The Necessity of Commensuration Bias in Grant Peer Review*

Remco Heesen[†‡]

October 10, 2018

1 Introduction

Peer review is one of the linchpins of the social organization of science. Whether as a grant proposal, manuscript, or conference abstract, just about every piece of scientific work passes through peer review, often multiple times. Yet philosophers of science have paid surprisingly little attention to peer review (exceptions include Zollman 2009, Avin forthcoming).

The linchpin role of peer review means that it is particularly important to understand *biases* in peer review. There is now a fairly large empirical literature studying gender bias, racial bias, prestige bias, publication bias, and many other forms of bias (see Lee et al. 2013, for a review). In addition to empirical questions, there are conceptual questions to be answered about defining, identifying, and distinguishing different biases and analyzing their

*This work was supported by an Early Career Fellowship from the Leverhulme Trust and the Isaac Newton Trust.

[†]Department of Philosophy, School of Humanities, University of Western Australia, Crawley, WA 6009, Australia. Email: remco.heesen@uwa.edu.au.

[‡]Faculty of Philosophy, University of Cambridge, Sidgwick Avenue, Cambridge CB3 9DA, UK.

potential effects (Lee et al. 2013, Saul 2013, Heesen 2018, Heesen and Romeijn 2018).

This paper focuses on a new type of bias identified by Lee (2015), which she calls *commensuration bias*. Commensuration is the activity of aggregating different quantities into a single number. Noting that many peer review processes ask reviewers to score submissions on some criteria as well as giving an overall score, Lee introduces commensuration bias to capture situations in which the act of commensuration is a locus at which bias gets introduced. She points at a number of phenomena that seem to fall under this label.

Lee distinguishes two types of commensuration bias. The first type, which is her primary focus, refers to peer review practices that privilege one of the individual criteria. Lee (2015, section 3) argues that current journal peer review practices overweight "intellectual significance" (narrowly interpreted as statistical significance), resulting in biased estimates of effect sizes in the published literature. Grant agencies, Lee goes on to argue, overweight methodological criteria relative to novelty, which in the aggregate ends up promoting conservatism.

A potential difficulty in identifying this type of commensuration bias is that it requires a substantive view on what counts as overweighting a criterion. For this reason I will invoke the first type of commensuration bias only in the particularly stark case where a peer reviewer gives a higher overall score to a grant proposal (or paper, but I will focus mostly on grant proposals) whenever it scores higher on the privileged criterion. This is especially problematic because it reduces the other criteria to tie-breakers, contrary to the (intuitive) idea that all criteria should receive some genuine weight.

Second, reviewer idiosyncrasies or biases may lead to "deviation from the impartial weighting of peer review criteria" (Lee 2015, 1273). To make this a little more precise, I will say that commensuration bias has occurred whenever two grant proposals receive identical scores on the individual criteria, but their overall scores differ. To illustrate this definition, consider the

following two fairly different ways of instantiating it.

Suppose a peer reviewer is (explicitly or implicitly) socially biased, that is, her judgment of the quality of a grant proposal is affected by prejudice based on the gender, race, or other social characteristics of the scientist or scientists responsible for the work. One point in the peer review process where such bias might have an effect is in commensurating criteria scores to an overall score. The reviewer might go so far as to rate one proposal (written by a woman, say) higher than another proposal (by a man) on all criteria, but nevertheless give a higher overall score to the latter proposal. This would be an example of the second type of commensuration bias.

For a different example, consider the principle that identical individual criteria scores should lead to identical overall scores. This entails that if two grant proposals' individual scores are held fixed, their overall scores are similarly fixed. Suppose a peer reviewer violates this in the following specific way: the overall scores of two proposals depend, in addition to their individual scores, also on the individual scores of one or more other proposals. This is an example of the second type of commensuration bias because it violates the principle of identical overall scores for identical individual scores. More generally, it militates against the widespread idea that a proposal's score can be determined by looking (only) at it.

My aim in this paper is to use social choice theory to argue that rather than being a fringe phenomenon, commensuration bias is impossible or at least very hard to avoid in any peer review context where multiple criteria or multiple reviewers are used. Section 2 sets up the social choice framework, focusing on a single reviewer scoring a set of grant proposals. Section 3 gives the main argument, based on a well-known impossibility theorem. Section 4 considers the case of multiple reviewers and section 5 combines the two cases studied in the preceding sections. Section 6 contrasts peer review of grant proposals (my main focus in this paper) with peer review at journals, and section 7 concludes.

2 Peer Review As an Aggregation Problem

Consider a peer reviewer a_1, tasked with scoring m grant proposals $x, y, \ldots$. Suppose that the funding agency asks her to score the proposals on k criteria $c_1, \ldots, c_k$. For example, the National Institutes of Health (2017) use the following criteria: "significance", "investigator(s)" (suitability of the applicants to carry out the research), "innovation", "approach", and "environment" (suitability of the research environment).

Reviewer a_1 reads the proposals and scores them on the various criteria. For any proposal x, I write $s_{1j}(x)$ for the score reviewer a_1 assigns to that proposal on criterion c_j. The scores $s_{1j}(\cdot)$ are assumed to be real numbers. The index "1" for reviewer a_1 is just a placeholder for now; other reviewers will be introduced in section 4.

In addition to the criteria scores, the reviewer is asked to give an aggregate or overall score to each proposal. At the National Institutes of Health (2017), for example, this is called the overall impact score. I write $s_1(x)$ for the overall score assigned to proposal x, which is again assumed to be a real number.

