Ergo AN OPEN ACCESS JOURNAL OF PHILOSOPHY

The Necessity of Commensuration Bias in Grant Peer Review

REMCO HEESEN
University of Western Australia and University of Groningen

Peer reviewers at many funding agencies and scientific journals are asked to score submissions both on individual criteria and overall. The overall scores should be some kind of aggregate of the criteria scores. Carole Lee identifies this as a potential locus for bias to enter the peer review process, which she calls commensuration bias. Here I view the aggregation of scores through the lens of social choice theory. I argue that, when reviewing grant proposals, it is in many cases impossible to avoid commensuration bias.

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 Avin 2019; Heesen & Bright 2021; Lee 2013; Zollman 2009).

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, Sugimoto, Zhang, & Cronin 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 potential effects (Heesen 2018; Heesen & Romeijn 2019; Lee et al. 2013; Saul 2013).

This paper focuses on a type of bias newly identified by Lee (2015), which she calls *commensuration bias*. Commensuration is the activity of aggregating

Contact: Remco Heesen <remco.heesen@uwa.edu.au>

https://doi.org/10.3998/ergo.2240

different quantities into a single number (see Espeland & Stevens 1998 for background on the concept). 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.

Building on Lee's work, I distinguish two types of commensuration bias. The first type, which is her primary focus and for which she provides substantial evidence, 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). Grant agencies, Lee goes on to argue, overweight methodological criteria relative to novelty. The idea here is that reviewers systematically overweight some criteria at the expense of others, regardless of the content of individual proposals (or papers, but I focus on grant proposals). I refer to this as *proposal-independent commensuration bias*.

A potential difficulty in identifying proposal-independent commensuration bias is that it requires a substantive view on what counts as overweighting a criterion. This is Lee's approach: she argues that significance is overweighted at journals by pointing to the lack of insignificant results in the literature and the negative consequences of this phenomenon (known as publication bias) for science in general and meta-analysis in particular (Lee 2015: 1276–77). For the case of grant review, she argues that present practice amounts to conservatism, contrary to the stated goals of grant agencies (Lee 2015: 1277–78).

Taking a more agnostic approach, I invoke proposal-independent commensuration bias only in the particularly stark case where a peer reviewer gives a higher overall score to a grant proposal whenever it scores higher on the privileged criterion, regardless of the other criteria. 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.

The second type of commensuration bias, mentioned by Lee (2015: 1273–74) and elaborated in Erosheva, Grant, Chen, Lindner, Nakamura, and Lee (2020), refers to reviewers weighting peer review criteria differentially when assigning scores to different submissions. Call this *proposal-dependent commensuration bias*. To make this a little more precise, I say that proposal-dependent 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 two different ways of instantiating it.

Suppose a peer reviewer is (explicitly or implicitly) prejudiced, that is, her judgment of the quality of individual grant proposals is affected by the gender, race, or other social characteristics of the 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 (as suggested by Lee 2015: 1274). 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 proposal-dependent commensuration bias. Erosheva et al. (2020) attempt to find evidence for this form of commensuration bias against black applicants at the National Institutes of Health (NIH), the American agency responsible for government grants for medical research, but they report mostly null results.

Alternatively, suppose that for a particular reviewer the overall score of a proposal depends, in addition to its individual scores, also on the individual scores of one or more other proposals. This is an example of proposal-dependent commensuration bias because it violates the principle of identical overall scores for identical individual scores. It also militates against the widespread idea that a proposal's merit 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 necessarily occurs (or, if not that, is at least very hard to avoid) in any peer review context where multiple criteria are used. More precisely, I argue that if the commensuration process used is flexible enough to deal with a range of possible combinations of submissions, it will be subject to at least one of the two types of commensuration bias. 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. Section 5 combines the two cases studied in the preceding sections, and Section 6 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 (2018) uses the following criteria: "significance", "investigator(s)" (suitability of the applicants to carry out the research), "innovation", "environment" (suitability of the research environment), and "approach" (see Hug & Aeschbach 2020 for a systematic review of criteria used in grant peer review).

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 are 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 (2018), 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 judged to be better than proposals with a lower score. This induced ranking is 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 $xI_1\, y$ denotes $s_1\,(x) = s_1\,(y)$ or "x and y rank equally".

