature* and *Science* between 2010 and 2015. *Nature Human Behaviour*, 2(9):637–644, 2018. URL https://doi.org/10.1038/s41562-018-0399-z.

David Christensen. Epistemology of disagreement: The good news. *The Philosophical Review*, 116(2):187–217, 2007. URL https://doi.org/10.1215/00318108-2006-035.

Robert T. Clemen. Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting*, 5(4):559–583, 1989. URL https://doi.org/10.1016/0169-2070(89)90012-5.

Stewart Cohen. A defense of the (almost) equal weight view. In David Christensen and Jennifer Lackey, editors, *The Epistemology of Disagreement: New Essays*, chapter 5, pages 98–119. Oxford University Press, Oxford, 2013.

Harry Collins, Robert Evans, and Martin Weinel. Expertise revisited, part II: Contributory expertise. *Studies in History and Philosophy of Science Part A*, 56:103–110, 2016. URL https://doi.org/10.1016/j.shpsa.2015.07.003.

Marquis de Condorcet. *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix.* Imprimerie Royale, Paris, 1785.

Jacques Distler and Skip Garibaldi. There is no "Theory of Everything" inside $E_8$. Manuscript, 2009. URL https://arxiv.org/abs/0905.2658.

Anna Dreber et al. Using prediction markets to estimate the reproducibility of scientific research. *Proceedings of the National Academy of Sciences*, 112(50):15343–15347, 2015. URL https://doi.org/10.1073/pnas.1516179112.

Adam Elga. Reflection and disagreement. *Noûs*, 41(3):478–502, 2007. URL https://doi.org/10.1111/j.1468-0068.2007.00656.x.

David M. Estlund. Opinion leaders, independence, and Condorcet's jury theorem. *Theory and Decision*, 36(2):131–162, 1994. ISSN 1573-7187. URL https://doi.org/10.1007/BF01079210.

Paul Feyerabend. *Against Method.* New Left Books, London, 1975.

Eskil Forsell et al. Predicting replication outcomes in the Many Labs 2 study. *Journal of Economic Psychology*, 75:102117, 2019. URL https://doi.org/10.1016/j.joep.2018.10.009.

Joshua S. Gans and George B. Shepherd. How are the mighty fallen: Rejected classic articles by leading economists. *Journal of Economic Perspectives*, 8(1):165–179, 1994. URL https://doi.org/10.1257/jep.8.1.165.

Todd A. Gibson. Post-publication review could aid skills and quality. *Nature*, 448(7152):408, 2007. URL https://doi.org/10.1038/448408d.

R. M. Hare. Rawls' theory of justice—I. *The Philosophical Quarterly*, 23 (91):144–155, 1973a. URL https://doi.org/10.2307/2217486.

R. M. Hare. Rawls' theory of justice—II. *The Philosophical Quarterly*, 23 (92):241–252, 1973b. URL `https://doi.org/10.2307/2218002`.

Stephan Hartmann and Jan Sprenger. Judgment aggregation and the problem of tracking the truth. *Synthese*, 187(1):209–221, 2012. URL `https://doi.org/10.1007/s11229-011-0031-5`.

Remco Heesen. Why the reward structure of science makes reproducibility problems inevitable. *The Journal of Philosophy*, 115(12):661–674, 2018. URL `https://doi.org/10.5840/jphil20181151239`.

Remco Heesen. The necessity of commensuration bias in grant peer review. Manuscript, 2019. URL `http://philsci-archive.pitt.edu/15930/`.

Remco Heesen and Liam Kofi Bright. Is peer review a good idea? *The British Journal for the Philosophy of Science*, forthcoming. URL `https://doi.org/10.1093/bjps/axz029`.

Remco Heesen, Liam Kofi Bright, and Andrew Zucker. Vindicating methodological triangulation. *Synthese*, 196(8):3067–3081, 2019. URL `https://doi.org/10.1007/s11229-016-1294-7`.

Irving L. Janis. *Victims of Groupthink: A Psychological Study of Foreign-Policy Decisions and Fiascoes*. Houghton Mifflin, Boston, 1972.

Serguei Kaniovski and Alexander Zaigraev. Optimal jury design for homogeneous juries with correlated votes. *Theory and Decision*, 71(4):439–459, 2011. URL `https://doi.org/10.1007/s11238-009-9170-2`.

Philip Kitcher. *The Advancement of Science: Science without Legend, Objectivity without Illusions*. Oxford University Press, Oxford, 1993.

Richard L. Kravitz et al. Editorial peer reviewers' recommendations at a general medical journal: Are they reliable and do editors care? *PLoS*

*ONE*, 5(4):e10072, 2010. URL https://doi.org/10.1371/journal.pone.0010072.

Thomas S. Kuhn. *The Essential Tension: Selected Studies in Scientific Tradition and Change*. University of Chicago Press, Chicago, 1977.

Imre Lakatos. *The Methodology of Scientific Research Programmes*. Cambridge University Press, Cambridge, 1978.

Carole J. Lee, Cassidy R. Sugimoto, Guo Zhang, and Blaise Cronin. Bias in peer review. *Journal of the American Society for Information Science and Technology*, 64(1):2–17, 2013. URL https://doi.org/10.1002/asi.22784.

Keith Lehrer and Carl Wagner. *Rational Consensus in Science and Society: A Philosophical and Mathematical Study*, volume 24 of *Philosophical Studies Series in Philosophy*. D. Reidel, Dordrecht, 1981.