The overall scores assigned by reviewer a_1 induce a ranking of the grant proposals: proposals with a higher score are implicitly judged to be better than proposals with a lower score. This induced ranking will be the main object of interest in the next section, so I introduce some notation for it. For any two proposals x and y, xP_1y denotes the proposition $s_1(x) > s_1(y)$, i.e., "reviewer a_1 ranks x strictly higher than y". Similarly, xR_1y denotes $s_1(x) \geq s_1(y)$ or "x ranks at least as high as y", and xI_1y denotes $s_1(x) = s_1(y)$ or "x and y rank equally".

In social choice theory, a reviewer's individual criteria scores (i.e., the collection of each of her criteria scores for each proposal) is called a *profile*. A function, defined on some given domain of profiles, which assigns to each profile a corresponding set of overall scores, will be called a *commensuration function*.

The first substantive question I address is: how much information is con-

tained in the individual criteria scores? In other words, which profiles should be treated as identical by the commensuration function? The question breaks down into two further questions. What kind of scale are the criteria scores measured on? And can scores be meaningfully compared across different criteria? I take the two questions in turn.

Numerical quantities are usually regarded as being measured on one of four types of scales: ordinal, cardinal, ratio, or absolute (e.g., Tal 2017, section 3.2). An *ordinal* scale orders the objects being measured by size (in this case: orders the grant proposals from best to worst on a given criterion) but the magnitude of differences is meaningless. In the present context this means that if two profiles differ only in that one is obtained from the other by applying a positive monotone transformation to the criteria scores the commensuration function should give them the same overall ranking; otherwise the ranking would be sensitive to meaningless differences in the way scores are represented numerically.

A *cardinal* (or interval) scale differs from an ordinal scale in that the size of differences is meaningful. If a criterion is measured on a cardinal scale only positive affine transformations can be applied without loss of information. The standard example is temperature: the Celsius, Fahrenheit, and Kelvin scales are related by positive affine transformations.

A *ratio* scale has a meaningful zero. As a result statements like "this object's measurement is twice that object's measurement" make sense when measurements are on a ratio scale. In this case multiplication by a positive constant is the only transformation that can be applied without loss of information. Standard examples are length and weight.

An *absolute* scale has a meaningful zero and a meaningful unit. This yields a unique scale in the sense that only the identity transformation can be applied without loss of information.

For the types of criteria considered here, it seems quite unrealistic to me to assume that scores on a more informative scale than a cardinal one. For ratio

or absolute scales to apply, there would have to be an empirically meaningful sense in which grant proposals could be said to have zero significance, or zero innovativeness, or for the investigator to have zero suitability to carry out the research. Or equivalently, statements like "This proposal is twice as innovative as that one" or "University X is twice as suitable for carrying out proposal x as Institute Y is for proposal y" would have to be among the types of claims peer reviewers make. For the types of criteria scored in the process of grant proposal peer review, however, I do not think that level of information is typically available. So I will assume throughout this paper that criteria scores (and overall scores) are measured on an ordinal or cardinal scale (for the formal results, it does not matter which).

How about intercriteria comparability? Here the question is whether statements like "proposal x is more significant than that the applicant of proposal y is suitable" are meaningful, or even something like "the difference between proposal x and proposal y's score on innovation is larger than the difference in their scores on approach".

If reviewers are given a particular numerical scale to score proposals on (say, a 1 to 5 scale), these types of statements could technically be used to compare proposals' scores on different criteria. But I do not think peer reviewers would typically regard such claims as useful or informative. They would more likely say something like "While we have scored the proposals on this scale, differences in scores should be interpreted more qualitative than that". So I assume that there is no intercriteria comparability.

To be clear, if some degree of intercriteria comparability could plausibly be taken to be implicit in the individual criteria scores, or if these scores could realistically be interpreted as being measured on a ratio scale or an absolute scale, then the results to be discussed in the next sections would not hold. See Sen (1970), List (2004, section 3), or Okasha (2011, section 6) for further discussion of measurability and intercomparability.

3 Aggregating an Individual Reviewer's Scores

What properties should a commensuration function have? In particular, what needs to be true for it to be free of commensuration bias?

Universal Domain (U) The domain of the commensuration function is the set of all possible profiles of criteria scores.

This requires that no combination of criteria scores is ruled out in advance. What this means is perhaps best explained by considering what the alternatives are. One way to circumvent (U) is to declare certain combinations of criteria scores impossible either descriptively ("any innovative proposal must have a suitable investigator by definition so we will never see a proposal with a high score for innovation but a low score for investigator") or normatively ("reviewers should avoid sending mixed messages by giving very high scores on some criteria and very low scores on others"). Another is to declare proposals with certain combinations of criteria scores unratable, giving them no overall score at all.

Neither of these routes is very attractive. While violating (U) perhaps does not constitute a bias in the same sense as violating the other requirements does, I think any reasonable commensuration function describing how a peer reviewer at a grant funding agency approaches commensuration should avoid ruling out combinations of criteria scores in advance, and hence should satisfy (U).

Weak Pareto (P) If a proposal scores higher than another proposal on all criteria it should get a higher overall score, i.e., $s_{1j}(x) > s_{1j}(y)$ for all criteria c_j entails xP_1y.