In social choice theory, the collection of the reviewer's individual criteria scores for all proposals is called a *profile*. A function, defined on some given domain of profiles, which assigns to each profile a corresponding set of overall scores, I call a *commensuration function*.

The first substantive question I address is: how much information is contained 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 (Tal 2020: 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 monotonic 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. Temperature as measured on the Celsius or Fahrenheit scales is the standard example.

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 assume that scores are given 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 assume throughout this paper that criteria scores (and overall scores) are measured on either an ordinal or a cardinal scale (for the formal results, it does not matter which).

How about intercriteria comparability? Here, the question is whether statements like "The significance of proposal x is higher than the suitability of the applicant of proposal y" 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 numerical scale to score proposals on (say, a 1 to 9 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 qualitatively 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, how flexible do we require a commensuration function to be, and what needs to be true for it to be free of proposal-independent commensuration bias and proposal-dependent 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.

Prima facie these alternatives are not very attractive. It seems that any reasonable 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) (Okasha 2011: 92 makes the same point in a similar context). But it has been argued that there can be principled reasons for thinking certain combinations of criteria scores will never come up (Morreau 2014) or are unlikely to come up (Marcoci & Nguyen 2017).

In response I would stress that there is an element of subjectivity in assigning criteria scores (if there were not, all reviewers would assign the same criteria scores, which is empirically false). Thus, even if there are good reasons not to expect certain combinations of criteria scores, those combinations may arise anyway due to subjective variation from what the criteria scores "should" be (assuming this notion of "should" makes sense at all). An unbiased commensuration function should be able to handle these possibilities. In fact, it should in principle be specifiable before the criteria scores are known (cf. Okasha 2015: 285–86).

An alternative response would focus on the fact that (U) can be significantly weakened without affecting any of what I write below. I do not pursue this response here as it would require introducing a level of technical detail inappropriate for this paper, which has a more applied focus. For relevant discussions, see Morreau (2014: section 7), Okasha (2015: 290–92), Zwart and Franssen (2007: section 5.1), and Gaertner (2001).

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 this happening with a gender biased reviewer, as described in Section 1. Similarly, a reviewer might violate (P) due to racial bias (Erosheva et al. 2020) 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 (Lee 2015: 1273–74 suggests reviewer idiosyncracy as a source of commensuration bias; for more on reviewer idiosyncracy see Bornmann & Daniel 2009; Cole, Cole, & Simon 1981; Lamont 2009). 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 proposal-dependent 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 proposal-independent commensuration bias. 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 (Okasha 2011: 95–96 discusses this suggestion in more detail), 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. In particular, a proposal's overall score should depend only on its criteria scores (this is

consistent with the idea that reading background literature or other proposals can improve the quality of a reviewer's judgment as suggested by Jayasinghe, Marsh, & Bond 2003: section 6.2 and Marsh, Jayasinghe, & Bond 2008: 163–64). 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 proposal-dependent commensuration bias.

As a technical note, there are multiple formulations of (I) in the literature. Here I follow the one by List (2004: 128), which derives from Sen (1970: 129).

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, from a mathematical perspective this is a restrictive requirement. 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. 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?

First, note 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 instructs reviewers to consider each proposal in isolation: "Don't compare one application to another—they should each be evaluated independently based on the review criteria" (National Institutes of Health 2015). Bundle considerations are instead considered at the second stage of review by the advisory council or board (National Institutes of Health 2018). 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 (though it has also been criticized, see Bailar & Patterson 1985; Heesen & Romeijn 2019; Lee et al. 2013). 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 & 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 juncture 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 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 identically scored 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—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.

Other objections to (I) can be dealt with in a similar way. Suppose for example that two proposals are given identical scores on all criteria, including a relatively weak score on approach. If one proposal is in an established field and the other in a novel field, arguably the approach score reflects less poorly on the latter,

given that methodological norms are less settled for its field. Would ranking such proposals differently be justified? If so, I maintain this should be reflected in the criterion scores as well. Perhaps approach should be scored relative to the standards of a field (so the score for the proposal in the more novel field should be higher because it does well relative to the looser norms of its field) or perhaps this information should be captured in other (existing or new) criteria.

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. If a reviewer is tempted to give identical criterion scores to two proposals but rank them differently, she either has a valid reason for doing so or not. If she does, there should be a criterion that lets her express that reason in her criterion scores (and if such a criterion does not exist it should be added). Otherwise, i.e., if she wants to rank proposals differently but cannot identify a relevant difference between the proposals, her ranking is expressing some bias. Since this bias is introduced at the commensuration step, we again end up with commensuration bias.