Jimmy Alfonso Licon. Sceptical thoughts on philosophical expertise. *Logos & Episteme*, 3(3):449–458, 2012. URL https://doi.org/10.5840/logos-episteme20123325.

A. Garrett Lisi. An exceptionally simple Theory of Everything. Manuscript, 2007. URL https://arxiv.org/abs/0711.0770.

Christin List and Robert E. Goodin. Epistemic democracy: Generalizing the Condorcet Jury Theorem. *Journal of Political Philosophy*, 9(3):277–306, 2001. URL https://doi.org/10.1111/1467-9760.00128.

Helen E. Longino. *Science as Social Knowledge: Values and Objectivity in Scientific Inquiry*. Princeton University Press, Princeton, 1990.

Maurice Mandelbaum. Review of A Theory of Justice by John Rawls. *History and Theory*, 12(2):240–250, 1973. URL https://doi.org/10.2307/2504913.

Carlo Martini and Jan Sprenger. Opinion aggregation and individual expertise. In Thomas Boyer-Kassem, Conor Mayo-Wilson, and Michael Weisberg, editors, *Scientific Collaboration and Collective Knowledge*, chapter 9, pages 180–201. Oxford University Press, Oxford, 2017.

Conor Mayo-Wilson, Kevin J. S. Zollman, and David Danks. The independence thesis: When individual and social epistemology diverge. *Philosophy of Science*, 78(4):653–677, 2011. URL `https://doi.org/10.1086/661777`.

Shinichi Mochizuki. Inter-universal Teichmüller theory I: Construction of Hodge theaters. Manuscript, 2012a. URL `http://www.kurims.kyoto-u.ac.jp/~motizuki/papers-english.html`.

Shinichi Mochizuki. Inter-universal Teichmüller theory II: Hodge-Arakelov-theoretic evaluation. Manuscript, 2012b. URL `http://www.kurims.kyoto-u.ac.jp/~motizuki/papers-english.html`.

Shinichi Mochizuki. Inter-universal Teichmüller theory III: Canonical splittings of the log-theta-lattice. Manuscript, 2012c. URL `http://www.kurims.kyoto-u.ac.jp/~motizuki/papers-english.html`.

Shinichi Mochizuki. Inter-universal Teichmüller theory IV: Log-volume computations and set-theoretic foundations. Manuscript, 2012d. URL `http://www.kurims.kyoto-u.ac.jp/~motizuki/papers-english.html`.

Michael Morreau. Democracy without enlightenment: A jury theorem for evaluative voting. *Journal of Political Philosophy*, forthcoming. URL `https://doi.org/10.1111/jopp.12226`.

Brian A. Nosek and Yoav Bar-Anan. Scientific utopia: I. Opening scientific communication. *Psychological Inquiry*, 23(3):217–243, 2012. URL `https://doi.org/10.1080/1047840X.2012.692215`.

P. H. Nowell-Smith. Review symposium: I—a theory of justice? *Philosophy of the Social Sciences*, 3(4):315–329, 1973. URL `https://doi.org/10.1177/004839317300300403`.

Cailin O'Connor and Justin Bruner. Dynamics and diversity in epistemic communities. *Erkenntnis*, 84(1):101–119, 2019. URL `https://doi.org/10.1007/s10670-017-9950-y`.

Samir Okasha. Theory choice and social choice: Kuhn versus Arrow. *Mind*, 120(477):83–115, 2011. URL `https://doi.org/10.1093/mind/fzr010`.

Douglas P. Peters and Stephen J. Ceci. Peer-review practices of psychological journals: The fate of published articles, submitted again. *Behavioral and Brain Sciences*, 5:187–195, 1982. ISSN 1469-1825. URL `https://doi.org/10.1017/S0140525X00011183`.

Richard Pettigrew. On the accuracy of group credences. In Tamar Szabó Gendler and John Hawthorne, editors, *Oxford Studies in Epistemology*, volume 6, chapter 6, pages 137–160. Oxford University Press, Oxford, 2019.

Felipe Romero. Can the behavioral sciences self-correct? A social epistemic study. *Studies in History and Philosophy of Science Part A*, 60:55–69, 2016. URL `https://doi.org/10.1016/j.shpsa.2016.10.002`.

Peter Scholze and Jakob Stix. Why *abc* is still a conjecture. Manuscript, 2018. URL `http://www.kurims.kyoto-u.ac.jp/~motizuki/SS2018-08.pdf`.

Daniel J. Singer. Diversity, not randomness, trumps ability. *Philosophy of Science*, 86(1):178–191, 2019. URL `https://doi.org/10.1086/701074`.

Eran Tal. Measurement in science. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Fall 2020 edition, 2020. URL `https://plato.stanford.edu/archives/fall2020/entries/measurement-science/`.

Johanna Thoma. The epistemic division of labor revisited. *Philosophy of Science*, 82(3):454–472, 2015. URL `https://doi.org/10.1086/681768`.

Marlene E. Turner and Anthony R. Pratkanis. Social identity maintenance prescriptions for preventing groupthink: Reducing identity protection and enhancing intellectual conflict. *International Journal of Conflict Management*, 5(3):254–270, 1994. URL `https://doi.org/10.1108/eb022746`.

Michael Weisberg and Ryan Muldoon. Epistemic landscapes and the division of cognitive labor. *Philosophy of Science*, 76(2):225–252, 2009. URL `https://doi.org/10.1086/644786`.

Kevin J. S. Zollman. The epistemic benefit of transient diversity. *Erkenntnis*, 72(1):17–35, 2010. URL `https://doi.org/10.1007/s10670-009-9194-6`.