If the reviewer unanimously ranks a proposal higher than another on all criteria, it would be quite strange for her to then turn around and give a lower overall score to the former proposal. In such a case one might reasonably say

that some kind of bias has influenced the way the reviewer has moved from the criteria scores to the overall scores.

For example, we could imagine a reviewer with (explicit or implicit) gender bias rating a proposal written by a woman higher on each of the criteria, but assigning it a lower overall score than a proposal written by a man, as described in section 1. In such a case we can properly speak of commensuration bias, i.e., a bias that affects specifically the step where multiple criteria scores are aggregated into an overall score, as the overall score can be seen to be (more) biased when compared to the criteria scores.

Similarly, a reviewer might violate (P) due to racial bias or prestige bias. Alternatively, she might violate (P) in a more idiosyncratic way, giving a higher score to some proposal with lower criteria score without an identifiable underlying bias. While this kind of arbitrariness is arguably less bad than commensuration bias motivated by social bias (as it need not track and therefore exacerbate wider social patterns of discrimination), it still counts as commensuration bias as it privileges one proposal over another despite better criteria scores, thus introducing bias at the commensuration step of the peer review process.

Non-Dominance (Dom) It is not the case that one criterion dominates all the others, i.e., there does not exist a criterion c_j such that for any profile and for any two proposals x and y, $s_{1j}(x) > s_{1j}(y)$ implies xP_1y.

Failure of (Dom) would be an extreme case of the first type of commensuration bias, described by Lee (2015, section 3). In such a case one criterion can overrule all others, which seems to go against the spirit of asking reviewers to score proposals on multiple criteria and then "weigh" these scores to come to an overall score. While there may be some cases where one criterion just is more important than the others, I take it that more typically the intention behind asking a reviewer to score proposals on multiple criteria is for her to lend real weight to each one. If (Dom) is violated, however, all but one of the

criteria are irrelevant to the overall score, except perhaps in a lexicographic sense, i.e., by acting as a tie-breaker.

Independence of Irrelevant Alternatives (I) The relative overall ranking of two proposals x and y depends only on the criteria scores of those two proposals. That is, if two profiles give the same criteria scores to x and y ($s_{1j}(x) = s'_{1j}(x)$ and $s_{1j}(y) = s'_{1j}(y)$ for all criteria c_j) then they should rank x and y the same (xR_1y if and only if xR'_1y).

This requirement follows from the following principle: in order to assess the merit of a particular proposal, one needs to read only that proposal (as opposed to a set of proposals; I do not mean to rule out reading background literature). In particular, a proposal's overall score should depend only on its criteria scores. Hence, if a particular proposal receives the same criteria scores on two different profiles, it should receive the same overall score on these profiles. So if two proposals x and y receive the same criteria scores on two profiles, both should get the same overall score, which entails they should be ranked the same (xR_1y if and only if xR'_1y). Violating (I) thus means violating the principle that overall scores should depend only on criteria scores, thereby instantiating the second type of commensuration bias.

Putting this in terms more familiar to social choice theorists, (I) says that how two proposals are ranked is not allowed to depend on how either of them ranks with respect to some third proposal. As is often pointed out, this is a fairly restrictive requirement which does a lot of the heavy lifting in the proof of Arrow's impossibility theorem. In the present context, it says that the reviewer should not take into account which bundles of proposals are likely to get funded based on her scores. The following example illustrates why one might take this to be an unreasonably restrictive requirement.

Consider two proposals x and y on fairly disparate topics. For example, suppose both proposals are submitted to the NIH, but proposal x concerns a comparative study of different antibiotics whereas proposal y focuses on

genetic determinants of cardiovascular disease, say. If there are a number of other strong proposals having to do with antibiotics but few or none focusing on genetics it may well seem reasonable to the reviewer to give a high overall score to proposal y, giving it a good chance to be funded. But if instead many other proposals focus on understanding the causes of various diseases but few actually study treatments the reviewer might want to give a high overall score to proposal x. In particular, we might imagine that proposals x and y are exactly the same in both scenarios, receiving the same criteria scores, but with proposal y scoring higher overall in the former scenario, and proposal x scoring higher overall in the latter.

The type of reasoning the reviewer seems to engage in here ("This proposal should get a high overall score because there are few other strong proposals in this area, whereas I will give that one a lower score to avoid funding too many proposals in that area.") is ruled out by (I). Yet I am sympathetic to a reviewer who would like to include such considerations—call them "bundle considerations"—in her scoring. So where does this leave the argument that violating (I) constitutes an instance of commensuration bias?

In response, I would first point out that the guidelines given to reviewers by grant agencies seem to rule out using bundle considerations in coming to overall scores. The NIH, for example, explicitly lists in its instructions to reviewers all factors that are supposed to enter into determining the overall score (called "impact score").

> The impact score for an application is based on each individual reviewer's assessment of the scored criteria plus additional criteria regarding the protection and inclusion of human subjects; vertebrate animal care and welfare; biohazards, and criteria specific to the funding opportunity. (National Institutes of Health 2017)

The "scored criteria" mentioned in this quote refers to the previously listed criteria: significance, investigator(s), innovation, approach, and environment.

So at least according to the NIH, reviewers are supposed to consider proposals exclusively on their own merit.