For this reason, I maintain that violating (I) is a form of commensuration bias. This appears to deviate from the NIH's position: "No formula is used to derive the overall impact score from the individual criterion scores, and reviewers are instructed to weigh the different criteria as they see fit in deriving their overall scores" (National Institutes of Health 2016). However, my position and that of the NIH are consistent if we assume that the NIH expects reviewers to (informally) incorporate additional criteria beyond the official five into their scoring.

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 / Sen 1970). *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, as generalized by Sen (1970: theorem 8*2). 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 are 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 evidence that this type of bias occurs, I have argued that commensuration bias necessarily occurs in a wide range of peer review processes of grant proposals.

One might be disappointed by this result, but there is a more optimistic interpretation. As is commonly suggested, rather than focusing on the impossibility, one can interpret Arrow's theorem (or other impossibility results) as giving a typology of possibilities. In light of the theorem, peer review will be biased in some way or other. The theorem's conditions can then be interpreted as ways in which peer review might be biased, which one can 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 is 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 now show by putting it into the same framework and demonstrating how Arrow's theorem comes up a second time.

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, the scores for any reviewer a_i are given by the function s_i. Once again the question arises 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 through reviewer instructions. For example, reviewers might be told explicitly which numerical scores are appropriate for proposals they think should definitely be funded, should be funded if possible, borderline cases, etc. This might 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 (not to mention busy reviewers who fail to read instructions). In order to set up the closest possible analogy with the case of commensuration by a single reviewer, I assume for the moment 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") are 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. The phrase "program director" 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 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 (and subject to her own subjective preferences and biases), 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, Franks, Feldman, Gerrity, Byrne, & Tierney 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 not to produce a final ranking in 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 proposal-dependent 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 proposal-dependent 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 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 / Sen 1970). *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 Zwart and Franssen (2007) and Okasha (2011), who applied social choice theory to the problems of verisimilitude and theory choice, respectively.

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. Incidentally, this is the process used by the NIH, which takes the (unweighted) average of reviewer scores to determine a proposal's "final overall impact score" (National Institutes of Health 2018).

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 (Morreau 2016 explores this in more detail). The next section raises the question whether combining the two frameworks allows one to avoid commensuration bias altogether.

5. Multiple Criteria and Multiple Reviewers

The following objection might be raised against the development in 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). Can the program director escape Arrow's theorem by considering reviewers' criteria scores?

Moreover, I noted that funding agencies may attempt to enrich the informational basis by instructing reviewers on how to use the numerical scales on which proposals are scored. This offers an escape route from the impossibility presented in theorem 2. 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 decide on a final ranking that determines which proposals get funded. The development in this section closely follows List (2004). It is worth noting that earlier work by McKelvey (1979) already established important difficulties in forming a ranking when evaluating alternatives on multiple dimensions. However, McKelvey assumes the existence of an infinity of alternatives with arbitrarily small differences between them, which seems unrealistic for grant proposals. For this reason I take his work to be less immediately relevant.

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. I make no assumption on intercomparability for now, but I return to this issue shortly.

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 reviewers a_i and criteria c_j 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 $xR'y$).

Following List (2004), I formulate three versions of a non-dictatorship condition. 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. 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 $f: \mathbf{R}^k \to \mathbf{R}$ 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 $f: \mathbf{R}^n \to \mathbf{R}$ 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 / Sen 1970). *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 are then available (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. 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.

An important question for further research is whether and how these results generalize to other selection processes, including in particular other forms of peer review. Lee (2015: section 3) argues for the existence of commensuration bias not just at funding agencies, but also at top scientific journals. However, the peer review process at journals differs in a number of respects from that at funding agencies. Key among these is that journals review and accept papers on a rolling basis. 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. This suggests that the framework used here would have to be adapted to apply to journal peer review. Nevertheless, there are enough similarities that one might expect to run into analogous problems.

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 (Avin 2019; Fang & Casadevall 2016; Guthrie, Ghiga, & Wooding 2018). In other work I have suggested that the role of peer review in science should be significantly reduced (Heesen & Bright 2021; Heesen & Romeijn 2019). These suggestions may come with other downsides, but they would surely suffice to eliminate commensuration bias in peer review.