Moreover, this way of thinking seems to be typical among funding agencies and among peer reviewers and academics more generally. It is common to speak of *the* quality of a paper or a proposal, in a way that strongly suggests that this is an inherent feature of the work not dependent on bundle considerations. And peer review is commonly thought to be about identifying this quality, e.g., it is "the means by which one's equals assess the quality of one's scholarly work" (Eisenhart 2002, 241), whereas bias may be defined as "any systematic effect on ratings unrelated to *the true quality* of the object being rated" (Blackburn and Hakel 2006, 378, emphasis mine). Bundle considerations reflect a deviation from this view, and more specifically from the NIH's reviewer instructions, and in that sense might be said to bias the process.

At this point one might correctly point out that I have changed the terms of the discussion. Previously I was making normative claims about what unbiased commensuration should look like, but now I am making a descriptive claim about what funding agencies like the NIH might perceive as bias, without arguing that they are normatively right to do so. In fact I have already suggested that I think they may be wrong to do so, and that bundle considerations may well be a reasonable factor for the reviewer to take into account in determining her overall scores.

However, even if one insists on the importance of bundle considerations, violating (I) introduces bias. The reason for this is that bundle considerations can be incorporated into the framework as an extra criterion. In the example above I sketched two scenarios in which one or the other of two identical proposals seemed preferable due to the available alternatives. But identical proposals need not receive identical criterion scores if one or more criteria explicitly reference bundle considerations—in fact this is a fallacy encouraged by the widespread view that a proposal's merit depends only on it. We

could either add a criterion (perhaps "uniqueness") or use one of the existing criteria (innovation) to reflect in the criteria scores the fact that one proposal stands out by being different.

Once *all* relevant factors are represented in the criteria scores, the idea that identical criterion scores should lead to identical overall scores once again seems reasonable, in fact almost a tautology. With all relevant factors represented in the criteria scores, ranking proposals with identical scores differently by definition means that an irrelevant factor has entered into the overall scores. Since this bias is introduced at the commensuration step, we again end up with commensuration bias.

Taking stock, I have argued that a peer reviewer asked to score grant proposals on both a set of criteria and overall should satisfy requirements (P), (Dom), and (I) if she is to avoid commensuration bias. If, moreover, the criteria are scored on ordinal or cardinal scales that are not intercomparable, and she is to provide overall scores regardless of what combination of criteria scores she decides to give, she faces the following problem.

Theorem 1 (Arrow 1951)**.** *If there are at least three proposals ($m \geq 3$), it is impossible for a commensuration function to simultaneously satisfy (U), (P), (Dom), and (I).*

This is Arrow's famous impossibility theorem. In the present context it says that it is impossible for a reviewer to score a set of at least three proposals without falling prey to commensuration bias. This interpretation of the theorem follows from the arguments given above that violating any of the four requirements constitutes a form of commensuration bias.

While some variations will be considered below, this is the main result of the paper. It is a significant strengthening of the conclusions of Lee (2015). Where Lee introduced the concept of commensuration bias and provided some evidence that this type of bias sometimes occurs, I have argued that commensuration bias necessarily occurs in a wide range of peer review processes of grant proposals (the applicability of this argument to journal peer

review is one of the variations to be considered below).

One might be disappointed by this result, but there is a more optimistic interpretation. As is commonly suggested for Arrow's theorem as well as other impossibility results, rather than focusing on the impossibility one can interpret the theorem as giving a typology of possibilities. In light of the theorem, peer review will be biased in some way or other. The conditions of the theorem can then be interpreted as ways in which peer review might be biased, which one can then evaluate relative to one another.

Is the type of commensuration bias that results from violating (P) to be preferred over the type that results from violating (I)? Or should the problem be avoided by effectively having only a single criterion—violating (Dom); by restricting the possible combinations of criteria scores—violating (U); or by broadening the informational basis so that criteria scores are measured on ratio scales or are somehow made intercomparable?

I have argued that the latter two options present major practical difficulties. But in concluding this section I want to emphasize that one can accept my main argument—that commensuration bias is a necessary feature of grant peer review as currently practiced—even if one disagrees about what can or should be done in light of this.

4 Aggregating Reviewers' Overall Scores

There is another problem of aggregation that comes up in the context of grant proposal peer review. This is the problem of aggregating the (overall) scores given to the proposals by multiple reviewers into a single final ranking that will be used to decide which proposals should be funded. The problem is structurally very similar to the problem of commensurating a single reviewer's criteria scores, as I will now show by putting it into the same framework and demonstrating how Arrow's theorem comes up a second time over.

Before I focused on a single peer reviewer. Now consider n reviewers

$a_1, \ldots, a_n$, again tasked with ranking m proposals. In this section I set aside the notion of criteria, or alternatively, I assume that the problem of aggregating the reviewers' judgments on the criteria into a single ranking of the proposals has somehow been solved.

Instead I assume only that each peer reviewer has scored the proposals. For reviewer a_1 these scores are given by the function s_1 discussed in section 2. Analogously, for any reviewer a_i their scores are given by the function s_i. Once again the question rises on what type of scale these scores are measured and whether they are intercomparable. For the same reasons given in section 2, I think the scores should be interpreted as being on a interval scale (or possibly merely an ordinal scale) as there does not seem to be a meaningful zero.

The issue of interreviewer comparability is less clear. Arguably some degree of comparability can be achieved by instructing reviewers on how to translate qualitative judgments into numerical scores. For example, reviewers might be told that the highest score should be reserved for proposals they think should be funded no matter what, the next highest for those that the reviewer thinks should be funded given availability of funds, the next highest after that for borderline cases, and so on. This might even be supplemented with further instructions regarding the circumstances under which a proposal should be viewed as falling into one of these categories. And funding agencies do in fact give these types of instructions to their reviewers.

On the other hand it is not at all clear that each reviewer will apply these instructions in the same way. Anecdotally at least, the notions of "soft" and "harsh" reviewers are familiar. And then there is the question of whether busy reviewers even read the instructions given by the agency. In order to set up the closest possible analogy with the case of commensuration by a single reviewer, for the moment I will assume that there is no interreviewer comparability. But I return to this issue in the discussion below and in the next section.

The program director receives the peer reviewers' scores. Her task is to give a single ranking of the proposals, such that depending on the funding available, a cutoff point can be chosen: proposals above the cutoff (often called "the payline") will be funded. It is not uncommon for the cutoff point to be chosen after the ranking exercise, so that a complete ranking is indeed needed. At many funding agencies, these decisions are made by a panel rather than a single program director. Where I write "program director" below this should not be read as excluding that possibility.

The final ranking is denoted R, where xRy denotes "x ranks at least as high as y in the final ranking". As before we have the associated relations I for proposals ranked equally and P to denote ranking strictly higher. If the program director is to be free of commensuration bias, the final ranking must be related to the individual reviewer scores in a sensible way.

A combination of reviewer scores—an n-tuple $(s_1, \ldots, s_n)$—is called a profile. We are interested in a function that assigns to a profile a corresponding final ranking. To distinguish it from the function discussed previously, I will call such a function an *aggregation function*.

Universal Domain (U) The domain of the aggregation function is the set of all possible profiles of reviewer scores.

As each reviewer is presumably ranking the proposals independently, there is little reason to think that any combination of reviewer scores can or should be excluded *a priori*. At least in the case of a top medical journal, peer reviewers have been found to agree with each other's judgments "at a rate barely exceeding what would be expected by chance" (Kravitz et al. 2010, 3). If this finding can be generalized to the case of grant proposal review, it would give a positive reason to expect reviewer scores to be all over the map. Since the program director generally does not have the freedom to decide simply not to produce a final ranking in particularly difficult cases, it seems that violating (U) is not a realistic way to avoid commensuration bias.

Weak Pareto (P) If a proposal scores higher than another proposal according to all reviewers it should be higher in the final ranking, i.e., $s_i(x) > s_i(y)$ for all reviewers a_i entails xPy.

If the program director were to go against a unanimous judgment from the reviewers that one proposal is better than another she would seem to have inserted her own opinion into the process, contrary to her task which is to passively aggregate the reviewer scores. This would be a form of the second type of commensuration bias as identical scores would not produce identical rankings.

Non-Dictatorship (D) It is not the case that one reviewer dominates all the others, i.e., there does not exist a reviewer a_i such that for any profile and for any two proposals x and y, $s_i(x) > s_i(y)$ implies xPy.

Just as requirement (P) rules out one form of bias for or against specific proposals, requirement (D) rules out a particularly strong bias in favor of one reviewer. Arguably, a certain respect for reviewers' time and expertise entails that they should be treated interchangeably. If two reviewers' scores were switched (i.e., all the same scores are reported but by different reviewers) this should not affect the final ranking; anything short of this is a form of the second type of commensuration bias.

This argument supports a requirement called "anonymity" (any two profiles in which the same scores are reported but by different reviewers should be treated the same by the aggregation function) which is also sometimes used in social choice theory and is strictly stronger than (D). I use the weaker requirement (D) here because it is all that is needed for the theorem below and to preserve the close analogy with the previous section. Contrary to anonymity, (D) allows reviewers to have specific areas of expertise or even for some reviewer's scores to count more heavily than others', as long as it is not the case that one reviewer can overrule the others on all proposals and regardless of how strongly the others disagree.

Independence of Irrelevant Alternatives (I) The relative final ranking of two proposals x and y depends only on the reviewer scores of those two proposals. That is, if two profiles give the same reviewer scores to x and y ($s_i(x) = s'_i(x)$ and $s_i(y) = s'_i(y)$ for all reviewers a_i) then they should rank x and y the same (xRy if and only if $xR'y$).

The discussion here is largely analogous to the discussion of requirement (I) in the previous section. Because the program director's task is simply to passively aggregate the reviewers' scores, and because "bundle considerations" are either ruled out by the background assumption that a particular proposal's merit depends only on the proposal itself or are already incorporated into the individual reviewers' scores, two proposals x and y that receive identical scores on two profiles should be perceived as being equally meritorious on either profile, and so should be ranked the same (either x outranks y on both profiles, or vice versa, or they are ranked equally). Any deviation from this—and hence any violation of requirement (I)—should be regarded as an instance of commensuration bias.

The argument for requirement (I) is stronger in this case than in the setting of the previous section. As I have imagined it here, the program director that comes up with the final ranking is supposed to be completely passive, which is to say she defers to the expertise of the peer reviewers and aggregates their scores with minimal insertion of her own opinions. Arguably then, any bundle considerations should be reflected in the reviewers' scores, and not in the process by which they are aggregated.

Structurally speaking, both the framework and the requirements just described are exactly the same as those discussed previously. It should be no surprise, then, that the same theorem holds.