Acknowledgments

Thanks to Liam Bright, Nick Baigent, Carole Lee, two anonymous referees, and audiences at the Asian Conference on the Philosophy of the Social Sciences (Nankai University) and the SILFS Postgraduate Conference (University of Urbino) for valuable discussion. Thanks to the evaluation committee for honoring this paper with the 2018 SILFS Prize for Philosophy of Science. This work was supported by the Dutch Research Council (NWO) under grant 016.Veni.195.141.

References

Arrow, Kenneth J. (1951). *Social Choice and Individual Values*. John Wiley & Sons.

Avin, Shahar (2019). Centralized Funding and Epistemic Exploration. *The British Journal for the Philosophy of Science*, *70*(3), 629–56. https://doi.org/10.1093/bjps/axx059

Bailar, John C. and Kay Patterson (1985). Journal Peer Review: The Need for a Research Agenda. *New England Journal of Medicine*, *312*(10), 654–57. https://doi.org/10.1056/NEJM198503073121023

Blackburn, Jessica L. and Milton D. Hakel (2006). An Examination of Sources of Peer-Review Bias. *Psychological Science*, *17*(5), 378–82. https://doi.org/10.1111/j.1467–9280.2006.01715.x

Bornmann, Lutz and Hans-Dieter Daniel (2009). The Luck of the Referee Draw: The Effect of Exchanging Reviews. *Learned Publishing*, *22*(2), 117–25. https://doi.org/10.1087/2009207

Cole, Stephen, Jonathan R. Cole, and Gary A. Simon (1981). Chance and Consensus in Peer Review. *Science*, *214*(4523), 881–86. https://doi.org/10.1126/science.7302566

Eisenhart, Margaret (2002). The Paradox of Peer Review: Admitting too Much or Allowing too Little? *Research in Science Education*, *32*(2), 241–55. https://doi.org/10.1023/A:1016082229411

Erosheva, Elena A., Sheridan Grant, Mei-Ching Chen, Mark D. Lindner, Richard K. Nakamura, and Carole J. Lee (2020). NIH Peer Review: Criterion Scores Completely Account for Racial Disparities in Overall Impact Scores. *Science Advances*, *6*(23), eaaz4868. https://doi.org/10.1126/sciadv.aaz4868

Espeland, Wendy N. and Mitchell L. Stevens (1998). Commensuration as a Social Process. *Annual Review of Sociology*, *24*(1), 313–43. https://doi.org/10.1146/annurev.soc.24.1.313

Fang, Ferric C. and Arturo Casadevall (2016). Research Funding: The Case for a Modified Lottery. *mBio*, *7*(2), e00422–16. https://doi.org/10.1128/mBio. 00422–16

Gaertner, Wulf (2001). *Domain Conditions in Social Choice Theory*. Cambridge University Press. https://doi.org/10.1017/CBO9780511492303

Guthrie, Susan, Ioana Ghiga, and Steven Wooding (2018). What Do We Know about Grant Peer Review in the Health Sciences? [version 2; referees: 2 approved]. *F1000Research*, *6*(1335), 1–23. https://doi.org/10.12688/f1000research.11917.2

Heesen, Remco (2018). When Journal Editors Play Favorites. *Philosophical Studies*, *175*(4), 831–58. https://doi.org/10.1007/s11098-017-0895-4

Heesen, Remco and Liam K. Bright (2021). Is Peer Review a Good Idea? *The British Journal for the Philosophy of Science*, *72*(3), 635–63. https://doi.org/10.1093/bjps/axz029

Heesen, Remco and Jan-Willem Romeijn (2019). Epistemic Diversity and Editor Decisions: A Statistical Matthew Effect. *Philosophers' Imprint*, *19*(39), 1–20. http://hdl.handle.net/2027/spo.3521354.0019.039

Hug, Sven E. and Mirjam Aeschbach (2020). Criteria for Assessing Grant Applications: A Systematic Review. *Palgrave Communications*, *6*(37), 1–15. https://doi.org/10.1057/s41599-020-0412-9

Jayasinghe, Upali W., Herbert W. Marsh, and Nigel W. Bond (2003). A Multilevel Cross-Classified Modelling Approach to Peer Review of Grant Proposals: The Effects of Assessor and Researcher Attributes on Assessor Ratings. *Journal of the Royal Statistical Society: Series A*, *166*(3), 279–300. https://doi.org/10.1111/1467–985X.00278

Kravitz, Richard L., Peter Franks, Mitchell D. Feldman, Martha Gerrity, Cindy Byrne, and William M. Tierney (2010). Editorial Peer Reviewers' Recommendations at a General Medical Journal: Are They Reliable and Do Editors Care? *PLoS ONE*, *5*(4), e10072. https://doi.org/10.1371/journal.pone.0010072

Lamont, Michèle (2009). *How Professors Think: Inside the Curious World of Academic Judgment*. Harvard University Press.