Theorem 2 (Arrow 1951)**.** *If there are at least three proposals ($m \geq 3$), it is impossible for an aggregation function to simultaneously satisfy (U), (P), (D), and (I).*

Given my arguments that violating each of the requirements constitutes commensuration bias, the theorem says that it is impossible to avoid commensuration bias, or alternatively that commensuration bias is a necessary feature of the type of peer review process studied here.

As an aside, I note that theorem 2 is directly analogous to Arrow's original theorem, in the sense that what is being aggregated are n voters' (here: peer reviewers') preference rankings of a set of options (here: proposals). By contrast, theorem 1 of the previous section involves a reinterpretation of Arrow's result, in which different criteria act as "voters". This reinterpretation is instead analogous to Okasha (2011), who applied social choice theory to the problem of theory choice.

The assumption of no interreviewer comparability is crucial to the theorem above, as noted in the following proposition.

Proposition 3. *If reviewer scores are comparable (i.e., are measured on the same scale), there exist aggregation functions that simultaneously satisfy (U), (P), (D), and (I).*

For example, if reviewer scores are measured on intercomparable interval scales, the four requirements are satisfied by a utilitarian rule that assigns a weight to each reviewer (with at least two reviewers receiving nonzero weight) and ranks a proposal above another if and only if the weighted average of the reviewer scores of the former is higher than the latter.

Since I have suggested that (some degree of) interreviewer comparability may hold in the case of grant peer review, whereas intercriteria comparability seems highly unlikely, an escape route from the version of Arrow's theorem discussed in this section appears that is not open to the version discussed in the previous section. This raises the question whether combining the two frameworks allows one to avoid commensuration bias altogether. This is the topic of the next section.

5 Multiple Criteria and Multiple Reviewers

The following objection might be raised against the argument of the previous section: the information given to the program director is needlessly impoverished. She was only given the reviewers' overall scores to work with, but at many funding agencies reviewers are asked to score proposals on a number of criteria as well as giving overall scores (as discussed in sections 2 and 3). Would the program director be able to escape Arrow's theorem by considering reviewers' criteria scores?

Moreover, in theorem 2 the program director treated reviewer scores as being given on separate (not intercomparable) scales. I noted that enriching the information available to her by treating the reviewer scores as being on the same scale offers an escape route from the theorem. This appears to be what funding agencies attempt to do when they instruct reviewers on how to use the numerical scales on which proposals are scored. Does interreviewer comparability provide an escape from both versions of Arrow's theorem?

This section addresses both of these points by considering a "double" aggregation framework in which multiple reviewers score proposals on multiple criteria. The program director needs to extract from these scores a single ranking that will determine which proposals get funded. While nothing in the framework forbids "intermediate" aggregations (either single reviewers' overall scores or aggregated reviewer scores on single criteria), I assume here that only the final ranking is ultimately of interest and is to be assessed in the light of possible commensuration bias. The development in this section closely follows List (2004).

Suppose there are n peer reviewers $a_1, \ldots, a_n$ scoring m proposals on k criteria $c_1, \ldots, c_k$. For any proposal x, let $s_{ij}(x)$ denote the score reviewer a_i assigns to x on criterion c_j. As before, assume that these scores are given on a cardinal or ordinal scale, i.e., there is no meaningful zero. For the moment, I make no assumption on intercomparability.

The final ranking determined by the program director is denoted by the

relation R and the derivative relations I and P, as in the previous section. A *double aggregation function* assigns a final ranking to any profile—an $n \cdot k$-tuple $(s_{11}, \ldots, s_{nk})$—in its domain, which is some given subset of all possible profiles.

In order to avoid falling prey to commensuration bias, a double aggregation function needs to satisfy a number of conditions. The first three of these are straightforward generalizations of the conditions given in previous sections. The arguments for why violating these requirements constitutes commensuration bias are unchanged from those given above. Note that the versions of (P) and (I) given here are somewhat weaker due to their antecedents being stronger, requiring agreement between all reviewers *and* all criteria.

Universal Domain (U) The domain of the double aggregation function is the set of all possible profiles of criteria scores.

Weak Pareto (P) If a proposal scores higher than another proposal on all criteria according to all reviewers it should be higher in the final ranking, i.e., $s_{ij}(x) > s_{ij}(y)$ for all criteria c_j and reviewers a_i entails xPy.

Independence of Irrelevant Alternatives (I) The relative final ranking of two proposals x and y depends only on the criteria scores of those two proposals. That is, if two profiles give the same scores to x and y ($s_{ij}(x) = s'_{ij}(x)$ and $s_{ij}(y) = s'_{ij}(y)$ for all reviewers a_i and criteria c_j) then they should rank x and y the same (xRy if and only if $xR'y$).

The generalization of the non-dictatorship/dominance condition is slightly less straightforward. Following List (2004), I formulate three such conditions. The first one requires that no single individual reviewer acts like a dictator, without specifying how her criteria scores are aggregated. The second one requires that no single criterion dominates the final ranking, without

specifying how individual reviewers' scores on that criterion are aggregated. Finally, the third and weakest version only rules out that a single score function (i.e., a single reviewer's scores on a single criterion) dominates the final ranking.

Non-Dictatorship (D) There does not exist a reviewer a_i and a strictly increasing function $f : \mathbb{R}^k \to \mathbb{R}$ such that for any profile and for any two proposals x and y, $f(s_{i1}(x), \ldots, s_{ik}(x)) > f(s_{i1}(y), \ldots, s_{ik}(y))$ implies xPy.

Non-Dominance (Dom) There does not exist a criterion c_j and a strictly increasing function $f : \mathbb{R}^n \to \mathbb{R}$ such that for any profile and for any two proposals x and y, $f(s_{1j}(x), \ldots, s_{nj}(x)) > f(s_{1j}(y), \ldots, s_{nj}(y))$ implies xPy.

Non-Double-Dictatorship (DD) There does not exist a reviewer a_i and a criterion c_j such that for any profile and for any two proposals x and y, $s_{ij}(x) > s_{ij}(y)$ implies xPy.

If there is neither interreviewer comparability nor intercriteria comparability the double aggregation problem reduces to a regular aggregation problem with $n \cdot k$ individuals. Hence Arrow's theorem applies, and in the present framework says the following.

Theorem 4 (Arrow 1951)**.** *If there is neither interreviewer comparability nor intercriteria comparability and there are at least three proposals ($m \geq 3$), it is impossible for a double aggregation function to simultaneously satisfy (U), (P), (I), and (DD).*

As I suggested above, however, it may be reasonable to expect some degree of interreviewer comparability, as reviewers may be instructed to score proposals in broadly similar ways. The following theorem applies to this scenario.

Theorem 5 (Roberts 1995 / List 2004)**.** *If there is interreviewer comparability but not intercriteria comparability and there are at least three proposals ($m \geq 3$), it is impossible for a double aggregation function to simultaneously satisfy (U), (P), (I), and (Dom).*

This answers the questions from the beginning of this section. Despite interreviewer comparability, and despite the broader informational basis provided by the presence of scores on multiple criteria from multiple reviewers, an analogue of theorem 1 of section 3 goes through. In this most general version of the model it still turns out that it is impossible to avoid commensuration bias.

Finally, although in my opinion not as relevant to the case of grant proposal reviewing, the previous theorem can be reinterpreted to apply when there is intercriteria comparability but not interreviewer comparability.

Theorem 6 (Roberts 1995 / List 2004)**.** *If there is intercriteria comparability but not interreviewer comparability and there are at least three proposals ($m \geq 3$), it is impossible for a double aggregation function to simultaneously satisfy (U), (P), (I), and (D).*

For the sake of completeness, I should mention that in the presence of both interreviewer and intercriteria comparability, all the criteria can be satisfied simultaneously. Possibilities similar to the one sketched at the end of section 4 then arise (see List 2004, sections 4.3 and 4.4, for more details). However, due to the absence of intercriteria comparability (as discussed in section 2), this does not make for a plausible response to the problem of commensuration bias in grant proposal review.

6 Commensuration Bias At Scientific Journals

Lee (2013, section 3) argues for the existence of commensuration bias not

just at funding agencies, but also at top scientific journals. So far I have focused exclusively on funding agencies. To what extent do my arguments extend to peer review practices at journals?

One important difference is that journals tend to use different peer reviewers for different papers. A single peer reviewer will not normally be asked to review more than a few papers per year for a given journal. So the framework as laid out here, in which a single reviewer scores all papers in a given set, does not apply as naturally to journal peer review. In contrast, at a funding agency it is more common that the same group of reviewers is used for all proposals (or at least all proposals on a particular topic, or submitted to a particular panel).

This point, however, need not be prohibitive to the application of the social choice theory framework to journal peer review. At least from a formal perspective, nothing prevents us from treating "Reviewer 1" (the first reviewer to provide scores on each paper) as a single reviewer, and similarly for "Reviewer 2" and so on. Assuming each paper has the same number of reviewers this fits the framework laid out above. Then we can still ask whether Reviewer 1 aggregates her criteria scores consistently in a way that is free from commensuration bias, and whether the journal editor aggregates the reviewer scores in an unbiased way.

If instances of commensuration bias are found, it may be less clear who is to blame: if Reviewer 1 evaluates papers inconsistently, is that the fault of any particular reviewer who acted as Reviewer 1 on one or more papers? But even so, it identifies a problem in the peer review process: if requirements (U), (P), (I), (D), or (Dom) are being violated, some papers are being treated unfairly by the journal. This may, and arguably should, be something a journal would want to fix. The question of blame is perhaps a bit of a red herring.

Another difference between journals and funding agencies is that almost all journals review and accept papers on a rolling basis, whereas funding

agencies usually evaluate all proposals in response to a particular call at once. I argued above that funding agencies need to come up with a full ranking of the proposals, as the payline is often not known until late in the process. Journals instead need to make binary decisions—accept or reject this paper—with a page limit in mind. Moreover, compared to paylines, page limits tend to be a little more fungible at printed journals (as backlogs can be grown or shrunk up to a point) and much more fungible at online-only journals.

So journals are probably better modeled as using some kind of threshold on overall scores (i.e., a paper is accepted if it scores above the threshold, with the threshold gradually adjusted over time in view of the page limit) rather than creating a ranking of batches of papers. For this reason I think the framework studied here does not apply neatly enough to journal peer review to support the kind of claims I made about commensuration bias in grant peer review. Adapting the framework to ask whether commensuration bias necessarily arises in journal peer review I leave for future research. I expect that similar results may be achieved in such a model.

7 Conclusion

This paper has argued that commensuration bias is a necessary feature of peer review at funding agencies, assuming it is organized broadly along the lines it currently is at for example the NIH.