Lee, Carole J. (2013). The Limited Effectiveness of Prestige as an Intervention on the Health of Medical Journal Publications. *Episteme*, *10*(4), 387–402. https://doi.org/10.1017/epi.2013.35

Lee, Carole J. (2015). Commensuration Bias in Peer Review. *Philosophy of Science*, *82*(5), 1272–83. https://doi.org/10.1086/683652

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

List, Christian (2004). Multidimensional Welfare Aggregation. *Public Choice*, *119*(1), 119–42. https://doi.org/10.1023/B:PUCH.0000024168.00362.af

Marcoci, Alexandru and James Nguyen (2017). Scientific Rationality by Degrees. In Michela Massimi, Jan-Willem Romeijn, and Gerhard Schurz (Eds.), *EPSA15 Selected Papers: The 5th Conference of the European Philosophy of Science Association in Düsseldorf* (321–33). Springer. https://doi.org/10.1007/978-3-319-53730-6_26

Marsh, Herbert W., Upali W. Jayasinghe, and Nigel W. Bond (2008). Improving the Peer-Review Process for Grant Applications: Reliability, Validity, Bias, and Generalizability. *American Psychologist*, *63*(3), 160–68. https://doi.org/10.1037/0003–066X.63.3.160

McKelvey, Richard D. (1979). General Conditions for Global Intransitivities in Formal Voting Models. *Econometrica*, *47*(5), 1085–112. https://doi.org/10.2307/1911951

Morreau, Michael (2014). Mr. Fit, Mr. Simplicity and Mr. Scope: From Social Choice to Theory Choice. *Erkenntnis*, *79*(6), 1253–68. https://doi.org/10.1007/s10670–013–9549-x

Morreau, Michael (2016). Grading in Groups. *Economics and Philosophy*, *32*(2), 323–52. https://doi.org/10.1017/S0266267115000498

National Institutes of Health (2015). Meeting Do's and Don'ts: Advice for Reviewers. Retrieved on February 25, 2021 from https://grants.nih.gov/grants/policy/review/meeting.htm

National Institutes of Health (2016). Scoring Guidance. Retrieved on February 25, 2021 from https://grants.nih.gov/grants/policy/review/rev_prep/scoring.htm

National Institutes of Health (2018). Peer Review. Retrieved on February 25, 2021 from https://grants.nih.gov/grants/peer-review.htm

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

Okasha, Samir (2015). On Arrow's Theorem and Scientific Rationality: Reply to Morreau and Stegenga. *Mind*, *124*(493), 279–94. https://doi.org/10.1093/mind/fzu177

Roberts, Kevin (1995). Valued Opinions or Opinionated Values: The Double Aggregation Problem. In K. Basu, P. K. Pattanaik, and K. Suzumura (Eds.), *Choice, Welfare, and Development: A Festschrift in Honour of Amartya K. Sen* (141–65). Oxford University Press.

Saul, Jennifer (2013). Implicit Bias, Stereotype Threat, and Women in Philosophy. In Katrina Hutchison and Fiona Jenkins (Eds.), *Women in Philosophy: What Needs to Change?* (39–60). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199325603.003.0003

Sen, Amartya K. (1970). *Collective Choice and Social Welfare*. Holden Day.

Tal, Eran (2020). Measurement in Science. In Edward N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2020 ed.). https://plato.stanford.edu/archives/fall2020/entries/measurement-science/

Zollman, Kevin J. S. (2009). Optimal Publishing Strategies. *Episteme*, *6*(2), 185–99. https://doi.org/10.3366/E174236000900063X

Zwart, Sjoerd D. and Maarten Franssen (2007). An Impossibility Theorem for Verisimilitude. *Synthese*, *158*(1), 75–92. https://doi.org/10.1007/s11229-006-9051-y