I already mentioned that one might view Arrow's theorem as giving a typology of possibilities. For those who are committed to a form of grant peer review as presently organized (with different criteria that are measured on ordinal or cardinal scales that are not intercomparable), future research could fruitfully investigate the different possibilities that arise when one of the requirements (U), (P), (Dom), or (I) is weakened. While I have argued that violating each of these makes for commensuration bias, this is not to

say that all forms of commensuration bias are equally bad.

Alternatively, one might consider more far-reaching reforms to peer review. One proposal that appears to be gaining some momentum is the idea to fund grant proposals by lottery, usually combined with some minimal screening through peer review (Fang and Casadevall 2016, Guthrie et al. 2017, Avin forthcoming). In other work I have suggested that the role of journal peer review in science should be significantly reduced (Heesen and Romeijn 2018, Heesen and Bright 2018). These suggestions may come with other downsides, but they would surely suffice to eliminate commensuration bias in peer review.

References

Kenneth J. Arrow. *Social Choice and Individual Values*. John Wiley, New York, 1951.

Shahar Avin. Centralised funding and epistemic exploration. *The British Journal for the Philosophy of Science*, forthcoming. doi: 10.1093/bjps/axx059. URL `http://dx.doi.org/10.1093/bjps/axx059`.

Jessica L. Blackburn and Milton D. Hakel. An examination of sources of peer-review bias. *Psychological Science*, 17(5):378–382, 2006. doi: 10.1111/j.1467-9280.2006.01715.x. URL `http://dx.doi.org/10.1111/j.1467-9280.2006.01715.x`.

Margaret Eisenhart. The paradox of peer review: Admitting too much or allowing too little? *Research in Science Education*, 32(2):241–255, 2002. ISSN 1573-1898. doi: 10.1023/A:1016082229411. URL `http://dx.doi.org/10.1023/A:1016082229411`.

Ferric C. Fang and Arturo Casadevall. Research funding: the case for a mod-

ified lottery. *mBio*, 7(2):e00422–16, 2016. doi: 10.1128/mBio.00422-16. URL http://dx.doi.org/10.1128/mBio.00422-16.

Susan Guthrie, Ioana Ghiga, and Steven Wooding. What do we know about grant peer review in the health sciences? *F1000Research*, 6:1335, 2017. doi: 10.12688/f1000research.11917.2. URL http://dx.doi.org/10.12688/f1000research.11917.2. Version 2.

Remco Heesen. When journal editors play favorites. *Philosophical Studies*, 175(4):831–858, 2018. ISSN 0031-8116. doi: 10.1007/s11098-017-0895-4. URL http://dx.doi.org/10.1007/s11098-017-0895-4.

Remco Heesen and Liam Kofi Bright. Is peer review a good idea? Manuscript, 2018.

Remco Heesen and Jan-Willem Romeijn. Epistemic diversity and editor decisions: A statistical Matthew effect. Manuscript, 2018.

Richard L. Kravitz, Peter Franks, Mitchell D. Feldman, Martha Gerrity, Cindy Byrne, and William M. Tierney. Editorial peer reviewers' recommendations at a general medical journal: Are they reliable and do editors care? *PLoS ONE*, 5(4):e10072, 2010. doi: 10.1371/journal.pone.0010072. URL http://dx.doi.org/10.1371/journal.pone.0010072.

Carole J. Lee. The limited effectiveness of prestige as an intervention on the health of medical journal publications. *Episteme*, 10(4):387–402, 2013. doi: 10.1017/epi.2013.35. URL http://dx.doi.org/10.1017/epi.2013.35.

Carole J. Lee. Commensuration bias in peer review. *Philosophy of Science*, 82(5):1272–1283, 2015. doi: 10.1086/683652. URL http://dx.doi.org/10.1086/683652.

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. ISSN 1532-2890. doi: 10.1002/asi.22784. URL http://dx.doi.org/10.1002/asi.22784.

Christian List. Multidimensional welfare aggregation. *Public Choice*, 119(1): 119–142, 2004. ISSN 1573-7101. doi: 10.1023/B:PUCH.0000024168.00362.af. URL http://dx.doi.org/10.1023/B:PUCH.0000024168.00362.af.

National Institutes of Health. Guidelines and review criteria, 2017. URL https://grants.nih.gov/grants/policy/review_templates.htm. Accessed August 22, 2018.

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

Kevin Roberts. Valued opinions or opinionated values: The double aggregation problem. In K. Basu, P. K. Pattanaik, and K. Suzumura, editors, *Choice, Welfare, and Development: A Festschrift in Honour of Amartya K. Sen*, pages 141–165. Oxford University Press, Oxford, 1995.

Jennifer Saul. Implicit bias, stereotype threat, and women in philosophy. In Katrina Hutchison and Fiona Jenkins, editors, *Women in Philosophy: What Needs to Change?*, chapter 2, pages 39–60. Oxford University Press, Oxford, 2013.

Amartya Kumar Sen. *Collective Choice and Social Welfare*. Holden Day, San Francisco, 1970.

Eran Tal. Measurement in science. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Fall 2017 edition, 2017. URL https://plato.stanford.edu/archives/fall2017/entries/measurement-science/.

Kevin J. S. Zollman. Optimal publishing strategies. *Episteme*, 6(2):185–199, Jun 2009. ISSN 1750-0117. doi: 10.3366/E174236000900063X. URL http://dx.doi.org/10.3366/E174236000900063X.